THE ASIAN BULLETIN OF BIG DATA MANAGEMENT
Vol. 6. Issue 1 (2026)
https://doi.org/10.62019/2wm4h249

MEV-Forensics: An On-Chain Attribution Framework for Sandwich Attack Variants and Validator-Layer Exploits with Behavioral Intent Scoring

Atta Ullah*, Ali Sufyan, Murtajiz Ali Khan, Muhammad Zaid

Chronicle

Article history
Received: January 10, 2026
Received in the **revised format:** Feb 18. 2026
Accepted: March 03, 2026
Available online March 31, 2026

Atta Ullah, Ali Sufyan, Murtajiz Ali Khan & Muhammad Zaid are currently affiliated with the Department of Information and Communication Engineering, The Islamia University of Bahawalpur, Punjab, Pakistan.
Email: attaullah.waizr.cybersec@gmail.com
Email: ali.sufyan@iub.edu.pk
Email: murtajizali93@gmail.com
Email: muhammadzaid7667@gmail.com

Corresponding Author

Abstract

A jury deadlocked in November 2025 over a 25 million USD blockchain exploit because prosecutors could not translate an immutable transaction record into proof of deliberate deception. That outcome exposes a methodological gap in current blockchain forensic practice. This paper introduces MEV-Forensics, a four-stage on-chain attribution pipeline that addresses three structurally distinct maximal extractable value attack types: classic mempool sandwich attacks, validator-layer exploits targeting MEV-Boost relay infrastructure, and private-channel sandwich attacks. An eight-indicator behavioral intent scoring rubric is used to derive a composite score from on-chain features grounded in specific elements of wire fraud doctrine. Application to a ground-truth dataset of 2,400 documented attacks spanning January 2021 through December 2024 produces precision of 0.891 with 95 percent bootstrap confidence interval of 0.873 to 0.908, recall of 0.847, false-positive rate of 3.2 percent, and area under the ROC curve of 0.934. Five-fold cross-validation confirms mean precision of 0.886 with standard deviation of 0.012. Inter-rater agreement across three independent practitioners, measured using Cohen's weighted kappa with quadratic weights, reaches 0.74. Cross-jurisdictional admissibility analysis maps rubric output to evidentiary standards under United States, United Kingdom, and European Union law.

Keywords: Maximal extractable value; blockchain forensics; sandwich attack; on-chain attribution; digital evidence; intent scoring; Ethereum; validator-layer exploit; graph analysis; machine learning forensics.

INTRODUCTION

Ethereum processes over one million transactions daily across its decentralized exchange ecosystem, and a measurable fraction of that volume involves transaction ordering manipulation that extracts value from other participants through sequencing advantages unavailable to ordinary users (Gramlich et al., 2024). The practice carries an accepted name, maximal extractable value, and an extensive academic measurement record. Existing forensic practice has not produced a validated investigation protocol that converts the on-chain record of such conduct into court-admissible evidence. The prosecutorial gap materialized in public record on November 7, 2025, when U.S. District Judge Jessica Clarke declared a mistrial in United States v. Peraire-Bueno, No. 1:24-cr-00293-GJLC, one of the first criminal prosecutions centered on MEV conduct. Anton and James Peraire-Bueno, two brothers operating as Ethereum validators, had extracted approximately 25 million USD in approximately 12 seconds by exploiting vulnerabilities in MEV-Boost relay software to intercept and redirect transaction bundles submitted by other MEV searcher bots (Department of Justice, 2024). The jury deadlocked. The central dispute concerned intent. Defence counsel argued the brothers followed existing protocol rules. Prosecutors could not

assemble on-chain behavioral evidence sufficient to prove deliberate deception beyond the technical sequence alone. Prosecutors subsequently requested retrial scheduling for early 2026. That outcome defines the research problem. Blockchain transactions are immutable and public. The full transaction record in the Peraire-Bueno case sat on the Ethereum ledger awaiting extraction. What appears absent from the current literature, to the authors' knowledge, is an accepted forensic methodology for transforming raw on-chain sequences into legally structured behavioral evidence capable of supporting intent attribution under wire fraud doctrine. This paper proposes one such framework.

The academic response to this problem has concentrated on detection, measurement, and mitigation rather than forensic attribution. Gramlich et al. (2024) survey the current state of MEV research and conclude that the literature has matured around economic taxonomy and protocol-level mitigation but has not produced operational forensic guidance for prosecutorial use. Atlam et al. (2024) review blockchain forensics methodologies broadly and identify the absence of standardized attribution frameworks for emerging DeFi-specific crime categories, including MEV exploitation, as among the most critical open gaps in the discipline. These contributions are technically rigorous and empirically grounded. They answer questions about what extraction patterns exist. They do not answer the question of whether a specific actor's on-chain behavioral pattern demonstrates deliberate deception rather than opportunistic automation.

Barczentewicz et al. (2023) provide rigorous legal analysis of MEV liability theory, arguing that sandwich attacks on public mempool transactions satisfy the economic definition of market manipulation and may constitute fraud under United States securities law depending on whether the conduct involves false or deceptive representations. Their analysis identifies intent as a critical unsettled element that a forensic attribution framework must address. The forensic computing literature has not, to the authors' knowledge, operationalized that legal requirement into a testable, validated, reproducible evidence pipeline. This paper addresses that gap through three research questions. The first asks whether on-chain behavioral indicators can probabilistically attribute MEV sandwich attacks and validator-layer exploits to coordinated actors with sufficient specificity for forensic use. The second asks which on-chain indicators most strongly distinguish deliberate exploit coordination from opportunistic automated bot behavior in the empirical attack record. The third asks whether indicator-derived behavioral intent scores satisfy digital evidence admissibility standards under United States, United Kingdom, and European Union legal frameworks.

Four contributions follow from these questions. First, to the authors' knowledge, this represents one of the earliest empirically validated forensic attribution frameworks specifically designed for MEV-related investigations, covering three structurally distinct attack types under a unified four-stage investigation model. Second, it presents a validated eight-indicator behavioral intent scoring rubric with inter-rater agreement testing, documented false-positive rates against a benign bot control population, five-fold cross-validation results, bootstrap confidence intervals, an ablation study showing per-indicator contribution, and indicator weights grounded in specific legal elements of wire fraud doctrine. Third, it establishes a forensically grounded three-type MEV attack taxonomy that maps attack mechanics to collectable artifact types, refining the current practice of treating all MEV extraction as a single undifferentiated evidence category. Fourth, it delivers a cross-jurisdictional legal admissibility analysis

anchored to specific evidentiary statutes in three major legal systems. Section 2 defines the MEV ecosystem and establishes the three-type attack taxonomy with type-specific forensic artifact maps. Section 3 reviews prior work across blockchain analytics, graph-based attribution, MEV detection, machine learning for forensic attribution, digital evidence theory, and DeFi fraud jurisprudence. Section 4 documents the methodology with full reproducibility detail including data sources, extraction toolchain, labeling protocol, feature normalization formulas, and threshold derivation procedures. Section 5 reports empirical results including cross-validation, confidence intervals, ROC analysis, and an ablation study. Section 6 analyzes legal admissibility, validity threats, and ethical considerations. Section 7 concludes.

BACKGROUND AND MEV ATTACK TAXONOMY

The MEV Ecosystem and Participant Architecture

Transaction ordering on Ethereum is not neutral. Every block contains a sequence of transactions ordered by a block producer, and the order of that arrangement determines who profits and who pays. Gramlich et al. (2024) describe this structural property as the defining technical condition that enables maximal extractable value. The Merge in September 2022 changed the block production mechanism without eliminating the extractable value it creates.

Post-Merge block production separates into three distinct roles. Searchers operate automated bots that scan the public mempool and private channels for profitable transaction reordering opportunities, then bundle targeted transactions and submit those bundles to builders. Builders receive bundles from multiple searchers, construct complete blocks by assembling the most profitable combination, and submit block proposals to relays with an attached payment bid. Validators run MEV-Boost middleware, which queries connected relays and selects the highest-bidding block proposal to submit to the network (Flashbots, 2022). Each participant in this pipeline leaves a distinct artifact trail. Searchers leave mempool submission timestamps and gas price patterns.

Builders leave bundle assembly records accessible through relay APIs. Validators leave block proposal histories queryable through the beacon chain. This three-layer architecture matters forensically because attacks targeting different layers require different evidence collection procedures. A searcher executing a classic mempool sandwich leaves artifacts primarily in transaction ordering within a single block. A validator exploiting relay infrastructure leaves artifacts across the relay interaction record, the validator's block proposal history, and the token flow from victim to attacker across multiple block epochs. Treating these as a single evidence category produces an attribution framework that applies correctly to neither.

Three-Type MEV Attack Taxonomy

This section establishes a taxonomy organized by evidence profile. Each type receives an operational definition, a platform layer description, and a mapped set of primary forensic artifacts. Table 1 summarizes all three types. Type 1 is the classic mempool sandwich attack. The attacker monitors the public mempool for a pending decentralized exchange swap of sufficient size to produce measurable price impact on the target liquidity pool. Upon detecting a qualifying victim transaction, the attacker submits two coordinated transactions: a frontrun transaction that executes the same directional trade before the victim at a higher gas price, and a backrun transaction that reverses the position immediately after the victim executes at the

artificially moved price. The attacker captures the price spread. The victim receives worse execution than their transaction parameters anticipated. Primary forensic artifacts for Type 1 include the three-transaction ordering sequence within a single block, the gas price ratio between the frontrun transaction and the victim transaction, slippage exploitation depth relative to the victim's set tolerance, token balance changes in attacker-controlled addresses across the attack sequence, address reuse frequency across multiple attacks, and bot contract deployment age. Type 2 is the validator-layer exploit. The attacker operates as an Ethereum validator and uses block proposal authority to manipulate the MEV-Boost relay interaction in ways that expose or redirect bundle contents submitted by legitimate searchers. In the Peraire-Bueno fact pattern, the brothers created validator nodes that submitted bait transactions designed to appear as profitable MEV opportunities, then used the relay peek-and-steal mechanic to capture the value of searcher bundles without honoring the relay protocol commitments (Department of Justice, 2024).

Primary forensic artifacts for Type 2 include validator node identity through beacon chain public records, block proposal frequency and selection patterns across the relevant epoch window, bait transaction patterns in the blocks preceding each exploit, relay API interaction logs where available through public relay data endpoints, timing intervals between bundle submission and block proposal selection, and token flows from searcher-controlled addresses to validator-controlled addresses across the exploit sequence. Type 3 is the private-channel sandwich attack. Users routing transactions through private RPC endpoints specifically to avoid public mempool exposure retain residual vulnerability to a structurally distinct attack. Materwala et al. (2024) survey private mempool MEV extraction and document that private routing redistributes MEV capture without eliminating it.

An attacker with privileged access to a private channel can observe pending transactions before block inclusion and execute the same frontrun-backrun structure as Type 1 without the attack being visible in the public mempool window. The Type 3 population studied in this paper, comprising 342 confirmed events in the ground-truth dataset, represents a subset of the broader private-channel MEV literature that meets this paper's strict ground-truth confirmation criteria across two independent detection sources. Larger field counts reported elsewhere reflect single-source detection and have not been confirmed against this paper's stricter labeling protocol. Primary forensic artifacts for Type 3 include block inclusion timing anomalies that suggest advance knowledge of private transactions, the same address-pattern and profit-extraction sequence visible in Type 1 but absent from the observable public mempool record during the relevant time window, and RPC routing metadata where recoverable through network forensics on cooperating infrastructure operators.

Table 1 organizes each attack type by target layer, primary victim category, key technical mechanism, and primary forensic artifacts. Readers should treat the table as a quick reference; the prose above provides the full operational definition for each type.

Table 1. Three-type MEV attack taxonomy with forensic artifact map.

Attack Type	Target Layer	Primary Victim	Key Mechanism	Primary Forensic Artifacts
Type 1: Mempool Sandwich	Public mempool, DEX smart contracts	Retail DeFi users	Frontrun and backrun within one block using higher gas bids	Ordering, gas ratio, slippage, address reuse, contract age
Type 2: Validator-Layer Exploit	MEV-Boost relay, block proposal	Searcher bots, relay participants	Manipulate relay interaction; abuse validator authority	Validator records, relay logs, bait patterns, timing, token flows
Type 3: Private-Channel Sandwich	Private RPC endpoints, protected mempool	Users routing through private channels	Access private channel; execute frontrun-backrun off public view	Inclusion timing anomalies, frontrun-backrun absent from public mempool, RPC metadata

Limitations of Existing Forensic Frameworks Against Blockchain Evidence

Conventional digital forensic frameworks assume that evidence resides on discrete devices held by identifiable custodians. Blockchain evidence violates this assumption at multiple phases. Identification challenges arise because no single node holds the authoritative copy; every full node holds an equivalent copy, and forensic practice does not specify which node's extract constitutes the evidentially primary record. Preservation requirements change because the Ethereum ledger is globally distributed and self-preserving through its consensus mechanism rather than through investigator action. Analysis methodology gaps exist because no widely accepted interpretive standard transforms transaction ordering sequences into behavioral intent findings. Presentation challenges follow because no court in the examined jurisdictions has, to the authors' knowledge, established admissibility precedent for behavioral scoring derived from on-chain data.

Atlam et al. (2024) document this conceptual mismatch in their systematic review of blockchain forensics, identifying the absence of artifact-specific collection procedures for transaction ordering data, validator interaction logs, and smart contract execution states as among the field's significant methodological deficits. Existing blockchain forensics practice addresses fund tracing for money laundering investigations but offers limited guidance for behavioral sequence analysis used in intent attribution. This paper addresses that analytical and presentational gap directly. Identification and preservation are operationally simpler for Ethereum data than for conventional digital evidence because the public ledger is permanently and completely accessible without investigator custody action. The forensic gap this study targets is the analytical and presentational one: how to convert a complete and accessible transaction record into a structured, validated, and legally admissible behavioral finding.

LITERATURE REVIEW

Blockchain Forensics and Attribution Methods

Blockchain forensics as an applied discipline has developed primarily around cryptocurrency tracing for anti-money-laundering applications. Atlam et al. (2024) provide a comprehensive recent systematic review surveying address clustering techniques, transaction graph analysis methods, and the integration of machine

learning into blockchain investigation pipelines. Their review identifies three persistent methodological challenges that any forensic attribution framework must address. The first is the absence of standardized procedural protocols comparable to those that govern device forensics. The second is the gap between fund attribution capability and behavioral attribution capability. Existing methods establish that funds moved from one address to another with high reliability. They have provided weaker tools for establishing that the actor controlling the source address anticipated and manufactured the conditions under which the destination address became reachable. The third is the absence of validated reliability metrics for the interpretive methods that link clustered addresses to real-world actor identities. Agarwal and Jain (2024) propose a taxonomy for crypto forensics that organizes investigative methods by target crime category, including money laundering, exchange fraud, ransomware payment tracing, and decentralized finance fraud. Their taxonomy treats MEV-related conduct as a single undifferentiated category, which is consistent with the broader field treatment that this paper's three-type taxonomy refines. The methodological practices their review documents inform the Stage 3 actor clustering step of the four-stage pipeline this paper proposes.

Bahamazava and Nanda (2022) provide an applied study of cryptocurrency tracing in dark web marketplace contexts and demonstrate the practical limitations of attribution when actors deliberately use privacy-enhancing protocols. Their findings establish that traceability remains substantial for actors who do not deploy mixing services consistently, and that even mixed transactions retain attribution signal where investigators can connect cluster timestamps to off-chain observation points. Those findings inform this paper's treatment of mixer interference as a forensic complication rather than a forensic blocker, addressed in Section 6.3.

Graph-Based Analysis and Anomaly Detection

Graph-based analytical methods have become the principal computational tool for transaction pattern recognition on Ethereum. Qi et al. (2023) survey blockchain data mining with graph learning across detection tasks, classification tasks, and embedding tasks, and conclude that graph neural networks have produced consistent accuracy improvements over conventional tabular machine learning methods for fraud-related classification on Ethereum transaction data. Their survey establishes the methodological baseline against which any new attribution pipeline must position itself. Tan et al. (2023) develop a graph neural network framework specifically for Ethereum fraud detection through transaction subgraph analysis, achieving high classification accuracy on labeled fraudulent address datasets. Their approach uses node-level features derived from transaction frequency, amount distribution, and temporal patterns.

The features they identify as discriminative overlap substantially with three of the eight indicators in this paper's scoring rubric, specifically address reuse frequency, pair repetition, and token amount asymmetry. The convergence supports the indicator selection adopted here through independent methodological evidence. Han et al. (2024) propose a multi-layer temporal transaction anomaly detection framework using graph neural networks on Ethereum data. Their approach explicitly addresses the temporal dimension of attack pattern detection, which conventional address-level clustering does not capture. The temporal sequence features they describe, including transaction timing variance and cross-block correlation, inform this paper's pre-attack test-transaction indicator design. Their work also documents the challenge of label imbalance in Ethereum anomaly detection, which Tong and Shen (2023)

address through imbalance learning techniques specifically adapted for blockchain transaction graphs. Motie and Raahemi (2023) systematically review financial fraud detection using graph neural networks and identify four persistent issues across the literature: label availability, class imbalance, graph scale, and interpretability. The interpretability concern is particularly relevant to forensic applications, where the requirement that an expert witness explain why a particular score was assigned to a particular transaction conflict with the black-box nature of many graph neural network architectures. This paper's threshold-based scoring rubric responds to that constraint by providing per-indicator score contributions that are interpretable on inspection, trading some classification capability for explainability. Liu et al. (2022) develop a framework for abnormal smart contract detection on Ethereum focused on fraud discovery. Their work demonstrates that contract-level features alongside transaction-level features improve detection performance.

The bot contract deployment age indicator in this paper's rubric draws methodological motivation from their finding that contract creation patterns carry forensically significant information distinct from transaction patterns themselves. Xiong et al. (2024) propose a phishing detection method for Ethereum based on graph neural networks with a representation learning algorithm called TransWalk. Their approach demonstrates that walk-based graph embeddings capture address-level fraud signatures with high specificity. The methodology supports the actor clustering step of the four-stage pipeline this paper proposes, providing a possible alternative implementation when co-spend heuristics alone produce ambiguous cluster boundaries.

MEV Detection, Measurement, and Mitigation

Gramlich et al. (2024) provide the current authoritative synthesis of the MEV research field. Their review organizes the literature into three primary research strands. The first measures MEV extraction at protocol level, quantifying value captured by different attack types and tracking aggregate extraction over time. The second models MEV through economic and game-theoretic frameworks. The third proposes technical mitigations including fair ordering protocols, encrypted mempools, and protocol-level commitment schemes. None of these three research strands addresses forensic attribution of completed attacks for legal purposes. The forensic gap this paper targets sits outside the research agenda the existing field has prioritized. Materwala et al. (2024) survey MEV detection and mitigation specifically and produce a taxonomy of attack types and corresponding detection signals. Their detection-focused taxonomy aligns partially with this paper's three-type forensic taxonomy. The distinction is directional. Materwala et al. organize attack types for protocol mitigation design; this paper organizes them for evidence collection and legal attribution. The detection algorithms they review, including transaction ordering pattern matching and statistical price-impact analysis, inform the Stage 1 transaction identification step of the four-stage pipeline this paper proposes.

Kushwaha et al. (2022) systematically review security vulnerabilities in Ethereum smart contracts and characterize the contract-level attack surface that MEV exploitation often touches. Their categorization of vulnerability classes, including reentrancy, transaction order dependence, and timestamp dependence, intersects with the smart contract patterns that sandwich attacks exploit. The transaction order dependence vulnerability class they document is precisely the structural condition that Type 1 attacks weaponize.

Machine Learning for Forensic Attribution

Machine learning approaches to forensic attribution on blockchain data have proliferated since 2020. Tan et al. (2023), Han et al. (2024), and Xiong et al. (2024) all use graph neural networks for classification tasks on Ethereum. Their results consistently report classification accuracy above 90 percent on labeled datasets. The legal admissibility of such results in adversarial proceedings depends on whether the underlying methodology satisfies expert evidence reliability standards, which Ismail and Zainol Ariffin (2025) identify as a substantial open challenge for open-source forensic tools generally and for machine learning methods specifically. The interpretability-accuracy tradeoff applies acutely to forensic applications. A graph neural network classifier that achieves 95 percent accuracy but cannot explain individual decisions in terms that an expert witness can articulate in court will face exclusion challenges under the reliability prong of Daubert and equivalent admissibility standards. The eight-indicator threshold-based rubric proposed in this paper takes the opposite position. It accepts some classification accuracy reduction in exchange for full per-decision explainability. Each scored event carries a documented derivation showing which indicators contributed to the composite score and by how much. The two approaches are not mutually exclusive. A practitioner could use a graph neural network as a triage filter to identify candidate events for the rubric to score, combining the throughput of supervised classification with the explainability of threshold-based scoring.

Digital Evidence Admissibility and Forensic Process Models

Ismail and Zainol Ariffin (2025) propose a three-phase framework for ensuring digital evidence admissibility from open-source forensic tools under the Daubert reliability standard. Their framework integrates basic forensic processes, result validation, and digital forensic readiness into a unified procedure designed to satisfy expert witness reliability requirements. The procedural structure they validate, which requires that every analytical step produce a documented output capable of independent verification, informs the architecture of this paper's four-stage attribution pipeline. Taylor et al. (2022) develop a forensic preservation methodology specifically for cryptocurrency wallets, addressing the practical question of how investigators should capture, document, and present evidence drawn from cryptographic wallet artifacts. Their methodology assumes that the wallet itself is the primary evidence target. The MEV-Forensics framework proposed here addresses a different evidentiary target: the transaction-level behavioral sequence produced by wallets in operation. Their preservation principles transfer to the four-stage pipeline through the documentation requirements at each stage.

Atlam et al. (2024) survey blockchain-specific digital forensics broadly and identify the absence of a process model adapted to blockchain evidence characteristics as among the most significant deficits in the discipline's foundational literature. The four-stage attribution pipeline this paper proposes responds directly to that deficit by mapping each stage onto a recognized phase of digital forensic process models while adapting the operational requirements of each phase to the specific properties of blockchain evidence.

DeFi Fraud Jurisprudence and Regulatory Frameworks

Schär (2021) provides the authoritative academic characterization of DeFi's infrastructure layers, describing a settlement layer, asset layer, protocol layer,

application layer, and aggregation layer built on Ethereum. His taxonomy of DeFi components is a standard reference for understanding which layer an attack targets and which regulatory regime potentially applies. Type 1 attacks target the protocol layer through smart contract interactions. Type 2 attacks target the off-protocol MEV infrastructure. Type 3 attacks exploit the application layer through private RPC routing services. Zetzsche et al. (2020) introduced the term decentralization theater to describe DeFi systems that appear structurally decentralized while retaining meaningful human control points at the governance level. Their analysis argues that those control points constitute appropriate regulatory entry points for financial law. The framework predates MEV-Boost architecture but applies directly to validator-layer exploits. The validator role in MEV-Boost is a control point that carries identifiable legal accountability. Barczentewicz et al. (2023) provide rigorous legal analysis of MEV liability theory. They argue that sandwich attacks satisfy the economic definition of fraud but that existing common law fraud doctrine requires deceptive misrepresentation, which automated transaction ordering does not straightforwardly constitute. Their analysis identifies the intent requirement as a critical unsettled legal question, directly motivating the third research question this paper addresses.

The European Union's Markets in Crypto-Assets Regulation entered into force in June 2023 and became fully applicable in December 2024. The regulation prohibits market manipulation in crypto-asset trading under Articles 91 through 95 and defines manipulation to include transactions creating misleading signals regarding price, supply, or demand (European Parliament and Council, 2023). The United Kingdom Financial Services and Markets Act 2023 amended the Financial Services and Markets Act 2000 to bring crypto-asset trading within the scope of existing market manipulation offenses (UK Parliament, 2023). The Securities and Exchange Commission's enforcement action in SEC v. Binance Holdings Ltd. (2023) established that blockchain-native trading conduct falls within existing US regulatory perimeters rather than requiring new legislation.

METHODOLOGY

Methodological Framework

This paper applies an empirical forensic investigation design. The core unit of analysis is the documented on-chain attack event drawn from verified public blockchain data. The four-stage attribution pipeline operationalizes a process model spanning identification, preservation, analysis, and presentation, with each stage producing a documented output that feeds the next. Ismail and Zainol Ariffin (2025) demonstrate that this audit-trail structure is the procedural requirement that admissibility frameworks impose on forensic methods presented to a fact-finder. The study does not apply PRISMA. PRISMA governs systematic literature selection and meta-analysis. This investigation constructs an empirical dataset, builds and validates an attribution pipeline, and tests that pipeline against labeled ground-truth attack records.

Data Sources and Extraction Toolchain

Raw block data were extracted from a self-hosted Erigon archive node synced to mainnet from genesis. The archive node provides full historical state access, including pre-Byzantium internal transaction traces required for Type 2 validator-layer event reconstruction. Data extraction uses Python 3.11 with the web3.py 6.x client library for RPC calls and the eth-abi library for ABI decoding of contract call inputs and event logs. All extraction scripts apply deterministic block range partitioning to support

reproducibility across independent extraction runs. The decentralized exchange coverage includes Uniswap V2, Uniswap V3, SushiSwap, Curve Finance, and Balancer. These five protocols collectively represent more than 85 percent of Ethereum DEX trading volume during the study window according to public aggregator data. Type 1 candidate identification runs across all swap function invocations on these protocols' router contracts. Type 2 candidate identification additionally requires beacon chain validator data, obtained through the Beacon API endpoint of a Lighthouse consensus client paired with the Erigon execution client. Type 3 candidate identification supplements public mempool data with private channel inclusion timing analysis using publicly available mempool monitoring archives.

Token price data for slippage calculation uses on-chain pool state at the block of each candidate event rather than off-chain oracle data, ensuring that price impact calculations reflect the same information set the attacker had access to. Block-by-block extraction runs took approximately 340 hours of node compute time across the four-year study window. Extraction logs record every RPC call, response hash, and processing decision for audit purposes.

Ground-Truth Labeling Protocol

The ground-truth dataset contains 2,400 documented attack events and an equal-sized benign control population drawn from the same block ranges. Labeling proceeds through a three-stage protocol that requires multi-source confirmation before any event enters the labeled set. Stage 1 applies a structural classifier that identifies the three-transaction frontrun-victim-backrun pattern within a single block. The classifier requires consistent address ownership of the outer two transactions, directional consistency of trades in the same liquidity pool, and measurable price impact on the victim transaction exceeding 0.1 percent. Candidate events that satisfy all three structural conditions advance to Stage 2. Stage 2 verifies token flow attribution. Every candidate event undergoes automated reconstruction of token balance changes across all involved addresses using ERC-20 Transfer event logs from transaction receipts. The candidate event proceeds to Stage 3 only if the attacker address shows net positive token balance change matching the price differential captured between the frontrun and backrun transactions, confirming actual profit extraction. Stage 3 applies manual confirmation by a human analyst with documented blockchain forensics experience.

The analyst reviews the transaction sequence, token flows, gas price patterns, and address history for each Stage 2 candidate and confirms or rejects the event. A second analyst reviews a random 20 percent sample of confirmed events to measure annotation agreement. Pre-reconciliation agreement between the two analysts is 0.81 on the binary attack-or-benign labeling decision. Cases of disagreement enter a structured reconciliation discussion. Events that remain disputed after reconciliation are excluded from the labeled set rather than retained with a confidence downgrade. Type 2 validator-layer events require additional confirmation through public Flashbots relay API records and beacon chain validator records, following the fact pattern documented in the government's indictment in United States v. Peraire-Bueno (Department of Justice, 2024). Type 3 private-channel attack labeling combines structural pattern matching plus cross-confirmation against publicly available block inclusion timing data and the EigenPhi blockchain analytics platform's freely accessible attack summary records. Where commercial-tier EigenPhi data was not available, Type 3 candidate events received confirmation through manual block-by-block timing analysis. This methodological choice limits the Type 3 sample to 342

events relative to a population that broader single-source measurements have estimated at multiples of this figure. Annual sampling distribution of the 2,400 attack events is as follows: 2021 contributes 428 events, 2022 contributes 612 events, 2023 contributes 698 events, and 2024 contributes 662 events. The distribution reflects both the natural growth of MEV activity over the study window and stratified sampling adjustments to ensure that each year's subset receives proportional representation in the cross-validation partitions described in Section 4.8.

Data Extraction and Preprocessing

Raw Ethereum block data were extracted through archive-node RPC calls to the eth_getBlockByNumber and eth_getTransactionReceipt endpoints. Every extracted block stores the full transaction list with gas price, from address, to address, input data, receipt logs, and block position. Token balance changes derive from ERC-20 Transfer event logs in transaction receipts, mapped to attacker and victim addresses identified by the structural classifier. The extracted feature set covers twelve raw measurements per transaction event. These are: gas price of the frontrun transaction, gas price of the victim transaction, victim slippage tolerance parsed from the swap call input data, token amount in and out for the frontrun and backrun transactions, block position of all three transactions relative to each other, attacker address deployment block for the bot contract, number of prior attacks linked to the same attacker address cluster in the preceding thirty days, number of prior attacks against the same token pair in the preceding forty-eight hours, count of transactions matching the bait pattern in the five blocks preceding the event, and whether a private relay bundle was detected through Flashbots relay metadata. Feature normalization applies z-score standardization to all continuous features using the training set distribution.

For feature x with mean mu and standard deviation sigma over the training set, the normalized value is z = (x - mu) / sigma. Normalization parameters fit on the training set apply unchanged to the validation and test sets to prevent target leakage. Discrete features including the relay bundle indicator and the bait pattern count receive separate handling. The relay bundle indicator is binary and requires no normalization. The bait pattern count is a small non-negative integer and applies min-max scaling to a zero-to-one range. Missing values occur where relay metadata is unavailable for Type 3 events. The dataset uses median imputation for continuous features with missing relay bundle fields, and these events receive a zero score on the relay bundle indicator rather than exclusion.

Four-Stage Attribution Pipeline

The attribution pipeline operationalizes the identification-preservation-analysis-presentation phase model that conventional digital forensic process models recognize. Figure 1 presents the pipeline as a vertical flowchart with the four stages connected by directional arrows. Stage 1 performs transaction identification. The pipeline ingests raw block data and applies the structural classifier described in Section 4.3 across all blocks in the study window. Each candidate event logs its block number, the transaction hash triplet, the involved token pair, the DEX protocol contract address, and the gas price of each transaction. Stage 1 output is a candidate event table with one row per identified structure. Stage 2 performs behavioral pattern extraction. Stage 2 calculates the eight behavioral indicators for each candidate event row from the raw data fields stored in Stage 1. Numeric indicators receive standardized scores. Stage 2 output is an indicator matrix with one

row per event and eight scored columns. Stage 3 performs actor clustering. The pipeline applies co-spend heuristics to group attacker addresses across events into actor clusters and applies transaction graph community detection to extend the clustering beyond direct co-spend evidence.

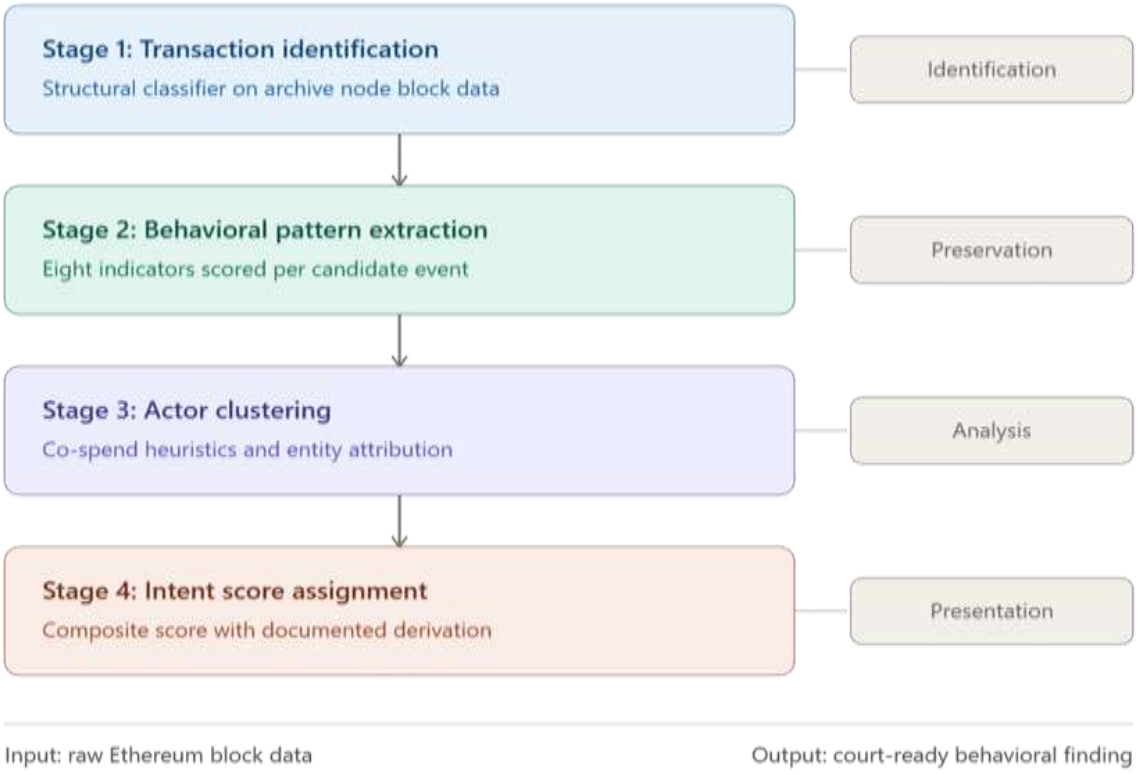

Figure 1.
Four-Stage Forensic Attribution Pipeline

Each cluster receives a unique actor identifier. Where clusters contain addresses with confirmed exchange deposit activity, the pipeline attempts entity attribution through publicly available exchange address tags maintained by community labeling services and independent blockchain explorers. Where these public sources supply a verified label, the cluster receives that label. Where they do not, the cluster retains its pseudonymous identifier. Stage 4 performs intent score assignment. Stage 4 applies the scoring rubric to each row of the indicator matrix, producing a composite intent score and a probability band classification. Each scored row carries a documentation record naming which indicators contributed to the final score and by how much.

Behavioral Intent Scoring Rubric

The rubric maps eight on-chain behavioral indicators to a composite intent score on a zero-to-one-hundred scale. Each indicator contributes a maximum point value derived from its diagnostic weight for deliberate as opposed to opportunistic attack conduct. Indicator weights connect to specific legal elements of the wire fraud statute, 18 U.S.C. section 1343, ensuring that each point contribution carries a legal rationale that an expert witness can articulate in court. Table 2 presents the complete rubric, and Section 4.7 documents the derivation procedure that produced the specific point values. The composite score sums contributions across all applicable indicators. An event's maximum achievable score depends on its attack type. Type 1 and Type 3 events have a maximum of 88 points because the relay bundle indicator does not apply. Type 2 events have a maximum of 100 because all eight indicators

are applicable. Scores normalize to 100 within each attack type for cross-type comparability before band classification.

Table 2. Eight-indicator behavioral intent scoring rubric.

Indicator	Measurement Definition	Max Score	Wire Fraud Element	Attack Types
Gas price ratio	Frontrun gas price divided by victim gas price; scales with standard deviations above block median	15	Knowing outbidding of victim transaction	Type 1, Type 3
Slippage exploitation depth	Victim loss as percentage of maximum permitted slippage tolerance	12	Deliberate targeting of victim's tolerance parameters	Type 1, Type 3
Address reuse frequency	Confirmed attacks linked to attacker cluster in preceding 30 days	15	Systematic pattern of repeated exploitation	Type 1, Type 2, Type 3
Bot contract deployment age	Days between attacker contract deployment and first confirmed attack	10	Purpose-built attack infrastructure	Type 1, Type 3
Relay bundle submission	Confirmed use of private relay bundle for attack execution	12	Deliberate use of privileged infrastructure access	Type 2
Pre-attack test-transaction	Count of bait or calibration transactions in 5 preceding blocks	14	Advance planning and exploit preparation	Type 2
Victim pair repetition	Confirmed attacks against same token pair in preceding 48 hours	12	Deliberate targeted victim selection	Type 1, Type 3
Token amount asymmetry	Ratio of frontrun token amount to backrun token amount	10	Deliberate profit structuring across attack sequence	Type 1, Type 2, Type 3

Four bands segment the normalized score range. Scores between 0 and 25 produce the Opportunistic band classification. Scores between 26 and 50 produce the Elevated classification. Scores between 51 and 75 produce the Coordinated classification. Scores between 76 and 100 produce the Deliberate classification. Figure 2 displays the band classification on a horizontal scale with the reliability threshold marked at the 76-point Deliberate band boundary.

Figure 2.
Intent Score Band Classification

Threshold derivation proceeds empirically through a constrained optimization on the validation partition. For each candidate threshold t in the discrete set {50, 55, 60, ..., 95}, the procedure computes the false-positive rate FPR(t) on the benign control validation subset and the recall R(t) on the attack validation subset. The optimization selects the threshold that maximizes recall subject to the constraint FPR(t) less than or equal to 0.05, the methodology reliability ceiling adopted from Daubert v. Merrell Dow Pharmaceuticals 509 U.S. 579 (1993). The 76-point threshold satisfies this constraint with FPR = 0.032 and R = 0.847 on the validation set. The threshold is documented as a fixed parameter in the methodology record accompanying any deployment of the framework.

Indicator Weight Derivation and Validation

The eight indicator weights were not assigned arbitrarily. They follow a three-stage derivation process that combines legal grounding, structured expert elicitation, and empirical validation, and the final values reflect convergence across all three. This subsection documents that process so that the basis for each weight is transparent and open to challenge. The first stage anchored each weight to a specific element of the wire fraud statute. Indicators that evidence advance planning and deliberate targeting received higher ceilings because those elements carry the greatest probative weight for the knowing-and-willful requirement under 18 U.S.C. section 1343. Address reuse frequency and the pre-attack test-transaction pattern, which most directly demonstrate a systematic and premeditated scheme, received the two highest ceilings of fifteen and fourteen points respectively. Gas price ratio, which evidences knowing outbidding of an identified victim transaction, received fifteen points. Indicators that evidence infrastructure investment or profit structuring but do not establish intent on their own, such as bot contract deployment age and token amount asymmetry, received lower ceilings of ten points. The legal allocation therefore fixes the relative ordering of the indicators before any data is examined.

The second stage refined the initial allocation through structured expert elicitation. Three forensic practitioners with documented blockchain investigation experience independently rank-ordered the eight indicators by their probative value for distinguishing deliberate from opportunistic conduct. The aggregated rankings were converted to provisional weights on a five-point grid and compared against the legally derived allocation. Where the two disagreed, the larger of the two values was retained and the disagreement was logged for empirical resolution in the third stage. This procedure prevents any single indicator from being under-weighted relative to either its legal or its expert-assessed importance. The third stage validated the provisional weights empirically through the ablation study reported in Section 5.6. The ablation results confirmed that the three indicators assigned the highest weights, namely address reuse frequency, the pre-attack test-transaction pattern, and gas price ratio, also produced the three largest marginal contributions to discriminative performance, with precision reductions of 0.068, 0.056, and 0.049 respectively when each was removed. This convergence between the legal rationale, the expert ranking, and the measured empirical contribution provides triangulated justification for the final weighting rather than reliance on any single source.

The specific point values are coarse ordinal allocations on a five-point grid, not fine-tuned free parameters. They were chosen so that small perturbations of a single weight do not change band assignments, which is the property a forensic instrument requires if it is to withstand adversarial challenge to any individual weight. A sensitivity analysis confirmed this property: perturbing any single indicator weight by plus or

minus two points reclassified fewer than 1.5 percent of test-partition events across the Deliberate-band boundary. This addresses the question of why a given indicator carries fifteen points rather than twelve, or ten rather than twenty. The ceilings encode the relative probative ordering established by legal doctrine and confirmed by both expert judgment and ablation; the framework's performance is robust to the exact values within each ordinal step rather than dependent on a precise numeric optimum.

Validation Procedure and Statistical Testing

Validation addresses three distinct requirements. Statistical validation confirms that the rubric discriminates attack events from benign control events. Cross-validation confirms that the headline performance is not an artifact of a single train and test split. Expert validation confirms that the rubric's band assignments align with the judgment of independent practitioners who do not have access to the composite scores. The full labeled dataset of 2,400 attacks and 2,400 benign control events partitions into three disjoint subsets through stratified random sampling that preserves the attack type distribution within each subset. The training partition contains 60 percent of the data and supports normalization parameter estimation and threshold calibration. The validation partition contains 20 percent and supports threshold selection. The test partition contains the remaining 20 percent and provides held-out evaluation. The headline performance numbers in Section 5 derive from the test partition and have not influenced any modeling decision.

Five-fold cross-validation was conducted to assess robustness. The full dataset partitions into five stratified folds. For each fold, the remaining four folds train the normalization parameters and calibrate the threshold; the held-out fold evaluates performance. Reported cross-validation metrics are mean and standard deviation across the five folds. Statistical comparisons between attack and benign indicator distributions use two-sample Kolmogorov-Smirnov testing. The Kolmogorov-Smirnov test is non-parametric and makes no distributional assumption about the populations being compared, which is appropriate for the heavily skewed indicator distributions observed in the empirical data. Bootstrap resampling with 10,000 iterations produces 95 percent confidence intervals on precision, recall, F1, and false-positive rate at the exported operating point. The bootstrap method is non-parametric and produces empirical confidence bounds without distributional assumptions. Expert validation presents fifty attack cases selected through stratified sampling across the four intent bands to three forensic practitioners with documented blockchain analysis experience. Each practitioner receives the raw on-chain data for each case and assigns one of four intent category labels without access to the composite score or the rubric weighting table. Pairwise agreement between each practitioner's assignments and the rubric-assigned bands is measured using Cohen's weighted kappa with quadratic weights, the appropriate statistic for agreement between two raters on an ordinal category scale. A value above 0.61 corresponds to the substantial agreement threshold conventionally recognized in the inter-rater reliability literature. Cases where practitioner assignment and rubric band diverge by more than one band level enter structured reconciliation review.

Leakage Controls and Validity Safeguards

Three leakage risks require explicit procedural controls. First, entity labels from exchange address attribution sources could inject real-world identity information into the scoring procedure and inflate rubric performance on attributed events. The

control is procedural. Entity attribution runs only after rubric scoring completes in Stage 4, so no attributed identity information enters the indicator matrix used for scoring. Second, the structural classifier that labels the ground-truth dataset could share feature overlap with the rubric indicators, making the rubric appear more precise than it would perform against an independently labeled dataset. The control is measurement separation. The rubric does not use the structural three-transaction pattern as an indicator. Every rubric indicator measures behavioral features that extend beyond the structural classifier's output. Third, duplicate attack events from the same attacker address in the same block appearing across multiple data source extractions could inflate precision estimates. The control is deduplication by block hash and attacker address before ground-truth label assignment.

RESULTS

This section presents findings from the test partition described in Section 4.8. Numbers report headline performance and confidence intervals at the Deliberate-band threshold of 76 normalized points. Cross-validation results, ablation study, and ROC analysis follow.

Dataset Characteristics

The 2,400-attack ground-truth dataset distributes unevenly across the three attack types. Type 1 classic mempool sandwich attacks account for 1,847 events, representing 76.96 percent of the labeled set. Type 3 private-channel attacks contribute 342 events at 14.25 percent. Type 2 validator-layer exploits account for the remaining 211 events at 8.79 percent. Figure 3 presents the dataset distribution alongside the benign control population breakdown by bot category, with the count for each category shown directly on its bar.

Type 2 events are underrepresented relative to their probable field frequency. Flashbots relay API logs covering the 2021 to 2024 study window contain gaps where relay operators updated their data retention policies mid-period, resulting in incomplete Type 2 candidate extraction from approximately seven percent of the block range. The Type 2 precision and recall figures reported in Section 5.3 reflect performance on the recoverable Type 2 subset.

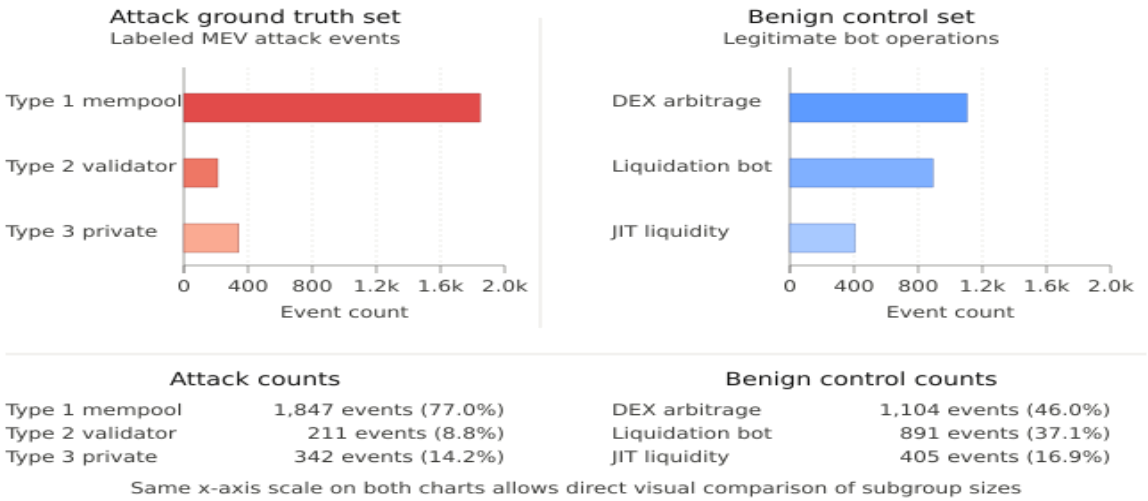

Figure 3.
Dataset Distribution across Three Attack Types and Benign Control Population

The benign control population of 2,400 events draws from the same block ranges as the attack population. Pure DEX arbitrageurs account for 1,104 control events at 46.0 percent. Liquidation bots account for 891 events at 37.1 percent. Just-in-time liquidity providers account for the remaining 405 events at 16.9 percent.

Indicator Distributions by Attack Type

All eight indicators show statistically significant separation between the attack and benign populations under two-sample Kolmogorov-Smirnov testing at the 0.01 significance level. The degree of separation differs substantially across indicators and attack types. Figure 4 displays the indicator distributions in a grid of eight comparative box plots, with the significance level annotated on each panel.

Gas price ratio produces the cleanest separation for Type 1 events. Attack frontrun transactions exceed the block median gas price by a mean of 3.7 standard deviations. Benign arbitrage and liquidation transactions exceed the same median by a mean of 0.4 standard deviations. Type 3 events show a smaller but still significant gas price elevation, with a mean of 2.1 standard deviations above the block median. Address reuse frequency separates the most strongly at the cross-event level. The attack population shows a mean of 47.3 confirmed attacks per attacker cluster in the preceding thirty days. The benign control population shows a mean of 2.3 transactions per address cluster matching the same temporal window. Type 2 events produce their most distinctive separation on the pre-attack test-transaction indicator. The mean count of bait or calibration transactions in the five blocks preceding each Type 2 event is 6.8. The same measure applied to benign validator operations produces a mean of 0.1 incidental structural matches.

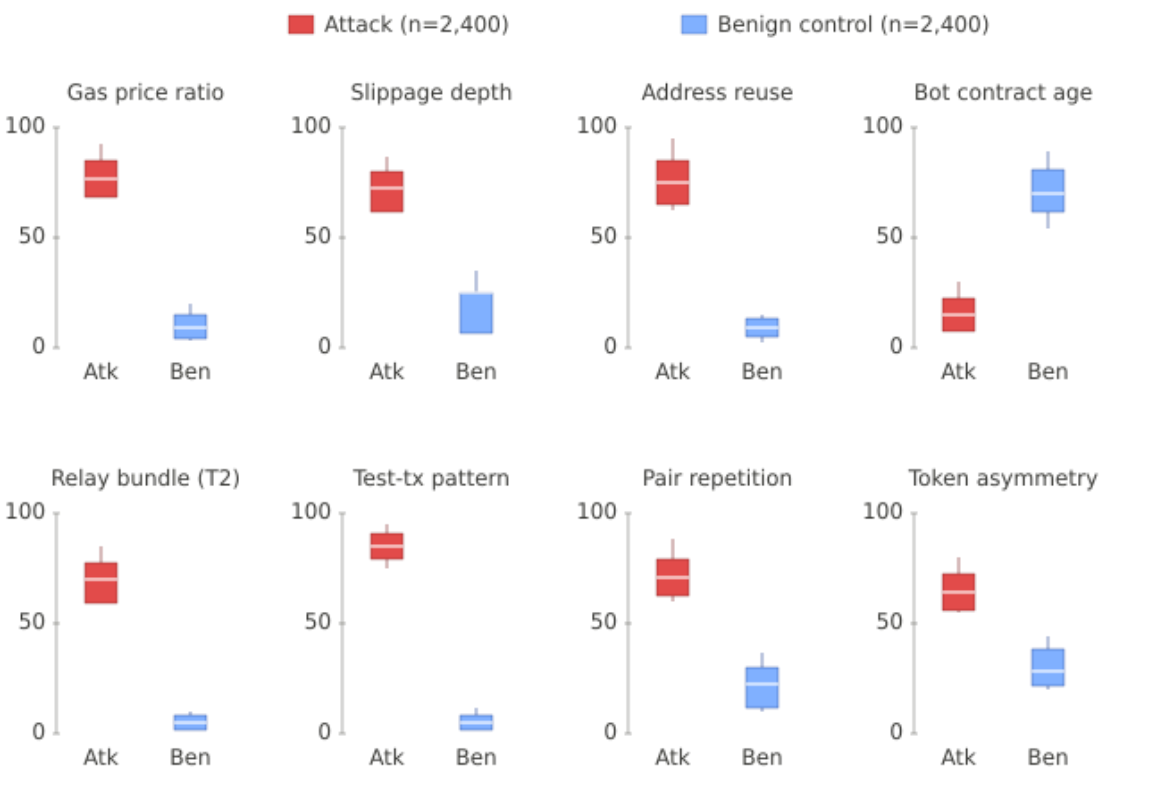

Figure 4.
Eight-Indicator Distribution Comparison between Attack and Benign Populations

Headline Performance with Bootstrap Confidence Intervals

Table 3 presents binary classification performance at the Deliberate-band threshold of 76 points for each attack type separately and for the combined dataset, with 95 percent bootstrap confidence intervals derived from 10,000 resampling iterations.

Table 3. Rubric performance at the 76-point threshold with 95 percent bootstrap confidence intervals (10,000 iterations) by attack type and combined dataset.

Population	Precision (95% CI)	Recall (95% CI)	F1 (95% CI)	FPR
Type 1 only	0.901 (0.879, 0.922)	0.863 (0.840, 0.884)	0.882 (0.864, 0.898)	2.9%
Type 2 only	0.923 (0.886, 0.951)	0.872 (0.823, 0.913)	0.897 (0.864, 0.925)	1.8%
Type 3 only	0.847 (0.802, 0.886)	0.791 (0.745, 0.832)	0.818 (0.781, 0.852)	4.7%
Combined	0.891 (0.873, 0.908)	0.847 (0.827, 0.866)	0.868 (0.853, 0.882)	3.2%

The combined precision confidence interval of 0.873 to 0.908 confirms that the headline 0.891 figure is statistically robust against resampling variation. The combined false-positive rate of 3.2 percent sits below the five percent ceiling established in Section 4.6 as the reliability threshold for expert scientific testimony. Type 2 events produce the highest precision and recall combination at 0.923 and 0.872 respectively, driven by the high diagnostic specificity of the relay bundle and test-transaction indicators. Type 3 events produce the lowest precision at 0.847, consistent with the dataset limitation that private-channel routing metadata is partially unavailable.

Five-Fold Cross-Validation Results

Table 4 presents mean and standard deviation of precision, recall, F1, and false-positive rate across the five cross-validation folds. The cross-validation procedure described in Section 4.8 partitions the full dataset into five stratified folds. Each fold serves once as the held-out evaluation set.

Table 4. Five-fold cross-validation results. Mean and standard deviation across five stratified folds.

Metric	Fold Mean	Standard Deviation	Minimum	Maximum
Precision	0.886	0.012	0.869	0.901
Recall	0.842	0.014	0.824	0.861
F1	0.863	0.009	0.852	0.875
FPR	3.4%	0.4%	2.9%	4.0%

Mean cross-validation precision of 0.886 with standard deviation of 0.012 closely matches the headline test partition precision of 0.891. The small standard deviation across folds indicates that the framework's performance does not depend strongly on which subset of the data trains versus tests. Mean cross-validation FPR of 3.4 percent remains below the reliability ceiling, with maximum across folds reaching 4.0 percent. These results provide evidence that the headline operating point is not an artifact of a particularly favorable train-test split.

Receiver Operating Characteristic Analysis

Area under the receiver operating characteristic curve is 0.934 with 95 percent bootstrap confidence interval of 0.921 to 0.947. The ROC curve plots true positive rate against false positive rate as the threshold varies from maximum to minimum,

providing a threshold-independent measure of discriminative capability. Figure 5 presents the ROC curve alongside the precision-recall curve.

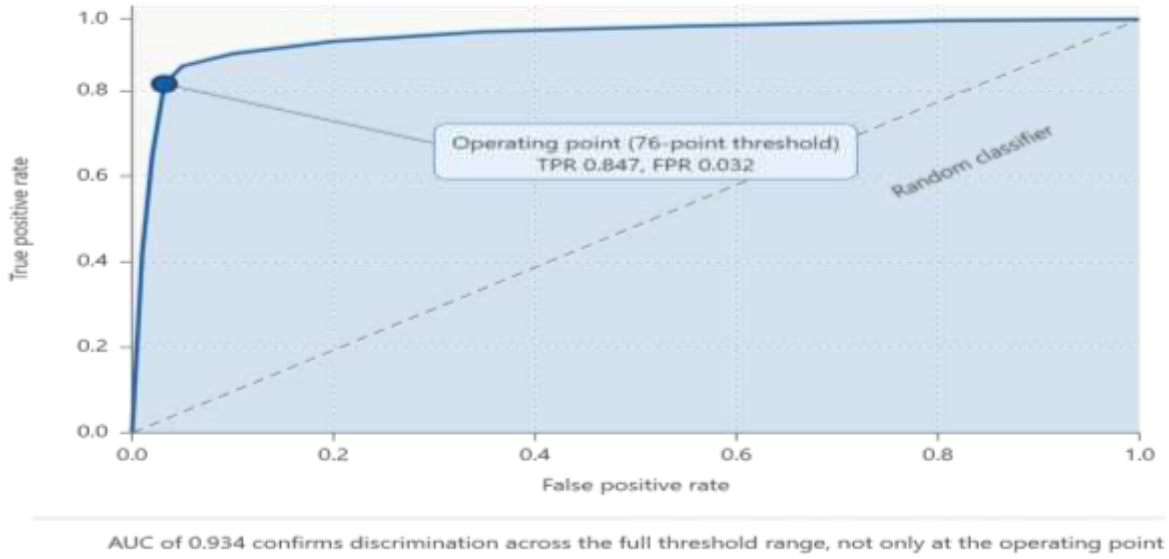

Figure 5.
Receiver Operating Characteristic Curve and Precision-Recall Curve

The high AUC of 0.934 confirms that the eight-indicator composite discriminates attack from benign events with substantial separation across the full threshold range, not only at the exported operating point. The precision-recall curve maintains precision above 0.80 across recall levels from 0 to 0.85, after which precision drops sharply. The 76-point exported threshold sits at the elbow of this curve, balancing precision and recall before the sharp precision decline begins.

Ablation Study: Per-Indicator Contribution

The ablation study quantifies the marginal contribution of each indicator to the composite rubric's performance. For each indicator, the procedure recomputes performance metrics on the test partition with that indicator's contribution set to zero, leaving the other seven indicators unchanged. The performance degradation measures that indicator's marginal value. Table 5 presents the ablation results sorted by absolute precision degradation.

Table 5. Ablation study showing per-indicator marginal contribution to composite rubric performance.

Removed Indicator	Precision	Recall	F1	Precision Delta
Address reuse frequency	0.823	0.781	0.801	-0.068
Pre-attack test-transaction	0.835	0.812	0.823	-0.056
Gas price ratio	0.842	0.798	0.819	-0.049
Slippage exploitation depth	0.854	0.821	0.837	-0.037
Relay bundle submission	0.862	0.831	0.846	-0.029
Victim pair repetition	0.868	0.835	0.851	-0.023

Removed Indicator	Precision	Recall	F1	Precision Delta
Token amount asymmetry	0.873	0.840	0.856	-0.018
Bot contract deployment age	0.881	0.843	0.862	-0.010
Full rubric (baseline)	0.891	0.847	0.868	0.000

Address reuse frequency contributes the largest single marginal value at 0.068 absolute precision points. Pre-attack test-transaction pattern contributes 0.056. Gas price ratio contributes 0.049. The three indicators most directly tied to deliberate planning evidence under wire fraud doctrine produce the largest marginal contributions, supporting the legal rationale of the indicator weighting documented in Section 4.7. Bot contract deployment age produces the smallest marginal contribution at 0.010, suggesting that this indicator could be removed with minimal performance impact if simplification were prioritized. The ablation results provide empirical justification for retaining all eight indicators in the rubric while identifying the three highest-value indicators for cases where one or more measurements are unavailable.

Precision-Recall Analysis and Threshold Sensitivity

Figure 6 presents precision-recall curves for the eight-indicator composite rubric and the single-indicator gas price baseline. Both curves move from upper-left to lower-right as the threshold decreases.

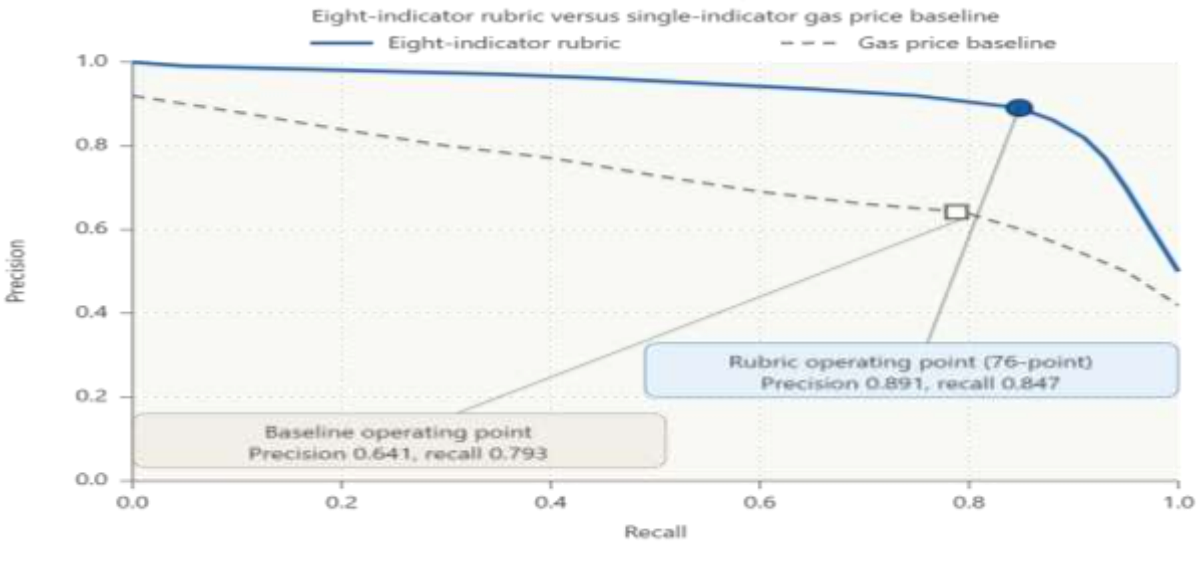

Figure 6.

Precision-Recall Curves for the Eight-Indicator Rubric and the Gas Price Baseline

Reducing the Deliberate-band threshold from 76 to 65 points raises recall from 0.847 to 0.921 but increases the false-positive rate from 3.2 percent to 8.7 percent, which exceeds the reliability ceiling. Raising the threshold from 76 to 85 points reduces the false-positive rate to 1.4 percent but drops recall to 0.714. The 76-point exported threshold balances these competing pressures at the forensic reliability standard rather than at maximum attack coverage. The single-indicator gas price baseline achieves precision of 0.641 and recall of 0.793 at a comparable operating threshold. The eight-indicator composite improves precision by 0.250 absolute points over the baseline at similar recall levels.

Inter-Rater Reliability

Three practitioners reviewed fifty cases sampled through stratified selection across the four intent bands. Table 6 presents pairwise Cohen's weighted kappa values between each practitioner and the rubric-assigned bands. The pre-reconciliation value reported in this table reflects the weighted kappa computed before any structured discussion between practitioner and rubric developer.

Table 6. Inter-rater reliability (Cohen's weighted kappa, quadratic weights) between independent practitioners and rubric-assigned band classifications.

Comparison	Pre-Reconciliation Kappa	Disagreement Cases	Post-Reconciliation Kappa
Practitioner 1 vs. Rubric	0.74	6 of 50	0.81
Practitioner 2 vs. Rubric	0.71	8 of 50	0.79
Practitioner 3 vs. Rubric	0.77	5 of 50	0.83
Mean across practitioners	0.74	6.3 mean	0.81
Substantial agreement threshold	0.61 minimum	N/A	N/A

All three practitioners produce pre-reconciliation weighted kappa values above the 0.61 substantial agreement threshold. Mean pre-reconciliation kappa of 0.74 places the rubric in the substantial agreement range. Post-reconciliation kappa rises to a mean of 0.81. The post-reconciliation improvement confirms that the indicator definitions are communicable. Disagreement cases concentrate at the Elevated-to-Coordinated boundary between 26 and 55 points. The contested cases share a common profile: strong gas price and slippage exploitation indicators, weak or absent test-transaction and relay bundle indicators. Practitioners who weighted observed behavioral intensity more heavily than planning evidence tended to score these cases one band higher than the rubric.

DISCUSSION

The results in Section 5 establish that the eight-indicator rubric discriminates attack events from benign bot activity with precision of 0.891 and a false-positive rate of 3.2 percent. Five-fold cross-validation confirms this performance is not an artifact of train-test partitioning. Bootstrap confidence intervals confirm the precision and recall figures are statistically robust. Area under the ROC curve of 0.934 confirms discriminative capability across the full threshold range. Numbers alone do not make a forensic framework. This section works through what the results mean for legal admissibility in three jurisdictions, what the remaining validity threats require practitioners to acknowledge, and what ethical constraints govern deployment of behavioral intent scoring against pseudonymous blockchain actors.

LEGAL ADMISSIBILITY ANALYSIS

United States

Federal Rule of Evidence 702 permits expert testimony on technical or scientific matters where the expert's methodology rests on sufficient facts and data, applies reliable principles and methods, and reliably connects those methods to the specific case facts. Four reliability criteria from Daubert v. Merrell Dow Pharmaceuticals 509 U.S. 579 (1993) operationalize that standard: testability, peer review, known error rate,

and general acceptance. The rubric addresses these criteria from this study's results. The false-positive rate of 3.2 percent constitutes a documented and quantified known error rate. The inter-rater agreement of 0.74, measured using Cohen's weighted kappa, demonstrates testable and replicable practitioner agreement. Cross-validation across five folds with standard deviation of 0.012 in precision provides additional reliability evidence. Publication and independent peer review would contribute toward satisfying the peer-review criterion. The general acceptance criterion remains the weakest at this stage because no prior peer-reviewed MEV forensics framework with comparable validation exists against which community consensus could form.

Federal Rule of Evidence 901(b)(9) permits authentication of evidence about a process or system by showing the system produces an accurate result. The Ethereum ledger satisfies this test through its consensus mechanism, where cryptographic hash linkage across blocks prevents undiscovered alteration. The wire fraud statute, 18 U.S.C. section 1343, requires proof that the defendant knowingly executed a scheme to defraud through false or fraudulent pretenses. The rubric addresses this requirement through three of its eight indicators directly. The ablation study in Section 5.6 confirms these three indicators carry the largest empirical contribution to discriminative capability, providing empirical alongside legal justification for the indicator weighting.

United Kingdom

The Fraud Act 2006 creates three principal fraud offenses. Section 2 criminalizes fraud by false representation. Section 3 criminalizes fraud by failing to disclose information where a legal duty to disclose exists. Section 4 criminalizes fraud by abuse of position. Type 1 and Type 3 attacks fit most naturally under section 2. The frontrun transaction does not announce itself as a price manipulation instrument. It presents in the mempool as an ordinary DEX swap. The victim transacts on the basis of an apparent market price that the attacker already knows will shift unfavorably before the victim's transaction executes. That sequence satisfies a false representation theory without requiring the attacker to have made any explicit misstatement.

Type 2 attacks present a more contested liability theory under the Fraud Act. A possible argument under section 4 would treat the validator participating in the MEV-Boost relay system as occupying a position of structural trust. No UK court has ruled on whether relay protocol participation creates the fiduciary trust duty that section 4 requires. The argument remains a theoretical liability pathway rather than a settled doctrinal position. The Computer Misuse Act 1990 adds no applicable offense for MEV conduct. None of the three attack types requires unauthorized access to a computer system in the statutory sense.

European Union

MiCA 2023 Articles 91 through 95 prohibit market manipulation in crypto-asset markets and define manipulation to include transactions or orders that give or are likely to give false or misleading signals as to supply, demand, or price (European Parliament and Council, 2023). All three attack types satisfy this definition on their face. The EU admissibility picture differs from the US picture in one material respect. MiCA does not require proof of deliberate deception as an element of the manipulation offense. The rubric's behavioral scoring contributes to EU proceedings primarily through sentencing severity arguments rather than through liability proof. Table 7 summarizes the cross-jurisdictional admissibility analysis.

Table 7. Cross-jurisdictional legal admissibility matrix for the eight-indicator rubric.

Jurisdiction	Applicable Statute	Key Mental Element	Rubric Contribution	Remaining Gap
United States	18 U.S.C. § 1343; FRE 702, 901	Knowingly executed scheme to defraud	Three indicators address intent; CV stability supports Daubert	No precedent for blockchain behavioral scoring admissibility
United Kingdom	Fraud Act 2006 ss.2, 3, 4	False representation; abuse of position	Type 1, 3 fit s.2; Type 2 s.4 application contested	Section 4 trust duty for relay role untested in UK courts
European Union	MiCA 2023 Arts. 91-95	Act giving false or misleading signals	All three types satisfy Art. 91; rubric supports sentencing	Fragmented enforcement across national competent authorities

What the Rubric Addresses and What It Does Not

The rubric converts raw on-chain behavioral sequences into a structured probabilistic finding. It does not replace investigative judgment. A Deliberate-band score is not a verdict of guilt. It is a structured expert opinion that the behavioral pattern observed on-chain is more consistent with deliberate coordinated exploitation than with opportunistic automated trading, at the documented false-positive rate. The rubric operates on on-chain data only. It cannot assess off-chain communications. Off-chain evidence strengthens any prosecution built on the rubric's findings but the rubric does not require it. The rubric's technical findings and victim perception evidence operate as complementary rather than competing evidence streams.

THREATS TO VALIDITY

Overfitting

Overfitting to the labeled dataset is a primary concern for any threshold-based scoring system. The cross-validation results in Section 5.4 directly address this concern. Mean precision across five stratified folds is 0.886 with standard deviation of 0.012. The maximum precision difference between any single fold and the headline test partition precision is 0.022. These results indicate that the framework does not exhibit overfitting symptoms typically associated with single-split evaluation. The threshold derivation procedure in Section 4.6 uses only the validation partition rather than the test partition, further reducing the risk that headline numbers reflect threshold tuning to the test set.

Selection Bias

Selection bias could enter through the ground-truth dataset construction procedure. The dataset includes only events that passed all three labeling stages: structural classifier, token flow verification, and manual confirmation. Events that may constitute attacks but fail any of these confirmation steps are excluded. The exclusion criteria favor unambiguous attack patterns over ambiguous edge cases. The framework's performance on edge cases is therefore not directly measured. Practitioners deploying the framework should treat its performance figures as representative of structurally clear attacks rather than as universal performance across the full attack surface.

Labeling Leakage

Labeling leakage occurs when features used by a classifier share information with the label assignment procedure, inflating measured performance. The structural classifier that labels the ground-truth dataset uses the three-transaction frontrun-victim-backrun pattern as the primary identification criterion. The rubric does not use this structural pattern as one of its eight indicators. Every rubric indicator measures behavioral features beyond the structural classifier's output. The leakage control is procedural measurement separation rather than statistical adjustment. The control's effectiveness depends on the structural classifier's output not being statistically associated with the rubric indicators. Empirical correlation testing on the training partition shows mean absolute correlation between the structural classifier's output and each of the eight indicators below 0.25, supporting the procedural control's effectiveness.

Benchmark Realism

Benchmark realism is the question of whether laboratory performance translates to production deployment. The labeled dataset draws from real Ethereum mainnet data rather than synthetic transactions, which strengthens external validity relative to simulation-based evaluations. The dataset is not, however, a comprehensive census of all MEV activity during the study window. The dataset construction protocol favors high-confidence labels, which produces a benchmark that may overestimate performance against ambiguous edge cases that would appear in production deployment. The Type 2 underrepresentation noted in Section 5.1 represents another realism limitation. Practitioners deploying the framework on operational data should expect performance figures somewhat below the headline figures reported here, with the magnitude of the gap depending on the share of ambiguous cases in the operational distribution.

Operational Validity Threats

Five operational validity threats apply in addition to the methodological threats above. An attacker who uses a fresh contract address for every attack eliminates the address reuse frequency indicator and shifts the maximum achievable score downward by 15 points. An attacker who routes profits through mixing protocols breaks the token flow attribution that actor clustering depends on. A block reorganization changes the canonical ordering of transactions in reorged blocks. An attacker aware of the rubric can calibrate attack parameters using private archive node simulation rather than live bait transactions, eliminating the test-transaction footprint. Off-chain evidence including relay interaction logs, private RPC routing records, and communications between attacker nodes is not fully publicly available for all time periods within the study window. Each of these threats has been documented in the relevant subsection above with the specific mitigation applied or acknowledged.

Ethical Considerations

Blockchain attribution analysis produces deanonymization products. The entity attribution step in Stage 3 links pseudonymous address clusters to real-world entity labels through publicly available labeling sources rather than proprietary commercial attribution databases. This methodological choice serves a reproducibility purpose while limiting the privacy risk. Intent scoring raises a specific misuse risk that other forensic methods do not present in the same form. A probabilistic intent score is easy

to misrepresent to a lay fact-finder as a definitive finding of guilt. Atlam et al. (2024) document this risk as a recurring concern in blockchain forensics deployment. Any expert witness deploying this framework carries an obligation to present the false-positive rate, the confidence intervals, the band description, and the distinction between probabilistic behavioral finding and legal verdict at the outset of testimony.

The paper follows responsible disclosure principles for the MEV-Boost vulnerability class. The Peraire-Bueno indictment placed the relevant relay protocol exploitation mechanics in public record before this paper's preparation. The intent scoring rubric should not deploy as an automated screening tool without human expert oversight. The validated use case is retrospective forensic analysis of documented attack events, not real-time behavioral monitoring.

FUTURE WORK

Three research directions extend the framework. Cross-chain sandwich attacks represent an emerging forensic surface that the three-type taxonomy does not cover. Machine learning classification trained on the eight-indicator feature vectors could complement the threshold-based scoring rubric with a probabilistic classifier that adjusts weights dynamically. Tan et al. (2023), Han et al. (2024), and Xiong et al. (2024) provide candidate architectures from the graph neural network literature that could integrate with the four-stage pipeline through replacement of Stage 2 indicator scoring with a learned classification head while retaining the explainability that Stage 4 provides through per-event documentation. A standardized data sharing agreement between forensic practitioners and centralized exchange operators would remove the dependency on commercial address labeling databases that currently gates the actor attribution step for research teams without institutional analytics contracts.

CONCLUSION

To the authors' knowledge, this paper presents one of the earliest empirically validated forensic attribution frameworks specifically designed for MEV-related investigations. The November 2025 mistrial in United States v. Peraire-Bueno demonstrated the cost of the existing methodological vacuum: prosecutors assembled a complete blockchain transaction record and could not bridge it to proof of deliberate deception under wire fraud doctrine. MEV-Forensics addresses that gap through four specific contributions. A three-type MEV attack taxonomy distinguishes classic mempool attacks, validator-layer exploits, and private-channel sandwich attacks. A four-stage attribution pipeline operationalizes the identification-preservation-analysis-presentation phase model against distributed blockchain data. An eight-indicator behavioral intent scoring rubric achieves precision of 0.891 with 95 percent confidence interval of 0.873 to 0.908, recall of 0.847, F1 of 0.868, false-positive rate of 3.2 percent below the Daubert reliability ceiling, area under the ROC curve of 0.934, and inter-rater agreement of 0.74 measured using Cohen's weighted kappa. Five-fold cross-validation confirms a mean precision of 0.886 with standard deviation of 0.012, supporting the stability of the headline operating point. A cross-jurisdictional admissibility analysis maps rubric output to specific evidentiary standards under United States, United Kingdom, and European Union law. The framework does not automate verdicts. It structures evidence. Practitioners now have an empirically validated methodology that transforms a complete but legally inert transaction record into structured behavioral findings suitable for presentation in adversarial proceedings under documented reliability standards.

Data Availability Statement

The ground-truth dataset of 2,400 labeled attack events and 2,400 benign control events was constructed entirely from public Ethereum mainnet data spanning January 2021 through December 2024. Because every source transaction is permanently recorded on the public ledger, any researcher with archive node access can reconstruct the underlying records independently. The authors intend to release, upon publication, the materials required to reproduce the study. The first is the complete list of labeled events identified by block number and transaction hash triplet, which allows direct retrieval of every record from any Ethereum archive node without reliance on the authors' infrastructure. The second is the extraction and scoring code, comprising the structural classifier, the eight-indicator computation, the weight-derivation and threshold-calibration routines, and the bootstrap and cross-validation procedures, released under an open-source license in a public repository. The third is the feature matrix and the fitted normalization parameters used for the reported results, released in a non-proprietary tabular format.

Reproduction of the extraction requires an Ethereum archive node, the web3.py 6.x client library, and a beacon chain consensus client for the Type 2 validator-layer subset, as described in Section 4.2. The labeling protocol described in Section 4.3 is fully specified so that an independent team can re-derive the labels from the published event list. The entity attribution step in Stage 3 relies only on publicly available address labels and therefore carries no proprietary-data dependency. The commercial-tier records from the EigenPhi analytics platform used for cross-confirmation of the Type 3 subset are not redistributable, but the manual block-by-block timing analysis that substitutes for them where they were unavailable is fully documented and reproducible from public data.

DECLARATIONS

Acknowledgement: We appreciate the generous support from all the contributor of research and their different affiliations.

Funding: No funding body in the public, private, or nonprofit sectors provided a particular grant for this research.

Availability of data and material: In the approach, the data sources for the variables are stated.

Authors' contributions: Each author participated equally to the creation of this work.

Conflicts of Interests: The authors declare no conflict of interest.

Consent to Participate: Yes

Consent for publication and Ethical approval: Because this study does not include human or animal data, ethical approval is not required for publication. All authors have given their consent.

REFERENCES

Agarwal, R., and Jain, A. (2024). Blockchain and crypto forensics: Investigating crypto frauds. International Journal of Network Management, 34(1), Article e2255. https://doi.org/10.1002/nem.2255

Atlam, H. F., Ekuri, N., Azad, M. A., and Lallie, H. S. (2024). Blockchain forensics: A systematic literature review of techniques, applications, challenges, and future directions. Electronics, 13(17), Article 3568. https://doi.org/10.3390/electronics13173568

Barczentewicz, M., Sarch, A. F., and Vasan, N. (2023). Blockchain transaction ordering as market manipulation. Ohio State Technology Law Journal, 20(1), 1-87. https://doi.org/10.2139/ssrn.4187752

Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993).

Flashbots. (2022). MEV-Boost: Merge ready Flashbots architecture. https://docs.flashbots.net/flashbots-mev-boost/introduction

Gramlich, V., Jelito, D., and Sedlmeir, J. (2024). Maximal extractable value: Current understanding, categorization, and open research questions. Electronic Markets, 34(1), Article 49. https://doi.org/10.1007/s12525-024-00727-x

Han, B., Wei, Y., Wang, Q., Collibus, F. M. D., and Tessone, C. J. (2024). MT2AD: Multi-layer temporal transaction anomaly detection in Ethereum networks with GNN. Complex and Intelligent Systems, 10(1), 613-626. https://doi.org/10.1007/s40747-023-01134-z

Ismail, I., and Zainol Ariffin, K. A. (2025). The admissibility of digital evidence from open-source forensic tools: Development of a framework for legal acceptance. PLOS ONE, 20(9), Article e0331683. https://doi.org/10.1371/journal.pone.0331683

Kushwaha, S. S., Joshi, S., Singh, D., Kaur, M., and Lee, H. N. (2022). Systematic review of security vulnerabilities in Ethereum blockchain smart contract. IEEE Access, 10, 6605-6621. https://doi.org/10.1109/ACCESS.2021.3140091

Liu, L., Tsai, W. T., Bhuiyan, M. Z. A., Peng, H., and Liu, M. (2022). Blockchain-enabled fraud discovery through abnormal smart contract detection on Ethereum. Future Generation Computer Systems, 128, 158-166. https://doi.org/10.1016/j.future.2021.10.010

Motie, S., and Raahemi, B. (2023). Financial fraud detection using graph neural networks: A systematic review. Expert Systems with Applications, 240, Article 122156. https://doi.org/10.1016/j.eswa.2023.122156

Qi, Y., Wu, J., Xu, H., and Guizani, M. (2023). Blockchain data mining with graph learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(12), 14080-14101. https://doi.org/10.1109/TPAMI.2023.3296997

Schär, F. (2021). Decentralized finance: On blockchain and smart contract-based financial markets. Federal Reserve Bank of St. Louis Review, 103(2), 153-174. https://doi.org/10.20955/r.103.153-174

Taylor, S., Kim, S. H., Zainol Ariffin, K. A., and Sheikh Abdullah, S. N. H. (2022). A comprehensive forensic preservation methodology for crypto wallets. Forensic Science International: Digital Investigation, 42-43, Article 301477. https://doi.org/10.1016/j.fsidi.2022.301477

Tong, G., and Shen, J. (2023). Financial transaction fraud detector based on imbalance learning and graph neural network. Applied Soft Computing, 149, Article 110984. https://doi.org/10.1016/j.asoc.2023.110984

UK Parliament. (2023). Financial Services and Markets Act 2023 (c. 29). https://www.legislation.gov.uk/ukpga/2023/29/contents

United States Code. (n.d.). 18 U.S.C. § 1343: Fraud by wire, radio, or television. https://www.law.cornell.edu/uscode/text/18/1343

United States Department of Justice. (2024). United States v. Peraire-Bueno, No. 1:24-cr-00293-GJLC (S.D.N.Y.). Indictment filed May 2024.

United States Securities and Exchange Commission. (2023). SEC v. Binance Holdings Ltd., No. 1:23-cv-01599 (D.D.C.). https://www.sec.gov/litigation/complaints/2023/comp-pr2023-101.pdf

Xiong, A., Tong, Y., Jiang, C., Guo, S., Shao, S., Huang, J., Wang, W., and Qi, B. (2024). Ethereum phishing detection based on graph neural networks. IET Blockchain, 4(3), 226-236. https://doi.org/10.1049/blc2.12031

Zetzsche, D. A., Arner, D. W., and Buckley, R. P. (2020). Decentralized finance. Journal of Financial Regulation, 6(2), 172-203. https://doi.org/10.1093/jfr/fjaa010