

Alliance Manchester Business School

BSc Information Technology Management for Business

Final Year Project (BMAN31260)

Operationalising Regulatory Red Flags for Bitcoin Anti-Money Laundering Detection

Constantine Samarikas

Project Tutor: Dr Hossein Jahanshahloo

1 May 2026

Abstract

The fundamental problem for regulators trying to prevent money laundering on Bitcoin is the identification of suspicious activity through qualitative red flags which require context and intuition. While automated detection methods use quantifiable features, they are still held back by the lack of labelled training data. This paper investigates how effectively qualitative indicators used by regulators can be converted into quantitative values used within an automated detection system under a typical Bitcoin anti-money laundering (AML) context. Thirty-five features were derived from the six Financial Action Task Force (FATF) red flags. These features were then deployed in a layered process of rule-based pseudo-labelling, K-Means and DBSCAN clustering, and supervised classification with LightGBM. This system processed over eight million transactions from twenty-four quarterly snapshots from 2019 to 2024, which were then augmented with confirmed labels from four external sources. When tested on 149,128 transactions with confirmed labels, it achieved F1 macro of 0.96 on a balanced evaluation set and 0.93 on the natural confirmed-label distribution, roughly twelve percentage points above the strongest combined rule baseline. Testing using the Elliptic++ data showed that the system matched or exceeded published results. It also correctly tracked changes over time within the Bitcoin AML landscape. However, a scope condition emerged. The system is more effective at detecting obfuscation-heavy laundering rather than illicit activity which does not leave structural patterns, such as gambling or blackmail. The results suggest that contextual regulatory red flags can be operationalised into meaningful detection signal, especially when integrated into a layered architecture, with the rule-based, supervised and unsupervised layers filling in.

Acknowledgements

I would like to thank my supervisor, Dr Hossein Jahanshahloo, for granting access to the Cardiff University Bitcoin Database that made this work possible, and for the guidance and feedback provided throughout the year. I would also like to thank Dr Tahir Syed, the module coordinator for BMAN31260, for the lecture materials and seminar structure that supported the research process.

Table of Contents

Abstract.....	2
Acknowledgements	3
Chapter 1: Introduction.....	7
Chapter 2: Background	8
2.1 Bitcoin, Pseudonymity and the AML Challenge	8
2.2 Regulatory Red Flags	9
2.3 Computational Approaches to Illicit Activity Detection	9
2.4 Limitations of Current Detection Approaches	10
2.5 Research Gap	11
2.6 Research Question and Objectives	12
Chapter 3: Methods	12
3.1 Research Design	12
3.2 System Overview	13
3.3 Data and Unit of Analysis	14
3.4 Feature Engineering and Indicator Operationalisation	15
3.5 Detection Logic	16
Pseudo-labelling logic	16
Unsupervised clustering.....	16
Supervised classification	17
3.6 Handling Mixers and CoinJoin	17
3.7 Evaluation Strategy.....	17

Confirmed-label mechanism validation	17
Temporal validation.....	18
External validation.....	18
3.8 Limitations of the Method	18
Chapter 4: Results	19
4.1 Dataset Overview	19
4.2 Pseudo-label Distribution and Rule Firing	20
4.3 Confirmed-label Enrichment	21
4.4 Mechanism Validation	21
4.5 External Benchmark Validation (Elliptic++).....	24
4.6 Cluster Analysis	25
4.7 Feature Importance	26
4.8 Discussion	27
4.8.1 Operationalising FATF red flags into detection signal	27
4.8.2 Performance under realistic AML conditions	29
4.8.3 Scope, contribution and limitations of the findings	31
Chapter 5: Conclusion & Limitations	33
5.1 Summary of contributions	33
5.2 Reflection and study-level limitations.....	34
5.3 Future work	34
References	36
Ethics Form	38
Appendices	41

Appendix A — Feature Specification	42
Feature Group 1 — Output statistics	42
Feature Group 2 — Address counts and ratios	42
Feature Group 3 — FATF rule signals	42
Feature Group 4 — Wallet context.....	42
Feature Group 5 — Output-address statistics	43
Feature Group 6 — Transaction metadata	43
A.1 Output statistics.....	43
A.2 Address counts and ratios	44
A.3 FATF rule signals.....	45
A.4 Wallet context	45
A.5 Output-address statistics	46
A.6 Transaction metadata.....	47
Appendix B — Per-Period Descriptive Statistics	49
B.1 Transaction volume and value per snapshot.....	49
B.2 Label distribution per snapshot	51
Appendix C — Rule Specification.....	54
Appendix D — Model Configuration and Hyperparameters	56
D.1 Supervised classifier (LightGBM)	56
D.2 Unsupervised clustering.....	57
D.3 Cross-validation protocol.....	57
D.4 Computational environment	58

Chapter 1: Introduction

Bitcoin's transparency through its public ledger, combined with the pseudonymity of its addresses, presents a challenge in preventing money laundering. All transactions are permanently and publicly visible, yet the addresses from which funds are sent and to which they are received are not linked to confirmed identities in the real world (Narayanan et al., 2016; Harlev et al., 2018). In response, regulators have provided qualitative indicators of suspicious activity, and researchers have developed ways of finding suspicious transactions by looking at quantitative transaction features. However, these two approaches haven't been combined. Regulatory indicators are practical to use but are dependent on the specifics of a situation and aren't easily measured. Quantitative methods work well with reliably labelled data but are limited by the scarcity of such examples and the increasing use of mixing services that distort wallet-level interpretation (Financial Action Task Force, 2020; Vlahavas et al., 2024; Arbabi et al., 2025).

This research addresses this gap. It investigates how well regulatory indicators of suspicious Bitcoin activity can be turned into a computational detection framework when used in typical anti-money laundering circumstances. A Python program converts six FATF red flags into thirty-five transaction-level features and then uses rule-based pseudo-labelling with K-Means and DBSCAN clustering and LightGBM supervised classification. The performance of this system is then assessed against confirmed labels from four separate sources. The system was tested on 8.34 million transactions from twenty-four three-month periods between 2019 and 2024, covering three different phases of AML application (the pre-Travel Rule era, the Travel Rule transition, and the post-FTX collapse era of intensified enforcement), and the core algorithm was validated against the Elliptic++ dataset, to ensure it worked independently of the FATF operationalisation.

Figure 1.1 — Conceptual framework

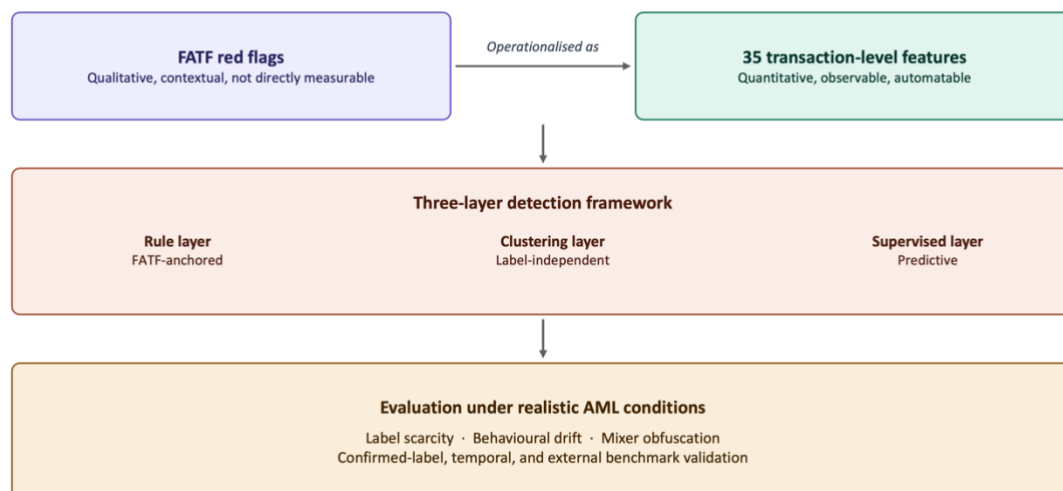


Figure 1.1 — Conceptual framework: from qualitative FATF red flags to a quantitative, layered detection framework evaluated under realistic AML conditions.

This report is organised in the following manner. Chapter 2 reviews scholarly research on detecting Bitcoin-related money laundering, pertinent regulatory signals and the weaknesses of present approaches, culminating in a defined lack of information and the central question this research aims to address. The methodology, comprising the data utilised, the development of the characteristics of the data, the three-stage detection procedure, the treatment of mixing services, and the evaluation of findings, is described in Chapter 3. Chapter 4 presents the results and discusses the findings in relation to the research question, using how the rules operated in each period and per-source analysis to explain what the system can and cannot achieve. Chapter 5 summarises the research's contributions, acknowledges the limitations of the study, and suggests areas for future research.

Chapter 2: Background

2.1 Bitcoin, Pseudonymity and the AML Challenge

Bitcoin has been touted as the future of finance and currency around the globe, through its open ledgers and public transactions (Narayanan et al., 2016). However, this transparency coexists with pseudonymous addressing, as wallet addresses are not inherently linked to

verified real-world identities (Harlev et al., 2018). This paradox of public ledger transparency and pseudonymous addresses makes Bitcoin traceable in structure but difficult in attribution, creating a distinctive anti-money laundering (AML) challenge (Financial Action Task Force, 2020). Although addresses are pseudonymous, transaction patterns recorded on the blockchain may still indicate suspicious activity and support the detection of potentially illicit flows before real-world attribution is established (Harlev et al., 2018). These characteristics have increased the importance of blockchain-based AML monitoring and financial forensic analysis for regulatory and investigative authorities (Financial Action Task Force, 2020).

2.2 Regulatory Red Flags

Regulatory bodies have responded to this monitoring challenge by codifying what suspicious activity looks like in qualitative terms. Compliance frameworks rest on the assumption that behavioural indicators, or red flags, are commonly present throughout suspicious transactions (Financial Action Task Force, 2020). The Financial Action Task Force (FATF) claim these indicators include unusual transaction patterns, concealment behaviour, or interactions with high-risk services (Financial Action Task Force, 2020). These indicators are useful because they reflect practical institutional understanding of what suspicious activity looks like, especially within virtual asset environments (Financial Action Task Force, 2020; Al Naqbi et al., 2025). Many of these indicators are broad and require context, which are not directly measurable as standalone variables or numbers which makes it difficult to create an automatic and standardized process of detection (Financial Action Task Force, 2020). While these red flags are useful for contextual risk identification, they are not automatically transferable into a computational detection system as is.

2.3 Computational Approaches to Illicit Activity Detection

Where regulatory frameworks define suspicion qualitatively, the computational literature has approached the same problem quantitatively, working from the measurable structure of on-chain activity rather than from regulatory indicators (Harlev et al., 2018; Zhang et al., 2020). Harlev et al. demonstrated that the public blockchain ledger already contains enough structural information to infer relationships between addresses, even without verified identity data (Harlev et al., 2018). This reasoning led to the development of heuristic clustering,

where transaction structures are exploited to group addresses that are likely controlled by the same entity (Zhang et al., 2020). This is where the computational literature begins, under the assumption that suspicious behaviour leaves observable structural traces on-chain and can be analysed before attribution (Zhang et al., 2020; Harlev et al., 2018).

The logical next step in automatically identifying issues is to employ machine learning for categorisation and anomaly identification (Harlev et al., 2018). Harlev and colleagues used supervised learning to categorise blockchain entities whose identities are unknown and demonstrated that patterns in address groupings can reveal the nature of the person or organisation using those addresses (Harlev et al., 2018). Weber et al. built on this by utilising graph convolutional networks, trained on a large collection of over 200,000 Bitcoin transactions, to categorise them as illicit or licit (Weber et al., 2019). Vlahavas et al., however, contend that supervised methods are restricted by the limited amount of trustworthy labelled data in blockchain environments, and that unsupervised techniques do not have this issue as they do not require labelled data for their operation (Vlahavas et al., 2024). Furthermore, their strategy differentiates between clustering at the transaction level and at the address level. Transactions are grouped by how they behave, irrespective of the originator, and this reveals unusual activity even when no information about the user is available (Vlahavas et al., 2024).

2.4 Limitations of Current Detection Approaches

These approaches, although effective on particular datasets, all share a fundamental shortcoming: they presume that certain structural characteristics are by their very nature illicit. Indicators such as clustering of transactions, many sources feeding into one destination (fan-in), many destinations from one source (fan-out), repeated sums of money, quick transfers between accounts, and engagement with mixing services are connected to money laundering, yet they are not exclusively used for criminal purposes (Financial Action Task Force, 2020; Arbabi et al., 2025). As the Financial Action Task Force emphasises, evaluating virtual asset environments requires careful consideration of circumstances, moving away from simple matching of rules. A pattern can be unusual without being illegal, and illicit activity can be concealed within normal patterns (Financial Action Task Force, 2020).

A further issue is reliance on categories and imperfect data for confirming outcomes. While structural details of certain transactions allow heuristic clustering to determine relationships between addresses (Zhang et al., 2020), this is based on inference and is not conclusive. This problem is particularly noticeable in supervised detection, where the accuracy of the results is directly tied to the quality of the assigned labels (Weber et al., 2019). Vlahavas et al. point to the shortage of substantial training data and categorised information in this field (Vlahavas et al., 2024); labels are scarce, are not immediately available, or are kept confidential. Consequently, high performance on test datasets does not necessarily mean the same performance during a dynamic, live investigation.

Mixers present the most obvious flaw. Arbabi et al. define mixing services as a way of hiding the transaction path by disconnecting incoming and outgoing funds (Arbabi et al., 2025), and Mariani and Homoliak's review shows mixers employ a range of techniques each affecting traceability differently (Mariani and Homoliak, 2025). The deeper problem is that current techniques assume observed structures map cleanly onto a user, when in practice both the patterns and the user can be deliberately engineered to defeat that interpretation.

2.5 Research Gap

Two strands therefore emerge from the literature. Suspicious activity on blockchains is not hidden and is in fact amenable to analysis, and financial regulations have already identified indicators of money laundering and asset concealment (Financial Action Task Force, 2020; Harlev et al., 2018; Weber et al., 2019; Vlahavas et al., 2024). However, these two areas are not currently integrated with one another. Regulatory indicators, while useful, are rather general, rely on subjective judgement and are specific to their circumstances; computational methods for grouping and following transactions are numerical, successful at what they do, but are limited in their focus and hampered by a lack of sufficient, trustworthy data for development (Financial Action Task Force, 2020; Vlahavas et al., 2024). What remains underdeveloped is the translation between those two domains: how regulatory defined suspicious indicators can be turned into an automatic detection framework.

This gap becomes even more significant under real AML conditions, where actors have the freedom to mix their transactions and actively hide their tracks (Arbabi et al., 2025; Mariani

and Homoliak, 2025; Zola et al., 2025). A dynamic model which can confidently detect illicit transactions within real AML conditions would eliminate the theoretical advantage that criminals have when acting one step ahead of the current computational detection methods. That is the main objective of this project, to develop a model that can confidently label illicit bitcoin transactions within realistic AML conditions, by bridging the gap between qualitative regulatory indicators and quantitative computational indicators.

2.6 Research Question and Objectives

To what extent can regulatory red flags for suspicious Bitcoin activity be operationalised into a computational detection framework under realistic AML conditions?

To answer this question this study will meet the following objectives:

1. Identify regulatory and computational indicators associated with suspicious activity
2. Translate selected qualitative indicators into quantitative features
3. Design and implement a detection framework
4. Evaluate the framework against labelled or known illicit transaction data under realistic conditions

Chapter 3: Methods

3.1 Research Design

This study uses a practical, empirical-computational design focused on building and testing a detection system, not on theorising or eliciting user perspectives. The central question is how far regulatory red flags for suspicious Bitcoin activity can be turned into a working detection framework under realistic AML conditions. This aligns with pragmatic research, where a method is judged by how well it performs against the problem rather than by fit with a single theory (Saunders, Lewis and Thornhill, 2019; Creswell, 2014). The design therefore both builds a deployable pipeline and evaluates it under the label scarcity, obfuscation and behavioural change identified in Chapter 2.

Reasoning is abductive: rules and features begin from FATF guidance, are tested against data, and refined where the two diverge (Dubois and Gadde, 2002). The supervised component is positioned within a weakly supervised learning framework, motivated by the shortage of high-quality illicit-transaction labels in the Bitcoin domain (Ratner et al., 2017; Harlev et al., 2018). Heuristic rules and a small set of confirmed labels jointly produce the hybrid training set used by the classifier; an approach designed for domains where complete ground truth is unavailable. The detailed rationale for combining rule-based, unsupervised and supervised layers is taken up in §3.5.

3.2 System Overview

The system processes original blockchain data to create a suspicion rating, a hybrid label, a grouping for each transaction, and a forecast from the prediction model. Six steps are completed in order: First, the basic data comes from CUBiD. Second, characteristics of each transaction are created. Third, each transaction is checked against four outside labelling services and, if an address matches, a verified label is added. Fourth, the transactions that do not have a verified label are given a 'pseudo' label based on six FATF red flags. Fifth, the combination of verified and 'pseudo' labels (where verified labels take precedence) is then used for both grouping transactions (using K-Means and DBSCAN) and classifying them (using LightGBM). Finally, the system's effectiveness is assessed in three different ways: by comparing results to the externally confirmed labels matched onto CUBiD, by using a chronological order to train and test, and by comparison to the Elliptic++ benchmark. A Streamlit dashboard is available to interactively look at transactions and their individual judgements, yet it isn't used in the functioning fraud detection.

Figure 3.1 — Detection pipeline architecture

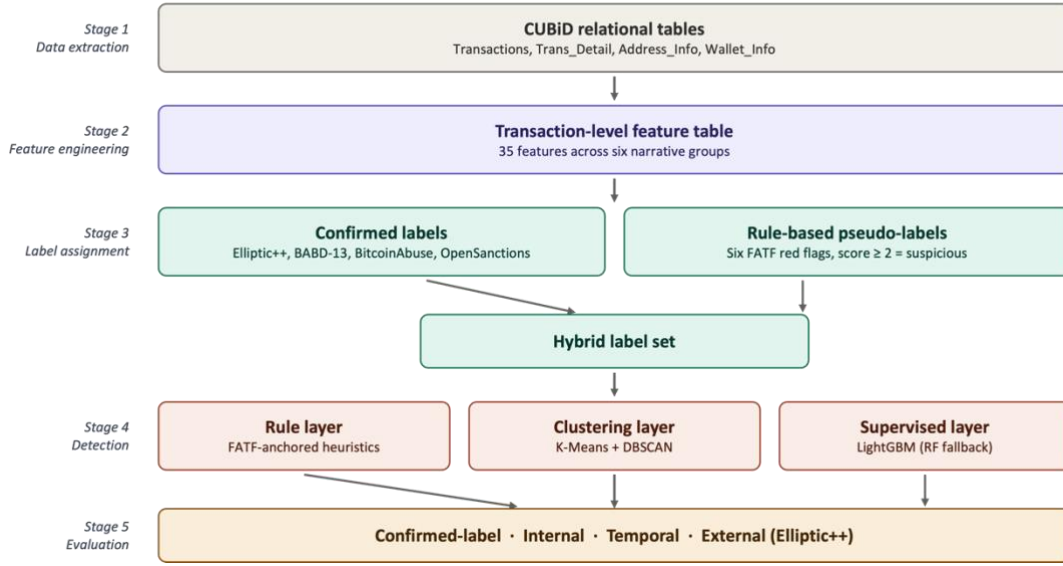


Figure 3.1 — Detection pipeline architecture: data extraction, feature engineering, label assignment (confirmed + rule-based pseudo-labels), three-layer detection (rule, clustering, supervised), and four-strategy evaluation.

3.3 Data and Unit of Analysis

The database utilised for this research is CUBiD (Jahanshahloo, 2023), which organises Bitcoin data in its raw form into relational tables. Four of these tables are used in this study: Transactions and Trans_Detail contain details of transactions and their individual inputs and outputs, and Address_Info and Wallet_Info provide historical information for each address and wallet. Monetary amounts are expressed in Satoshi, where 1 BTC equals 100,000,000 Satoshi (Jahanshahloo, 2023). The basic unit for this analysis is each individual transaction, and it is identified by Block_No and Trans_No. This corresponds to the research question's breadth and permits wallet and address details to provide context for each transaction without assigning a transaction's score to the entire wallet.

CUBiD was chosen instead of publicly available data collections such as Elliptic and BABD-13 for two specific reasons. It offers complete relational coverage of the Bitcoin blockchain at the transaction, address and wallet levels and this is crucial for the cross-layer feature engineering as detailed in §3.4. Most public data sources only show one of these levels; Elliptic is limited to transaction-level characteristics and BABD-13 to address-based activities. Furthermore, CUBiD permits the selection of data from multiple time periods at a

specified block interval, a necessity for the temporal evaluation approach. Elliptic++ and BABD-13 are used as comparison datasets and will be discussed in §3.7.

To facilitate the temporal evaluation strategy described in §3.7, the data is sampled as separate quarterly collections, not as one continuous timeframe. Twenty-four samples were taken from CUBiD representing March, June, September and December of each year from 2019 to 2024. The six-year period was intentionally chosen to encompass three separate Anti-Money Laundering environments: before the introduction of the Travel Rule, the period of the Travel Rule's implementation where CoinJoin and Tornado Cash became widely used, and the period following the collapse of FTX and the increased enforcement that occurred after the 2022 sanctions against Tornado Cash (Al Naqbi et al., 2025). Using the same month for each sample manages any seasonal effects, while the range of time periods assesses how well the results hold true when there are genuinely different regulatory and behavioural circumstances, rather than simply within a uniform period.

3.4 Feature Engineering and Indicator Operationalisation

The feature engineering process produces a single row per transaction with 35 features, organised internally across ten technical categories but presented here under six narrative groups: output statistics, address counts and ratios, FATF rule signals, wallet context, output-address statistics and transaction metadata.

The six FATF rule signals operationalised in Group 3 (equal-value, peeling chain, fresh outputs, high entropy, high fan-out, and CoinJoin strength) were selected from the broader FATF (2020) red flag set on two criteria: each had to map onto an observable on-chain pattern at the transaction level, and each had to be measurable without external attribution data such as identity verification. The choice of transaction-level rather than wallet-level operationalisation follows from the research question, which concerns whether suspicious activity can be detected before real-world attribution is established.

These groupings are based on existing work in computational studies: transaction classification and entity profiling have both used distributions of output values and address counts (Weber et al., 2019; Vlahavas et al., 2024; Harlev et al., 2018), similar structural characteristics are used in heuristic clustering (Zhang et al., 2020), and timing and fee

characteristics are used to identify batch and automated activities (Weber et al., 2019). The complete specification of the features is found in Appendix A.

3.5 Detection Logic

The detection phase utilises three analytical approaches. A system based only on rules is easy to understand and in line with FATF requirements, but it is not adaptable as it cannot identify patterns that are not included in its pre-set logic. A system based only on supervised learning can identify interactions between features but is affected by the biases in the data used to train it, a significant issue in the context of Anti-Money Laundering (AML) for Bitcoin. A system based only on clustering identifies groupings of behaviour without needing labels but does not make a direct prediction. By using all three, the clarity of the rule-based system, the flexible predictive capability of the supervised system and the viewpoint independent of labels in the clustering system are retained, and each system can be tested against the others.

Pseudo-labelling logic

The six FATF rules in Group 3 are implemented as binary flags; each one increases a transaction's suspicion score by one when activated. A score of two or more indicates a suspicious transaction; a score of one is considered unclear and is excluded from training because a single flag could represent normal activity as easily as criminal activity; and a score of zero is considered legitimate. This requirement for a combination of factors is in line with the guidance from FATF, which states that risk arises from a combination of behaviours, not a single indicator (Financial Action Task Force, 2020). The six rules, their thresholds, and the FATF indicator each operationalises are specified in Appendix C.

Unsupervised clustering

K-Means and DBSCAN are applied to the full feature matrix to identify natural behavioural groupings, blind to the rule-based labels. K-Means partitions transactions into a fixed number of groups by feature similarity; DBSCAN identifies dense behavioural regions and flags sparse outliers as noise. Clustering provides a second perspective on the dataset that does not

depend on the rules: a transaction labelled clean by the rules but landing in a heavily flagged cluster raises questions about its legitimacy that the rules alone could not surface.

Supervised classification

A LightGBM model is trained on a hybrid label set, with confirmed labels from external sources used wherever available and rule-derived pseudo-labels used elsewhere; uncertain rows are excluded throughout. The model takes the engineered features as inputs and the hybrid label as the target, evaluated through 5-fold stratified cross-validation. This approach follows precedent in the literature for supervised classification on labelled blockchain transaction data (Weber et al., 2019), with the supervised layer adding the ability to learn non-linear feature interactions that simple thresholds cannot capture.

3.6 Handling Mixers and CoinJoin

CoinJoin transactions merge multiple users' inputs and outputs into a single transaction (Narayanan et al., 2016; Mariani and Homoliak, 2025), making any single wallet's profile unreliable as evidence about the transaction itself. To prevent false classifications, the CoinJoin strength score estimates the likelihood of mixing involvement; when this exceeds 0.1, wallet-level features are suppressed (set to NaN) before training. The CoinJoin score is itself retained as a feature, allowing the model to recognise that mixing is likely involved while ignoring the now-untrustworthy wallet context for that transaction. The threshold of 0.1 is deliberately conservative: the cost of a false wallet association is substantially higher than the cost of losing wallet context for some transactions. This handling addresses one of the realistic AML conditions referenced in the research question; without it, the model would be vulnerable to exactly the obfuscation issues raised in §2.4.

3.7 Evaluation Strategy

Confirmed-label mechanism validation

The core of the framework's testing is its assessment against externally confirmed labels. Every address in the CUBiD data set is examined in relation to four resources: Elliptic++ (Elmougy and Liu, 2023), BABD-13 (Xiang et al., 2024), the BitcoinAbuse public dump, and OpenSanctions. A transaction will be assigned the label from whichever of these sources has

a corresponding incoming or outgoing address. BABD-13 and Elliptic++ both offer labels for both illicit and licit categories, whereas BitcoinAbuse and OpenSanctions only identify illicit activity, and this biases the inherent balance of labels to emphasise illicit transactions. Two evaluations are presented. The first, termed 'natural', uses the labels obtained as described. The second, 'augmented', randomly selects from the CUBiD transactions which have no matches and presumes these are legitimate, following the general understanding that under one percent of Bitcoin transactions are for illicit purposes. Each evaluation uses five-fold stratified cross-validation, alongside per-rule precision and recall, a per-source breakdown, and a threshold sensitivity sweep. Importantly, these external resources are used only to assess the system and are not needed by the deployable pipeline itself.

Temporal validation

In order to determine how well the system performs when criminal activity changes, the data is divided by time into the twenty-four three-month periods detailed in §3.3. The model learns from the first sixteen of these periods, from March 2019 to December 2022, and is then assessed on the last eight, March 2023 to December 2024. This simulates the situation of an active anti-money laundering system encountering previously unknown data, while covering several considerably different periods in the history of crypto anti-money laundering.

External validation

To provide an independent 'ground truth' for benchmarking the algorithmic methodology, the Elliptic++ dataset (Elmougy and Liu, 2023) is employed. A similar model, possessing the same architecture, is both trained and evaluated using Elliptic++'s data and labels. The outcomes of this are then compared against the results disclosed by Elmougy and Liu (2023), verifying the algorithmic backbone against a separately labelled data set and a public measure of success, while specifically excluding Elliptic++ data from the CUBiD process.

3.8 Limitations of the Method

The framework has three remaining limitations.

Firstly, despite the addition of confirmed labels from external sources as detailed in §3.7, rule-based pseudo-labelling remains the primary source of training data. As external sources

only cover a portion of CUBiD transactions, the majority of unconfirmed transactions rely on rule-derived pseudo-labels, meaning the supervised model ultimately replicates the logic of the rules used to label them. While the Elliptic++ benchmark provides a separate, non-circular verification of the algorithmic approach, its features are not the same as the Financial Action Task Force (FATF)-defined operational feature set; therefore, it verifies the algorithm itself rather than the specific implementation central to the research question. Secondly, the thresholds for the rules are set based on judgement informed by FATF guidance, not on optimisation against confirmed labels. They are calibrated to find a balance between incorrectly flagging transactions and missing detections but have not been adjusted. Thirdly, the chronological training and testing split only assesses changes over time within the sample period. Longer-term changes are not covered; criminal activity continues to change in response to detection, regulation and changes within the Bitcoin system as a whole, and a framework incorporating ongoing retraining would be needed for continued operational use.

Chapter 4: Results

4.1 Dataset Overview

The active pipeline handles 8,343,135 transactions from twenty-four quarterly snapshots of CUBiD, covering the period from March 2019 to December 2024. Each of these snapshots consist of a single day from each year, March 8, June 8, September 8, December 8, and generates approximately 152 blocks. Within the dataset are 3,642 distinct blocks, 12.6 million unique addresses serving as inputs, 13.6 million unique addresses used as outputs, and approximately 7.0 million unique wallets. Total transaction volume is around 44.6 million BTC.

Transaction numbers for each representation are illustrated in Figure 4.1. Numbers were relatively consistent at between 250,000 and 330,000 per representation in the earlier years but increased considerably during 2023 and 2024. Specifically, the Q3 2024 representation contains 910,083 transactions, or more than three times the amount from 2019.

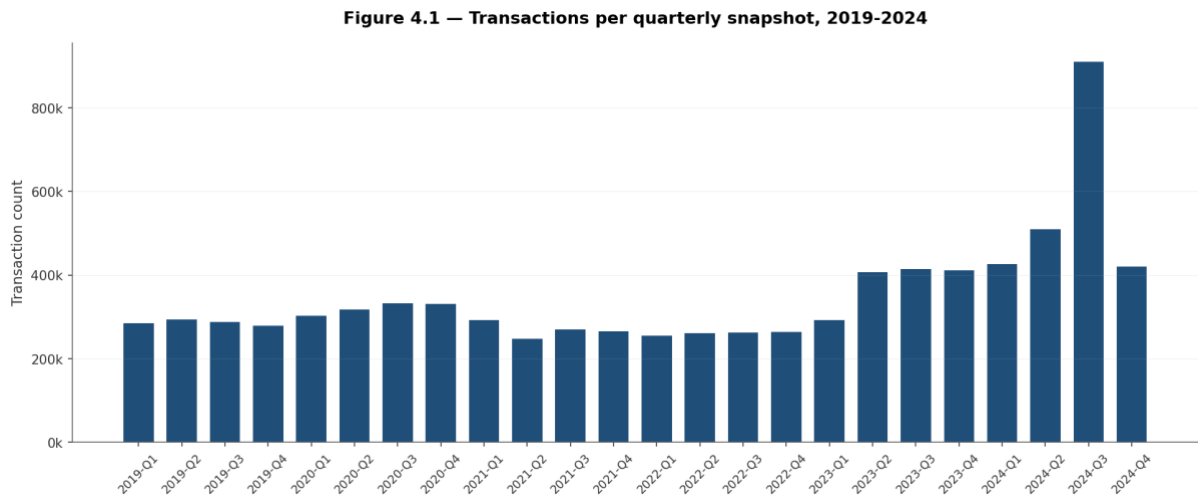


Figure 4.1 — Transactions per quarterly snapshot, 2019–2024.

4.2 Pseudo-label Distribution and Rule Firing

Each transaction receives a label through the hybrid scheme in §3.7: confirmed where any input or output matches an external source, pseudo-labelled by FATF rule scores otherwise. Confirmed labels cover 1.78% of the 8,343,135 transactions (112,861 illicit and 36,267 licit), pseudo-labels cover 58.95% (2,162,043 illicit and 2,762,049 licit), and the remaining 39.19% receive an ambiguous score of one and are excluded from supervised training. The full breakdown is in Appendix B.

The six FATF rules fire at the following rates across the dataset: rule_peeling 40.06%, rule_equal_value and rule_mixing each 24.94%, rule_high_fan_out 1.93%, rule_high_entropy 0.64%, and rule_fresh_outputs 0.00%. Two observations warrant comment. First, rule_fresh_outputs fires zero times: CUBiD's Address_Info table operates as a global registry rather than a snapshot-restricted index, so every output address has prior history relative to its block by construction. This is a structural property of the data, not a flaw in the rule's design. Second, the aggregated firing rates conceal a substantial temporal trajectory visible in Figure 4.2.

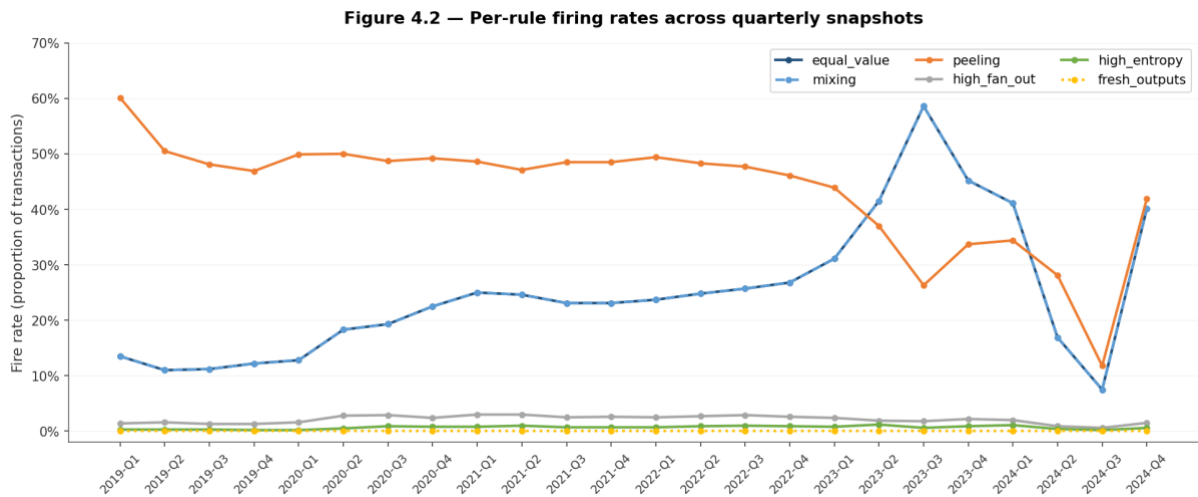


Figure 4.2 — Per-rule firing rates across quarterly snapshots, 2019–2024.

The `equal_value` and `mixing` rules rise from 13.5% in 2019-Q1 to a peak of 58.6% in 2023-Q3 before collapsing to 7.5% in 2024-Q3. The `peeling` rule shows the inverse pattern, beginning at 60.1% in 2019-Q1 and reaching its low of 11.8% in 2024-Q3. This evolution maps onto well-documented changes in the Bitcoin AML landscape, the implications of which are taken up in §4.8.

4.3 Confirmed-label Enrichment

Cross-referencing CUBiD addresses against the four external sources produced 149,128 transactions with externally verified labels: 112,861 confirmed illicit and 36,267 confirmed licit. BABD-13 contributes the largest matched volume (57,236 transactions, both classes), roughly three times the Elliptic++ count (19,828, both classes), reflecting both broader address coverage and closer temporal overlap with the 2019–2024 sample. BitcoinAbuse contributes 94,868 illicit-only matches and OpenSanctions 1,332 illicit-only, which precludes their use as standalone subsets in supervised cross-validation but allows their inclusion in pooled evaluation and rule validation. The full source breakdown is reported in Appendix B.

4.4 Mechanism Validation

The supervised classifier was evaluated against the confirmed-label subset using 5-fold stratified cross-validation under two configurations: a natural evaluation using the confirmed-label distribution as observed, and an augmented evaluation that supplements the confirmed-

licit pool with a random sample of unmatched CUBiD transactions presumed licit, producing a balanced 50/50 set. Table 4.1 reports both.

Metric	Augmented (50/50)	Natural (confirmed-only)
F1 macro	0.9617 $\pm$ 0.0008	0.9276 $\pm$ 0.0009
F1 weighted	0.9617	0.9451
Precision (illicit)	0.9753	0.9915
Recall (illicit)	0.9475	0.9334
Precision (licit)	0.9489	0.8247
Recall (licit)	0.9760	0.9752
ROC-AUC	0.9942	0.9906
PR-AUC	0.9949	0.9971
n samples	225,722	149,128

Table 4.1 — Mechanism validation: 5-fold stratified CV metrics on the augmented and natural evaluation sets.

Both configurations achieve F1 above 0.92 with ROC-AUC and PR-AUC values above 0.99, indicating that the FATF-operationalised feature set produces strong discrimination between confirmed illicit and licit transactions. The 3-point gap between augmented (0.96) and natural (0.93) F1 reflects the difficulty added by the natural distribution's class imbalance: with confirmed licit transactions outnumbered roughly three to one by confirmed illicit, the model has fewer licit examples to learn from, which shows up clearly in the asymmetry between illicit precision (0.99) and licit precision (0.82) on the natural set. Recall is preserved on both classes (0.93 illicit, 0.98 licit), so the model is not missing illicit activity to inflate precision; the cost is borne by a higher rate of false positives on the smaller licit class. For an AML use

case this is the more tolerable error mode, since a flagged-but-licit transaction goes for human review while a missed illicit transaction passes through unflagged. Figure 4.3 visualises the corresponding confusion matrices

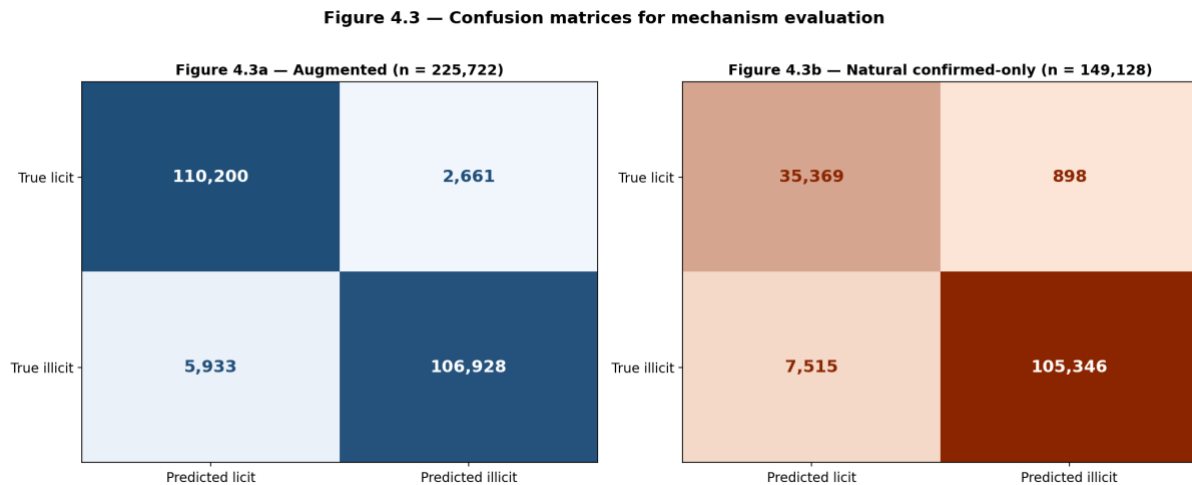


Figure 4.3 — Out-of-fold confusion matrices for the augmented and natural evaluation sets.

To compare the rule layer against the supervised model, each FATF rule was evaluated as a standalone classifier against the augmented set. Table 4.2 reports per-rule precision and recall alongside the supervised baseline and an any-rule combined baseline.

Classifier	Precision	Recall	F1
rule_peeling	0.9127	0.6019	0.7254
rule_high_entropy	0.8378	0.0541	0.1016
rule_high_fan_out	0.8004	0.1567	0.2620
rule_equal_value	0.7249	0.0782	0.1411
rule_mixing	0.7249	0.0782	0.1411
rule_fresh_outputs	—	—	Rule never fires

Classifier	Precision	Recall	F1
ANY rule fired (combined)	0.875	0.808	0.8403
Supervised LightGBM (augmented)	0.9753	0.9475	0.9617

Table 4.2 — Per-rule and supervised-model performance against the augmented confirmed-label set.

The peeling chain rule is the strongest individual indicator (F1 0.7254). The any-rule combination raises recall to 0.808 at precision 0.875, producing F1 of 0.8403. The supervised model raises F1 to 0.9617, a gap of roughly twelve percentage points over the strongest combined rule baseline, with the substantive interpretation discussed in §4.8.

Cross-validation was also run separately for each label source where both classes are present, to assess whether performance varies by category of confirmed illicit activity. Elliptic++ alone (n=19,828) achieves F1 macro 0.9747 with ROC-AUC 0.9944, while BABD-13 alone (n=57,236) achieves F1 macro 0.9101 with ROC-AUC 0.9767. BitcoinAbuse and OpenSanctions are excluded as single-class subsets. The Elliptic++-only result substantially exceeds BABD-13's despite identical training and CV protocols. The interpretation of this gap is taken up in §4.8. Full per-source breakdown in Appendix B.

4.5 External Benchmark Validation (Elliptic++)

To benchmark the algorithmic methodology against independently labelled data, the same supervised architecture was trained and evaluated on Elliptic++'s own features and labels. Table 4.3 reports the results alongside the published baselines from Elmougy and Liu (2023).

Level	Metric	This study (LightGBM)	Elmougy & Liu (RF)	Δ
Transaction-level	Precision	0.9909	0.986	+0.005
Transaction-level	Recall	0.9682	0.727	+0.241
Transaction-level	F1 macro	0.9792	0.836	+0.143
Actor-level	Precision	0.9114	0.921	-0.009
Actor-level	Recall	0.9798	0.802	+0.178
Actor-level	F1 macro	0.9423	0.858	+0.084

Table 4.3 — *Elliptic++ benchmark, this study compared with Elmougy and Liu (2023).*

The supervised architecture matches the published precision (0.99 vs 0.99 transaction-level) while substantially improving recall (0.97 vs 0.73), producing F1 differentials of +0.14 and +0.08 at the transaction and actor levels respectively.

4.6 Cluster Analysis

K-Means clustering with $k = 3$ was applied to the full feature matrix to identify behavioural groupings independent of the labelling scheme. Cluster 1 is the most informative: 855,945 transactions, 44,509 confirmed illicit and only 221 confirmed licit, producing 99.5% illicit purity. Cluster 2 contains 747 transactions but no confirmed labels. Cluster 0 absorbs the remaining 7.5 million transactions along with the residual 65.5% illicit purity that remains after cluster 1 has separated out a large block of obfuscation-typical illicit activity.

The recovery rate is striking: 44,509 of the 112,861 confirmed illicit transactions in the dataset, roughly 39%, fall into a single cluster discovered without any access to labels during

clustering. Two implications follow. First, the FATF feature space carries enough discriminative signal that obfuscation-typical activity self-organises into a dedicated region without supervision, which speaks to the validity of the operationalisation independently of the labels. Second, this is a partial floor on what the framework can detect at near-zero false-positive cost: cluster 1's 99.5% illicit purity means a transaction landing in this cluster is almost never licit, so even before the supervised classifier is applied, an investigator looking only at unsupervised cluster membership would catch four in every ten confirmed illicit transactions with very few false alarms. The remaining illicit volume in cluster 0 reflects the variety of illicit behaviour that does not present obfuscation signatures strongly enough to separate without supervision. DBSCAN at $\text{eps} = 1.5$, $\text{min_samples} = 5$ produced 42,162 microclusters with 85.9% noise and a largest cluster of 6,267. The configured parameters are too tight for meaningful density-based clustering at this dataset scale.

4.7 Feature Importance

Feature importance from the trained LightGBM model is inspected through both number of splits (how often each feature is used as a split point) and gain (predictive lift contributed when used).

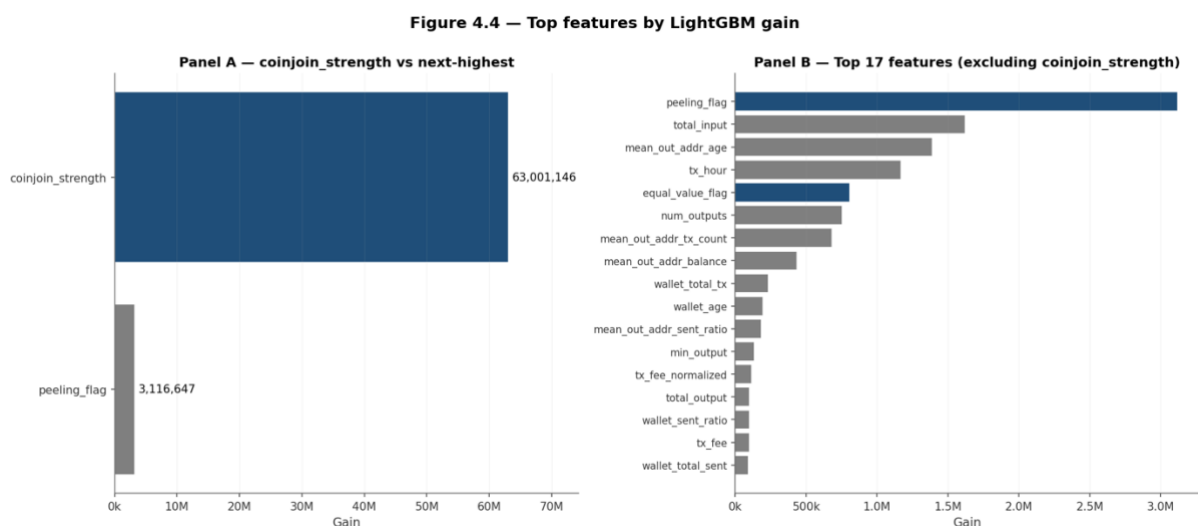


Figure 4.4 — Top features by LightGBM gain. Panel A illustrates the dominance of `coinjoin_strength` relative to the next-highest feature; Panel B shows the next 17 features at readable scale.

Aggregating these by feature group reveals a striking division of labour: the FATF rule signals group accounts for only 6.3% of model splits but contributes 94.4% of total gain,

driven almost entirely by `coinjoin_strength`. By contrast, the output-address statistics group accounts for 35.1% of splits but only 3.0% of gain. The other groups such as, output statistics (16.6% splits / 0.29% gain), address counts (13.7% / 1.7%), transaction metadata (12.7% / 0.23%), wallet context (9.0% / 0.31%), wallet temporal (2.9% / 0.04%), input statistics (2.1% / 0.03%), and graph topology (1.6% / 0.02%), each contribute frequent splits with only modest gain.

The pattern is not subtle. A small number of FATF rule features are used rarely but, when they are used, they carry almost all of the model's predictive weight. The remaining features are used many times more often, but each split contributes only a fractional improvement. This means the model is not learning structural patterns from scratch and treating the rules as one input among many; it is leaning heavily on the rule signals when they fire and using the structural features mainly to refine predictions in the cases where the rules do not fire decisively. The substantive interpretation of this division of labour, and what it implies for the relationship between regulatory red flags and computational detection, is taken up in §4.8. The full per-group breakdown is in Appendix D.

4.8 Discussion

4.8.1 Operationalising FATF red flags into detection signal

Table 4.1 illustrates that evaluating the headline mechanism achieved a macro F1 score of 0.9617 when tested against the expanded dataset and 0.9276 using only the natural confirmed-only set. A Receiver Operating Characteristic Area Under the Curve (ROC-AUC) of over 0.99 was obtained for each of these scenarios. Consequently, FATF red flags can be operationalised into computational detection signal that holds up against externally confirmed labels. Furthermore, the results for Elliptic++ in Table 4.3 (a macro F1 score of 0.9792 at the transaction level, 0.9423 at the actor level) both broaden this support to a dataset with independent labels and provide a basis for comparison with an established academic comparison point. Figure 4.5 displays these figures alongside the published baseline

measurements that were examined in Chapter 2

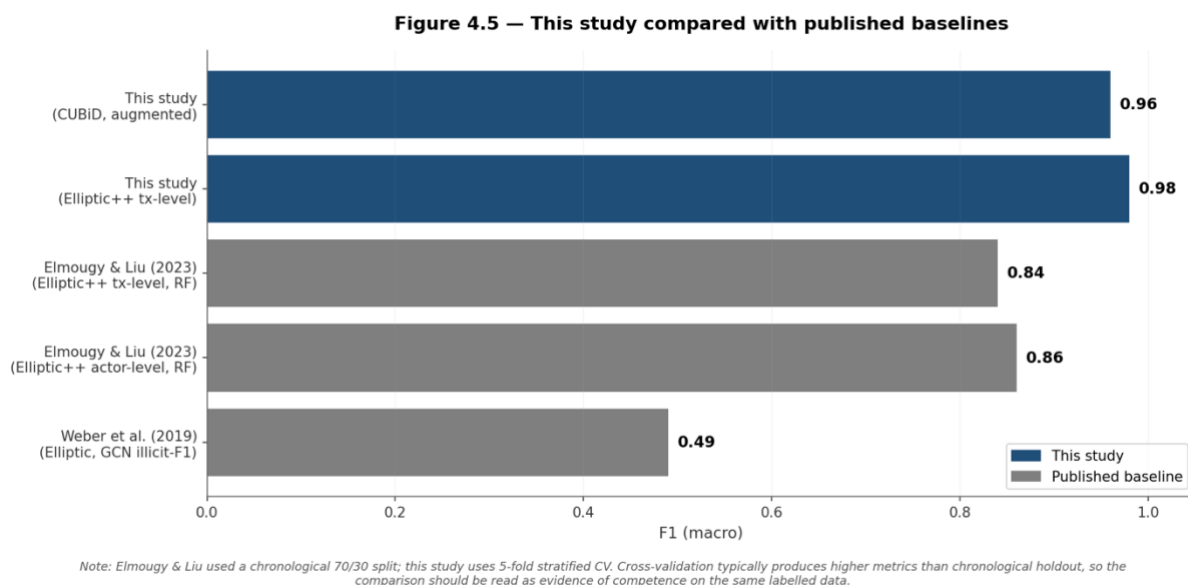


Figure 4.5 — This study's F1 (macro) compared with published baselines from Elmougy and Liu (2023) and Weber et al. (2019).

The comparison with Elmougy and Liu (2023) must be taken with the consideration that this study uses 5-fold stratified cross-validation, while their reported numbers come from a chronological 70/30 split that typically produces lower metrics. The result is therefore evidence of competence on the same labelled data rather than a strict head-to-head improvement. Even with that consideration, two observations follow. First, the framework's algorithmic backbone performs at a level consistent with the supervised classification literature discussed in §2.3 (Harlev et al., 2018; Weber et al., 2019). Second, the Elliptic++ result implies that this competence does not depend on the FATF-operationalised feature set in particular, the architecture also performs on Elliptic++'s own features, which is important for interpreting what the FATF operationalisation specifically contributes.

That contribution is more directly visible in the rules-vs-supervised comparison reported in Table 4.2. The best-performing single rule (peeling chain) reaches F1 of 0.7254, and the any-rule combination reaches 0.8403, while the supervised model trained on the same features lifts F1 to 0.9617, a gap of roughly twelve percentage points over the strongest combined rule baseline. The supervised layer therefore does substantial work; rules applied as standalone classifiers fall short of the integrated framework. Yet feature importance shows that the same

rules are doing most of the predictive work inside the model. As Figure 4.6 illustrates, the FATF rule signals group accounts for only 6.3% of LightGBM splits but 94.4% of total gain, driven almost entirely by `coinjoin_strength`.

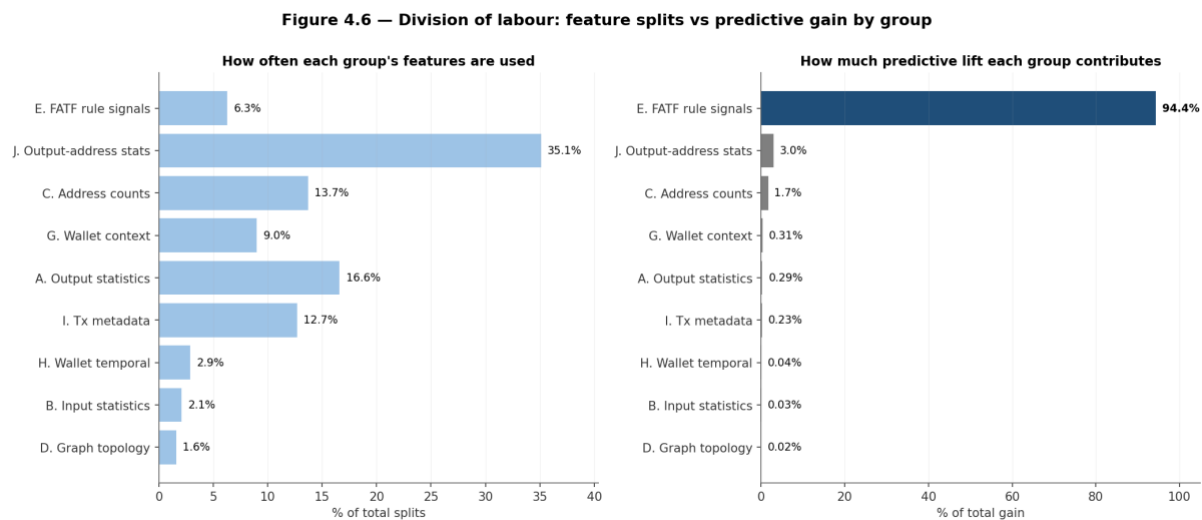


Figure 4.6 — Division of labour: how often each feature group's features are used by the model versus how much predictive lift each group contributes.

Figure 4.6 allows for two understandings. The FATF features behave as decisive signals: when a transaction activates a strong FATF indicator, the model leans heavily on that signal to make its prediction. Further, structural and contextual elements, specifically, statistics for the destination address, the number of addresses and the wallet's background, are used often, though each one provides only a small improvement to the prediction. Their purpose is to improve predictions when the FATF signals are weak or missing. This is in line with FATF recommendations that risk arises from a combination of behaviours, not one indicator in isolation (Financial Action Task Force, 2020), and it directly responds to the research gap outlined in §2.5. This gap concerns the relationship between qualitative, regulator-defined indicators and quantitative, computational detection; that relationship isn't a simple 'yes or no', but a layered system. Within this system, rule signals carry decisive weight when they apply, and structural characteristics come into play where they don't.

4.8.2 Performance under realistic AML conditions

Operating effectively against confirmed data is a requirement but is not enough on its own. The research question is specifically about performance in real-world AML situations, where

criminal activity changes, mixers conceal the source of funds and classifications of transactions are delayed in relation to the activity itself.

First, the per-period rule firing trajectory detailed in §4.2 closely mirrors established changes in the Bitcoin AML landscape. Figure 4.7 shows this pattern alongside significant legal enforcement.

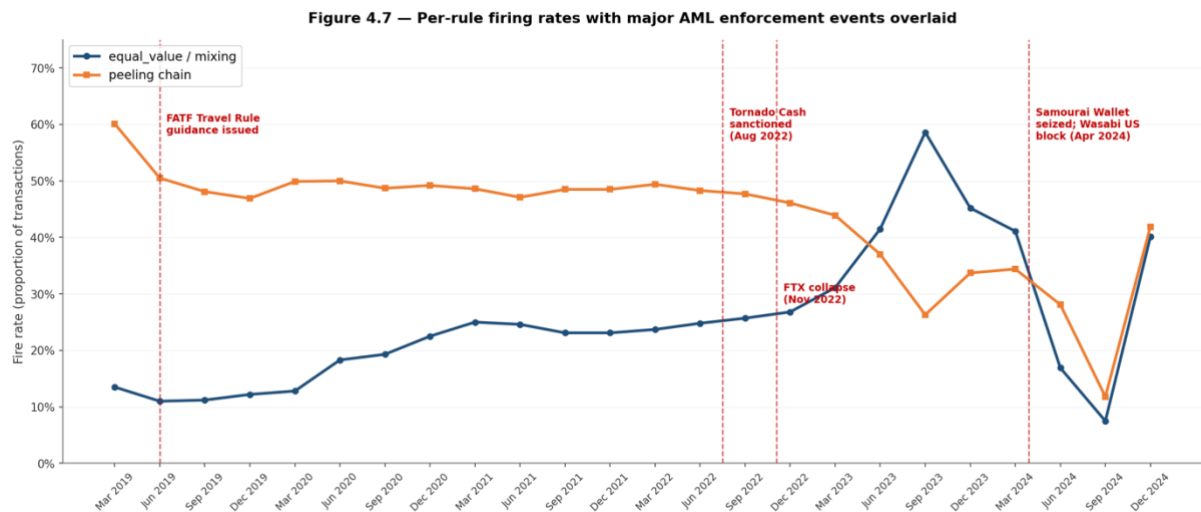


Figure 4.7 — Per-rule firing rates with major AML enforcement events overlaid.

The pattern shown in Figure 4.7, which displays the per-period rule firing rates (originally in Figure 4.2) with enforcement events shown on top of it, follows the shape of that landscape very closely. The 'equal_value' and 'mixing' rules both increased consistently from 13.5% in the first part of 2019 to a maximum of 58.6% in the third quarter of 2023, then dropped to 7.5% by the third quarter of 2024. Two aspects of this pattern are particularly notable. The peak in 2023 is approximately one year following the sanctions against Tornado Cash in August 2022, which is consistent with illicit mixing activity moving to other CoinJoin services, such as Wasabi and Samurai, and subsequently those services also facing enforcement in April 2024. The decline at the end of 2024 corresponds with the seizure of Samurai Wallet and Wasabi's exit from the American market, both of which considerably diminished the CoinJoin infrastructure that is visible on the blockchain. The peeling-chain rule shows the inverse pattern, declining from 60.1% in early 2019 to 11.8% by the third quarter of 2024, consistent with the broader shift toward bundled mixing techniques in place of the simpler peeling-chain layering that earlier laundering operations relied on. The

framework's measurements therefore track what Arbabi et al. (2025) and Mariani and Homoliak (2025) describe as an evolving mixer landscape, providing empirical evidence that the FATF-operationalised features are sensitive to behavioural shifts rather than static across the sample window.

Second, the temporal validation result, train F1 of 0.995, test F1 of 0.996, drift of +0.001, requires careful interpretation alongside Figure 4.7. Taken at face value, the near-zero drift would suggest the framework generalises across years of evolving Bitcoin behaviour.

However, the pseudo-labels used for the majority of training and test data are generated by the same deterministic FATF rules in both periods. A classifier trained on rule-derived labels in one period and tested on rule-derived labels in another will perform well as long as the rule logic is internally consistent, regardless of whether the underlying illicit behaviour has shifted. The honest reading is therefore that the temporal stability shown in §4.8.2 reflects rule-label self-consistency rather than independent generalisation to novel illicit behaviour. The stronger evidence of behavioural sensitivity is the cross-epoch rule firing trajectory in Figure 4.7, which shows the framework registering real shifts in obfuscation activity. The temporal F1 is best read as confirming that the framework does not significantly degrade across the sample window, with claims about generalisation to unseen illicit patterns deferred to future work.

Third, the mixer-handling mechanism described in §3.6 succeeded by the measure that matters: rather than becoming a source of noise or false positives, `coinjoin_strength` became the most predictive single feature in the model, contributing gain of approximately 63 million, roughly 20 times the next-highest feature. This is consistent with Zola et al.'s (2025) observation that even mixer behaviour leaves recurring topological forms that can be exploited structurally. The conservative threshold of 0.1 on the CoinJoin score, combined with the suppression of wallet-level features for likely-mixed transactions, prevented the framework from attributing merged users' wallet histories to any one transaction, directly addressing the obfuscation issue raised by Arbabi et al. (2025) and Mariani and Homoliak (2025) in §2.4.

4.8.3 Scope, contribution and limitations of the findings

The findings extend prior work by quantifying what FATF operationalisation contributes when integrated into a supervised pipeline, but the contribution is bounded in ways the per-source analysis makes visible. Figure 4.8 summarises the per-source performance gap alongside the approximate composition of each source's illicit pool.

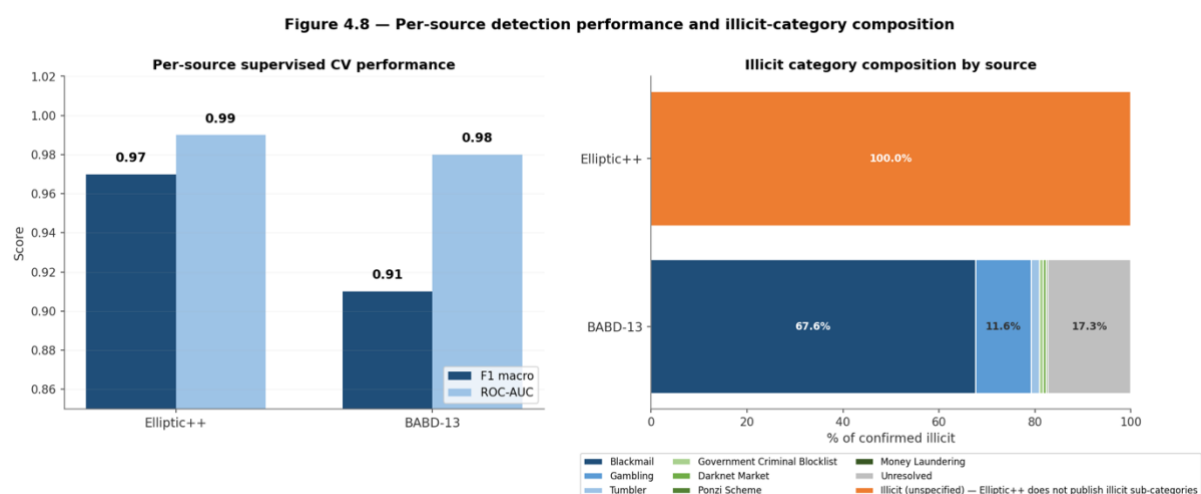


Figure 4.8 — Per-source supervised CV performance and approximate illicit-category composition. Categories are illustrative, derived from each source's documentation.

As Figure 4.8 demonstrates, Elliptic++ by itself had a notably higher F1 score of 0.97 compared to BABD-13's score of 0.91, despite both being trained and subjected to cross-validation in the same way. The chart on the right of Figure 4.8 provides a possible explanation for this: Elliptic++'s confirmed illicit addresses are largely associated with those running dark market services, ransomware operations, and money laundering, all of which show the kind of deliberately confusing structure that the Financial Action Task Force (FATF) list of features was intended to identify. BABD-13's illicit transactions, however, are dominated by blackmail at 67.6% while gambling sits at 11.6% and tumblers at only 1.7%. With only 1.7% of BABD-13's transactions being obfuscation related, these first two categories do not have a consistent structural pattern as they are often, on paper, simple transactions. The framework, therefore, works better for illicit behaviour that leaves a structural trace, rather than categorically illicit activities that may not. This is a scope condition that strengthens, rather than weakens, the answer to the research question: regulatory red flags can be operationalised effectively for the obfuscation-heavy laundering

activities that Anti-Money Laundering (AML) enforcement most frequently targets but are less useful as the underlying illicit category moves further from that profile.

Two further points are worth making. The unsupervised cluster analysis in §4.6 identified cluster 1 with 99.5% illicit purity and recovered roughly 39% of confirmed illicit transactions without access to labels during clustering, addressing the label-scarcity concern raised by Vlahavas et al. (2024): the framework retains utility even when labels are sparse. Conversely, the rule_fresh_outputs zero-firing result noted in §4.2 shows that operationalisation depends on what the underlying data can express, a different source, or a redesigned rule, would be needed to capture that signal.

Chapter 5: Conclusion & Limitations

5.1 Summary of contributions

This study set out to test the extent to which regulatory red flags for suspicious Bitcoin activity can be operationalised into a computational detection framework under realistic AML conditions. The work delivered five contributions. First, a layered detection framework was built that translates six FATF red flags into thirty-five transaction-level features and combines rule-based pseudo-labelling, K-Means and DBSCAN clustering, and LightGBM supervised classification within a single deployable pipeline. Second, mechanism validity was established against 149,128 externally confirmed labels: F1 macro of 0.96 on the augmented set and 0.93 on the natural set demonstrated that the framework produces detection signal against verified ground truth rather than only against self-generated labels. Third, the cross-epoch rule firing trajectory was shown to track real world Bitcoin AML developments over time, providing empirical evidence that the framework is sensitive to behavioural shifts rather than static across the sample window. Fourth, the supervised architecture matched or exceeded published baselines on Elliptic++, with the protocol caveats discussed in §4.8.1. Fifth, the per-source analysis identified a meaningful scope condition: the framework is more effective for obfuscation-heavy laundering activity than for categorically illicit categories that do not produce equivalent structural signatures.

5.2 Reflection and study-level limitations

Three reflexive limitations bound the interpretation of these findings beyond the methodological caveats already noted in §3.8. First, the temporal validation result of near-zero drift cannot be read as independent evidence of generalisation to novel illicit behaviour, because the same deterministic FATF rules generate pseudo-labels in both the training and test windows, producing a measure of rule-label self-consistency rather than behavioural transfer. The cross-epoch rule firing trajectory is the stronger evidence of behavioural responsiveness. Second, confirmed-label coverage from public sources concentrates on earlier Bitcoin eras, with BABD-13 and BitcoinAbuse coverage tapering from 2021 onward; the mechanism evaluation is therefore most evidence-rich for the 2019–2021 sample and least so for the contemporary period where AML enforcement is most active, a structural property of the public-label landscape rather than a flaw in the framework. Third, the framework operates on transaction-level features and cannot detect coordinated multi-transaction laundering schemes that only become visible through graph traversal across wallets and intermediaries. This is consistent with the research question's transaction-level scope but limits the kinds of laundering pattern the framework can identify.

5.3 Future work

The most straightforward development would be to empirically determine the best thresholds for the rules using actual, confirmed data. Currently, the thresholds for the rules are based on the judgements and guidelines from the Financial Action Task Force (FATF), rather than being refined by testing; with the 149,128 confirmed transactions available and a method for evaluating threshold sensitivity already planned, it would be possible to substantially improve the ability of the rule layer to find all relevant transactions from each source, in particular for the BABD-13 categories which are the least successful at present.

Additionally, the framework could be extended to look at the bigger picture beyond each individual transaction. By including details of the links between wallets, statistics on how funds move across multiple transactions, and the way groups of clusters are organised, the system could identify laundering that involves multiple parties in a way that analysing each transaction on its own could not. This is strongly related to the graph convolutional network

work of Weber et al. (2019) and the topological clustering of Vlahavas et al. (2024); the idea is to include information from the 'graph' of transactions alongside the features defined by FATF, and not to use the graph information instead of those features. To overcome the issue of data becoming outdated, as discussed in §3.8, the system could be continuously retrained or updated with a 'sliding window' of the most recent data.

The framework demonstrates that regulatory red flags and computational detection are not competing approaches but complementary layers of the same problem. The translation between them is feasible and measurable against confirmed ground truth, though bounded by the visible patterns of the underlying illicit activity. Whether that translation extends to behaviours without structural signatures, to multi-transaction laundering schemes, and to assets beyond Bitcoin remains an open question worth pursuing.

References

- Al Naqbi, N., Amin, F. and Saeed, S. (2025) 'Global trends and insights into cryptocurrency-related financial crime: A systematic review of money laundering and counter-measures', *Cogent Economics & Finance*, 13(1), pp. 1–27.
- Arbabi, M.S., Shojaeinasab, A. and Najjaran, H. (2025) 'Mixing services in Bitcoin and Ethereum ecosystems: A review', *IET Blockchain*, 5(1), pp. 1–18.
- BitcoinAbuse (2022) *BitcoinAbuse Database — Public CSV Dump*. [Online] Available at: <https://www.bitcoinabuse.com/> (Accessed: March 2026).
- Creswell, J.W. (2014) *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. 4th edn. Thousand Oaks: SAGE.
- Dubois, A. and Gadde, L.-E. (2002) 'Systematic combining: an abductive approach to case research', *Journal of Business Research*, 55(7), pp. 553–560.
- Elmougy, Y. and Liu, L. (2023) 'Demystifying fraudulent transactions and illicit nodes in the Bitcoin network for financial forensics', in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. New York: ACM, pp. 3979–3990.
- Financial Action Task Force (2020) *Virtual assets red flag indicators of money laundering and terrorist financing*. Paris: FATF. Available at: <https://www.fatf-gafi.org/> (Accessed: November 2025).
- Harlev, M.A., Yin, H.S., Langenheldt, K.C., Mukkamala, R. and Vatrappu, R. (2018) 'Breaking bad: De-anonymising entity types on the Bitcoin blockchain using supervised machine learning', in *Proceedings of the 51st Hawaii International Conference on System Sciences*. pp. 3497–3506.
- Jahanshahloo, H. (2023) *CUBiD: Cardiff University Bitcoin Database. User's Manual*. Cardiff University Business School.
- Mariani, G. and Homoliak, I. (2025) 'SoK: A survey of mixing techniques and mixers for cryptocurrencies', *ACM Computing Surveys*, 57(8), pp. 1–38.

Narayanan, A., Bonneau, J., Felten, E., Miller, A. and Goldfeder, S. (2016) Bitcoin and Cryptocurrency Technologies: A Comprehensive Introduction. Princeton: Princeton University Press.

OpenSanctions (2024) *Consolidated Targets Dataset (Daily Build)*. [Online] Available at: <https://www.opensanctions.org/datasets/> (Accessed: March 2026).

Ratner, A., Bach, S.H., Ehrenberg, H., Fries, J., Wu, S. and Ré, C. (2017) 'Snorkel: rapid training data creation with weak supervision', *Proceedings of the VLDB Endowment*, 11(3), pp. 269–282.

Saunders, M., Lewis, P. and Thornhill, A. (2019) *Research Methods for Business Students*. 8th edn. Harlow: Pearson Education.

Vlahavas, G., Karasavvas, K. and Vakali, A. (2024) 'Unsupervised clustering of Bitcoin transactions', *Financial Innovation*, 10, pp. 1–27.

Weber, M., Domeniconi, G., Chen, J., Weidele, D.K.I., Bellei, C., Robinson, T. and Leiserson, C.E. (2019) 'Anti-money laundering in Bitcoin: Experimenting with graph convolutional networks for financial forensics', *arXiv preprint arXiv:1908.02591*.

Xiang, Y., Lei, Y., Bao, D., Liu, T., Yang, T., Ren, W., Li, J. and Choo, K.K.R. (2024) 'BABD: A Bitcoin address behavior dataset for pattern analysis', *IEEE Transactions on Information Forensics and Security*, 19, pp. 2171–2185.

Zhang, Y., Wang, J. and Luo, J. (2020) 'Heuristic-based address clustering in Bitcoin', *IEEE Access*, 8, pp. 210582–210597.

Zola, F., Seguro-Gil, L., Bruse, J.L., Galar, M. and Orduna-Urrutia, R. (2025) 'Topological analysis of mixer activities in the Bitcoin network', *Future Generation Computer Systems*, 165, pp. 1–16.

Ethics Form

The signed ethics form (pages 38–40 of the submitted version) is omitted from this copy because it contains signatures and personal details.

Appendices

Appendix A — Feature Specification

Feature Group 1 — Output statistics

Mean, standard deviation, maximum, minimum, and Shannon entropy of output values, capturing structuring or layering behaviour where outputs are near-identical or one large output is paired with smaller remainders.

Feature Group 2 — Address counts and ratios

Number of input and output addresses, fan-in/fan-out, and input-output ratio, capturing fund dispersion and the high fan-out patterns flagged by FATF (Financial Action Task Force, 2020).

Feature Group 3 — FATF rule signals

Feature Group 3 contains six binary or combined signals that directly translate Financial Action Task Force (FATF) regulatory warnings into operational terms: the equal-value flag identifies outputs that are almost exactly the same (which corresponds to structuring), the peeling chain flag identifies a primary output with small amounts remaining (layering), fresh output ratio measures the proportion of output addresses that are new, high value entropy looks for unusual distributions, high fan-out identifies the spreading of funds, and CoinJoin strength is a combined measurement of how much mixing is occurring, based on equal-value outputs, and the number of inputs and outputs (Arbabi et al., 2025; Mariani and Homoliak, 2025; Zola et al., 2025).

Feature Group 4 — Wallet context

Feature Group 4, wallet context, considers wallet age, the total number of transactions, the total amount sent, the final balance, balance velocity, and balance range. These provide the history of the entity to allow a distinction between the activity of a new wallet and one that has been used for some time.

Feature Group 5 — Output-address statistics

Feature Group 5, output-address statistics, looks at the average transaction count, balance, age, and the fraction of zero-balance accounts across the addresses that are receiving funds. This is a way of looking at where the money is going, rather than who is sending it.

Feature Group 6 — Transaction metadata

Feature Group 6, transaction metadata, includes the confirmation timestamp, the day of the week, and the transaction fee, offering a time-based context and assisting the model in differentiating automated transactions from those completed manually.

A.1 Output statistics

Feature	Definition / Computation	Source	Deployed	Gain rank	Group
total_input	Sum of all Input_Value per transaction	TD	Y	3	A
min_output	Minimum Output_Value per transaction	TD	Y	13	A
total_output	Sum of all Output_Value per transaction	TD	Y	15	A

Feature	Definition / Computation	Source	Deployed	Gain rank	Group
std_output	Standard deviation of Output_Value (ddof=0)	TD	Y	20	A
value_entropy	Shannon entropy in bits: $-\sum p \cdot \log_2(p)$ where $p = \text{output_value} / \text{total_output}$	TD	Y	23	A
mean_output	Mean of Output_Value per transaction	TD	Y	28	A
max_output	Maximum Output_Value per transaction	TD	Y	29	A

A.2 Address counts and ratios

Feature	Definition / Computation	Source	Deployed	Gain rank	Group
num_outputs	Count of distinct Output_Address per transaction	TD	Y	7	C/D
io_ratio	$\text{num_inputs} / \text{num_outputs}$ (clipped, min denominator = 1)	D	Y	19	C/D
value_ratio	$\text{total_input} / \text{total_output}$ (clipped, min denominator = 1e-9)	D	Y	22	C/D
n_distinct_output_wallets	Count of distinct Output_Wallet per transaction	TD	Y	27	C/D
num_inputs	Count of distinct Input_Address per transaction	TD	Y	30	C/D

Feature	Definition / Computation	Source	Deployed	Gain rank	Group
n_distinct_input_wallets	Count of distinct Input_Wallet per transaction	TD	Y	35	C/D

A.3 FATF rule signals

Feature	Definition / Computation	Source	Deployed	Gain rank	Group
coinjoin_strength	Composite score: $0.6 \times \text{equal_value_flag} + 0.2 \times \min(\text{num_inputs}/10, 1) + 0.2 \times \min(\text{num_outputs}/10, 1)$	D	Y	1	E/F
peeling_flag	1 if $\text{max_output} \geq 80\%$ of total_output AND $\text{num_outputs} \geq 2$; else 0	D	Y	2	E/F
equal_value_flag	1 if coefficient of variation (std/mean) of output values < 0.01 ; else 0	D	Y	6	E/F
fresh_output_ratio	Fraction of Output_Address values not present in Address_Info (unseen addresses)	TD+AI	Y	34	E/F

A.4 Wallet context

Feature	Definition / Computation	Source	Deployed	Gain rank	Group
wallet_total_tx	N_Transactions from Wallet_Info for the sending wallet	WI	Y	10	G/H

Feature	Definition / Computation	Source	Deployed	Gain rank	Group
wallet_age	Last_Active_Block – First_Active_Block (clipped ≥ 0)	WI	Y	11	G/H
wallet_sent_ratio	N_Sent_Transactions / N_Transactions (clipped 0–1)	WI	Y	16	G/H
wallet_total_sent	Total_Sent from Wallet_Info for the sending wallet	WI	Y	18	G/H
wallet_final_balance	Final_Balance from Wallet_Info	WI	Y	21	G/H
wallet_size	N_Addresses from Wallet_Info for the sending wallet	WI	Y	24	G/H
wallet_balance_ratio	Final_Balance / Total_Received (clipped 0–1)	WI	Y	25	G/H
wallet_balance_velocity	(last_balance – first_balance) / n_observations, streamed from balance-block table	WBB	N	—	G/H
wallet_balance_range	max_balance – first_balance across all observed blocks	WBB	N	—	G/H
wallet_balance_std	Standard deviation of balance across all observed blocks (Welford online)	WBB	N	—	G/H

A.5 Output-address statistics

Feature	Definition / Computation	Source	Deployed	Gain rank	Group
mean_out_addr_age	Mean (Last_Active_Block – First_Active_Block) across output addresses	AI	Y	4	J
mean_out_addr_tx_count	Mean N_Transactions across output addresses	AI	Y	8	J
mean_out_addr_balance	Mean Final_Balance across output addresses	AI	Y	9	J
mean_out_addr_sent_ratio	Mean (Total_Sent / Total_Received) across output addresses, clipped 0–1	AI	Y	12	J
frac_out_addr_zero_balance	Fraction of output addresses with Final_Balance = 0	AI	Y	32	J
mean_out_addr_bal_velocity	Mean output-address balance velocity (requires balance-block tables)	—	N	—	J

A.6 Transaction metadata

Feature	Definition / Computation	Source	Deployed	Gain rank	Group
tx_hour	Hour of day (0–23) from transaction datetime	T	Y	5	I
tx_fee_normalized	tx_fee / block_avg_fee (block-mean fee from Block_Header)	T+BH	Y	14	I
tx_fee	Transaction fee (Fees column from Transactions)	T	Y	17	I

Feature	Definition / Computation	Source	Deployed	Gain rank	Group
lock_time_flag	1 if Lock_Time > 0; else 0	T	Y	26	I
tx_day_of_week	Day of week (0=Monday, 6=Sunday) from transaction datetime	T	Y	31	I
is_night	1 if tx_hour $\in$ [0, 5]; else 0	T	Y	33	I

Appendix B — Per-Period Descriptive Statistics

This appendix reports per-snapshot descriptive statistics across the 24 quarterly samples. Each row corresponds to a single CUBiD snapshot day (8 March, 8 June, 8 September, or 8 December for each year 2019–2024). The data is split into two tables for readability: Table B.1 covers transaction volume and value statistics; Table B.2 covers the label distribution under the hybrid scheme. Together they support §4.2 (label distribution and rule firing) and §4.3 (confirmed-label enrichment).

B.1 Transaction volume and value per snapshot

Period	Tx count	Mean tx (BTC)	Median tx (BTC)	Mean fee (sat)	Mean inputs	Mean outputs
2019-Q1	285,264	4.6030	0.077237	12,106	2.04	2.48
2019-Q2	294,369	3.1585	0.026409	24,923	1.86	2.27
2019-Q3	288,507	2.4279	0.016467	7,538	1.72	2.15
2019-Q4	278,853	2.6169	0.018681	5,201	1.76	2.14
2020-Q1	302,694	2.0342	0.022469	5,333	1.77	2.23
2020-Q2	317,275	3.0996	0.041244	13,150	2.41	2.54
2020-Q3	332,668	8.1266	0.050000	25,269	2.26	2.72

Period	Tx count	Mean tx (BTC)	Median tx (BTC)	Mean fee (sat)	Mean inputs	Mean outputs
2020-Q4	330,970	4.8912	0.031325	21,812	2.25	2.78
2021-Q1	291,955	6.7142	0.018419	31,014	3.02	3.08
2021-Q2	246,906	10.2384	0.027754	14,276	3.12	3.32
2021-Q3	270,483	12.2771	0.018033	5,437	2.43	3.06
2021-Q4	266,110	16.2787	0.015006	3,888	2.41	3.12
2022-Q1	255,771	34.7071	0.015000	4,138	2.79	3.24
2022-Q2	261,597	7.2140	0.018230	4,139	2.59	3.20
2022-Q3	262,863	18.7818	0.025471	5,854	2.74	3.36
2022-Q4	263,838	3.5690	0.024592	5,638	2.67	3.18
2023-Q1	292,493	2.9662	0.014507	8,976	2.24	2.97
2023-Q2	407,006	2.6015	0.004227	17,087	2.04	2.56
2023-Q3	414,799	1.5523	0.001822	8,644	1.95	2.57

Period	Tx count	Mean tx (BTC)	Median tx (BTC)	Mean fee (sat)	Mean inputs	Mean outputs
2023-Q4	411,456	2.4018	0.003999	37,076	1.79	2.57
2024-Q1	426,928	2.8147	0.002676	11,255	1.79	2.66
2024-Q2	509,927	0.9988	0.002864	30,499	1.69	2.26
2024-Q3	910,083	0.5942	0.000315	695	1.26	1.46
2024-Q4	420,320	1.0534	0.001452	3,621	2.18	2.23

B.2 Label distribution per snapshot

Period	Confirmed illicit	Confirmed licit	Pseudo illicit	Pseudo licit	Uncertain
2019-Q1	11,729	4,819	39,600	68,021	161,095
2019-Q2	15,561	5,130	34,330	107,527	131,821
2019-Q3	14,357	4,418	33,566	108,835	127,331
2019-Q4	9,351	3,428	35,026	107,906	123,142
2020-Q1	10,201	3,408	39,721	105,880	143,484

Period	Confirmed illicit	Confirmed licit	Pseudo illicit	Pseudo licit	Uncertain
2020-Q2	11,830	3,856	60,093	88,579	152,917
2020-Q3	10,026	2,883	67,159	94,689	157,911
2020-Q4	7,604	2,610	77,424	85,036	158,296
2021-Q1	5,414	1,931	77,279	69,932	137,399
2021-Q2	3,551	552	64,523	63,568	114,712
2021-Q3	3,283	412	66,061	71,490	129,237
2021-Q4	1,530	561	65,356	70,275	128,388
2022-Q1	1,328	431	64,204	63,479	126,329
2022-Q2	962	431	69,098	64,977	126,129
2022-Q3	969	317	72,352	64,163	125,062
2022-Q4	1,134	255	74,817	65,833	121,799
2023-Q1	969	208	95,208	68,029	128,079

Period	Confirmed illicit	Confirmed licit	Pseudo illicit	Pseudo licit	Uncertain
2023-Q2	642	144	173,563	79,640	153,017
2023-Q3	363	127	247,618	57,411	109,280
2023-Q4	555	223	191,546	80,178	138,954
2024-Q1	594	52	180,349	97,619	148,314
2024-Q2	280	21	88,915	277,090	143,621
2024-Q3	309	19	71,615	730,869	107,271
2024-Q4	319	31	172,620	71,023	176,327

Appendix C — Rule Specification

Rule	FATF indicator	Threshold / computation	Input feature(s)	Rationale and limitations
rule_equal_value	Structuring / CoinJoin splitting	$CV(output_value) = \text{std}/\text{mean} < 0.01$	equal_value_flag	FATF (2020) flag indicators on structured deposits and CoinJoin-style equal-value splitting; the 1% CV threshold captures near-identical output amounts within a tx.
rule_peeling	Layering / peeling chain	$\max(output_value) \geq 80\% \times \text{total_output}$ AND $\text{num_outputs} \geq 2$	peeling_flag	FATF (2020) layering indicator; a single dominant output paired with small change outputs is the canonical peeling-chain pattern.
rule_high_fan_out	Fund dispersion	$\text{num_outputs} \geq 10$	num_outputs	FATF (2020) dispersion indicator; ten or more distinct recipient addresses in a single transaction is unusual for legitimate use and characteristic of cash-out / mixer disbursement.
rule_fresh_outputs	Address churn / layering	$\text{fresh_output_ratio} \geq 0.80$	fresh_output_ratio	FATF (2020) churn indicator. Structurally inert in the multi-period dataset because CUBiD's Address_Info is global (covers blocks 677–873k), so every output address has prior history relative to its block by construction. Limitation flagged in §3.2 and §4.2.
rule_mixing	Mixer involvement	$\text{coinjoin_strength} \geq 0.5$	coinjoin_strength	FATF (2020) mixer indicator. The composite combines equal-value splitting (60% weight) with input/output count signals (20% each), matched against the 0.5 cutoff.

Rule	FATF indicator	Threshold / computation	Input feature(s)	Rationale and limitations
rule_high_entropy	Anomalous value distribution	value_entropy ≥ 3.0 bits	value_entropy	FATF (2020) anomalous-value indicator. Three bits of Shannon entropy implies output values spread across roughly 8 distinct magnitudes — suggestive of structuring or value obfuscation.

Appendix D — Model Configuration and Hyperparameters

This appendix documents the supervised classifier, the unsupervised clustering algorithms, the cross-validation protocol, and the computational environment used to produce the results in Chapter 4. All hyperparameters are fixed at the values below; no hyperparameter search was performed.

D.1 Supervised classifier (LightGBM)

LightGBM is the primary classifier. Random Forest is the fallback when LightGBM is unavailable; in the reported results LightGBM was used throughout.

LightGBM hyperparameter	Value	Random Forest fallback	Value
n_estimators	500	n_estimators	200
learning_rate	0.05	max_depth	10
max_depth	8	class_weight	balanced
num_leaves	63	random_state	42
class_weight	balanced	n_jobs	-1
random_state	42	preprocessing	SimpleImputer(median) + StandardScaler

LightGBM hyperparameter	Value	Random Forest fallback	Value
n_jobs	-1		
verbose	-1		

D.2 Unsupervised clustering

Algorithm	Parameter	Value
K-Means	n_clusters	3
K-Means	n_init	10
K-Means	random_state	42
DBSCAN	eps	1.5
DBSCAN	min_samples	5
DBSCAN	algorithm	ball_tree
DBSCAN	n_jobs	-1

D.3 Cross-validation protocol

Setting	Value
Method	StratifiedKfold (sklearn.model_selection)
Number of folds	5
Shuffle	True
Random state	42
Stratification	By binary illicit / licit class label
Mechanism-validation scoring	Macro and weighted F1; per-class precision and recall; ROC-AUC; PR-AUC; out-of-fold confusion matrix
Class-weighting	class_weight='balanced' in both LightGBM and Random Forest; no additional resampling at the data level

D.4 Computational environment

Component	Value
Python	3.13.12
LightGBM	4.6.0
scikit-learn	1.8.0

Component	Value
pandas	2.3.3
numpy	2.4.3
matplotlib	3.10.9
CPU / RAM	Apple M3 Pro (12-core) / 18 GB
OS	macOS (Darwin 25.3.0, arm64)