Natural Resources for Human Health

eISSN: 2583-1194

Original Research

Received 22 July 2026
Revised 18 Aug 2026
Accepted 05 Sep 2026

Natural Resources for Human Health 2026, Vol. 6 (20s): 517–529

KEYWORDS:
artificial intelligence, financial crime, anti-money laundering, network analytics, graph intelligence, fraud detection, temporal validation, explainable AI, coordinated crime.

A Multi-Layered Artificial Intelligence Framework for Detecting Coordinated Financial Crime Networks

Mozumder Maliha Shamsuddin

Student, University of Chittagong, malihasarkerima1996@gmail.com,
Orchid ID:- https://orcid.org/0009-0003-4332-7302

ABSTRACT: Coordinated financial crime involves distributed transactions, accounts, intermediaries, and payment vehicles, which form complex adaptive networks that cannot be inferred using transaction-level analytics. In this work, we designed a multi-layer artificial intelligence system and demonstrate its elements, including transaction attributes, contextual features, graph structure, temporal properties, and explainable risk scores in the task of detecting coordinated financial crime. We evaluated the system on two publicly available synthetic datasets. In Dataset I, Cash, Cheque, and Cross-border transactions were labeled as suspicious, whereas ACH Transfers, Credit Cards, Debit Cards, E-Wallet, and Mobile Payments were considered normal. This study explored the performance of a hierarchical detection model that employs a set of diverse evidence in a multi-layer system and the challenges of the model performance assessment in terms of leakage, network effects, temporal factors, and human-related aspects. The study highlighted the opportunities and limitations of the multi-layer approach.

1. INTRODUCTION

Financial systems nowadays are a part of the digital age where the rise of online banking, real-time payments, mobile apps, cross-border transactions, cryptocurrencies, and automations take place. While it may mean easier access and convenience for the user, at the same time it raises a question of growing criminal use which spans across many accounts, counterparties, financial institutions, devices, and jurisdictions. A single out-of-place transaction may be no issue in itself; what may be the issue is that it is a part of a bigger picture that involves layering, mules, and round trips, which all together are seeking to hide the real beneficiaries.

For a long time, we have had a rule-based approach with threshold scoring and scenario analysis for money laundering detection. The former has a great number of exceptions, and the latter is easy to get around by redesigning transactions to fall below the set thresholds. We are now seeing the introduction of AI into the picture to identify complex patterns in high-dimensional transaction data. However, we are also seeing that transaction-level machine learning models do not completely solve the issue of coordinated crime that by its very nature is a relational issue. The key signal may not be the attributes of any given transaction, but rather the relationships between transactions for example in complex account networks and circular chains as well as their time-based evolution.

The paper reports and analyzes what we present as a tiered approach to financial crime detection. We consider it to be the problem of tiered analysis: the analysis of transaction data, entity data, their relationships, and other analytics. We put it all together as a single coherent system that draws on the strengths of each tier while also recognizing its limitations. Our study contributions are in the design of these systems that we then decompose into transaction intelligence, behavioral and contextual analysis, the entity and relationship layer, graph-level signals, time-based features, risk fusion, explainability, and manual review.

One of the key contributions of the paper is the emphasis on ensuring that one's evaluation methodology does not obfuscate the capabilities of their model. As such, we include a target-leakage audit to highlight potentially misleading features for an overly optimistic assessment of one's predictive power. We stress the importance of distinguishing between the signal in the

data and modeling decisions and advocate for a separate assessment of the transaction classification and network structure when the available data does not permit a joint test.

The problem mentioned also concerns the separation between a suspicious observation and a substantiated financial-crime network: in practice, an alert could have been triggered because the account does not follow its own history, or involves multiple parties in anomalous exchanges, or the set of transfers follows a particular pattern. All these considerations represent different types of evidence and should be assessed separately: as such, it is preferable to focus on who participates, the direction and order of transfers as well as the accompanying business logic for the observed coordinated pattern. Hence, a reliable detector must contemplate general screening and event reconstruction capabilities to explain the source of an alert — and requires multiple complementary analytical layers than a single transaction classifier.

2. RELATED LITERATURE AND RESEARCH CONTEXT

Due to the relational nature of money laundering activity, graph-based approaches have gained increasing attention in recent years for anti-money laundering research. Different types of nodes could be defined, for instance: accounts, customers, companies, wallets, devices, etc., and edges could represent transfers or other relations. Centrality, communities, reciprocity, cycles, and time-evolving subgraphs can all be leveraged to detect illicit activity in ways that go beyond what tabular data can tell.

The Elliptic Bitcoin set is a widely used resource for the study of graph-based illicit transaction detection. It includes over 200k transaction nodes and over 200k directed payment-flow edges, of which only a small number are labeled. Also, many studies that have used the Elliptic framework and similar anti-money laundering (AML) transaction data report that graph learning is very useful for financial forensics. At the same time, it is noted that temporal drift, label sparsity, imbalance, and data leakage may play a role in degrading graph model performance, which is why they should be addressed in the design of experiments.

While AMLSim and the IBM synthetic AML datasets represent another set of research that simulates banking transactions and laundering patterns (e.g., fan-in, fan-out, and cycles), it should be noted that synthetic data can be vital to the task due to the difficulty in sharing real-world data. On the other hand, the data generation procedures for synthetic graphs can sometimes render the task too easy. One solution is to perform a data auditing step that determines whether labels are encoded in any easily predictable field, such as the payment channel, amounts range, account identifier, or other features.

Explainability and governance are no less important: a bank cannot blindly trust the "unexplainable" score as a proof of money laundering and would rather prefer an algorithm that offers grounds for investigation. In particular, the latter option enables investigators to prioritize alerts based on the score's explanation. Indeed, one would be interested in scrutinizing the evidence that suggests the suspicious behavior, the network role of the account or entity, and the patterns that support the suspicion.

Thus, the current research's findings can be arranged into complementary analytical tasks. In particular, transaction classifiers are concerned with whether a single record resembles suspicious instances identified in the past, and network analytics reveal whether a network occupies a rare configuration. Meanwhile, temporal methods highlight time-sensitive patterns that are characteristic of the laundering schemes. Each approach's performance is dependent upon the data characteristics, and all three tasks have distinct failure modes. For instance, a graph-based method might be overly reliant on the presence of multiple transactions between the same accounts, which undermines the instances with only one transaction. In turn, a classification model's performance depends on whether the suspicion labels reflect practical relevance. For instance, an apparently effective classifier might only identify routine transaction categories, which is unsuitable for the given task. Therefore, one should always be aware of what the data can support. In particular, the two experiments utilized in this article illustrate that each analytical task should be evaluated against separate evidence and targets.

The literature review also highlights a critical challenge in evaluation design: the target of prediction should be on par with the evidence provided. In particular, a transaction record, an account, and a network of interconnected instances are not the same objects. While suspicious-transfer labels are adequate for transaction scanners, they could miss the network-level criminal conduct. Likewise, a network-level anomaly discovered at the account level could have little to do with the organization of laundering schemes. Accordingly, evaluation should specify the units of analysis, the presence of labels, and the time frame. It is therefore important to distinguish between detecting an outlying instance and identifying a coordinated criminal network.

3. RESEARCH PROBLEM AND OBJECTIVES

The problem is that patterns of coordinated financial crime cut across transactions and entities, potentially appearing innocuous at either level, individually. We develop a framework for synthesizing diverse sets of financial information in a multi-layer algorithmic system that identifies anomalous transactions and networks while being transparent in its modeling uncertainties and acknowledging the limitations of the inputs.

Our contributions are focused on the development of a multi-layer discovery framework. We tested one transaction-level classification layer on a temporal holdout and inspected the labeled sample for features that might have leaked into the training set; built a directional financial interaction graph from another public sample and inspected features that could indicate coordination, while evaluating the aggregation of transaction and graph levels without overstatement. Our analytical window is limited because we want to focus on specific research questions. The first is to understand how transaction attributes affect classifier performance on a temporal holdout and what spurious proxy variables contribute to apparent predictive power. The second goal is to evaluate what information about inter-account network structure can be extracted from such a sample and which accounts would be prioritized for review by a transparent network ranking. Both experiments are independent and neither result is taken as evidence that the other experiment generalizes. Thus, we emphasize the areas that require a joint discovery framework: a validated account linkage mechanism, network-level ground truth, time-aware assessment and a demonstration of value of the findings to investigators.

Another research goal is an emphasis on interpreting negative results. If the transaction classifier fails to find suspicious instances, we might wonder whether the selected features carry any useful information and if the chosen decision threshold is reasonable given the suspicious activity density. In case the graph-level indicators identify a dense core of nodes, we question this evidence by evaluating what generic network mechanisms could produce such patterns. In either case, we propose to consider alternative explanations that would explain the observation. Such an approach to analysis would reduce the risks of exaggerated claims of practical utility and invite investigators to engage the discovery framework as a whole, rather than a single component. It might be especially important in the case when the costs of a false positive or the missed network would be exceptionally high.

4. PROPOSED MULTI-LAYERED ARTIFICIAL INTELLIGENCE FRAMEWORK

4.1 Data and Entity Layer

In the first stage financial information is collected from many sources and we see the need for standardization of entity, transaction, account, time, location, device, and context attributes. As it pertains to entity resolution this is of great import it is not unusual for the same entity to be represented in different records. For a live system which is in use we see that probabilistic matches may benefit from including a level of confidence which in turn may require that human intervention take place in order to avoid the issue of false identification of our customers.

Graph and behavior analysis' credibility does in this data. As for customer which is but a single record going in at at the wrong node we may see false associations, also an issue of unsolved alias which in turn may cause a live player to break into many nodes. Also we see that transaction reversals, imported records that put in more then once, and time zones out of sync may report flow that isn't there or report it as going in a different direction. Also it is of great importance that an operational entity layer is in place which will house the raw data with normalized identifiers and which also reports how each connection was made. At the institutional or data system which is foreign to you we see that confidence levels and what to do in correction of errors is key. In our present synthetic studies we did not look at entity resolution performance; we present this as a need for future work instead of a new empirical report.

4.2 Transaction Intelligence Layer

The transaction layer which displays statistical or behavioral anomalies. While we may train supervised classifiers on past positive instances, unsupervised methods in some cases will bring up outlying which past data didn't. Examples of transaction layer attributes are transaction values, channels, currencies, timing, geographies, and user based variables. Our paper's empirical study uses a simple probabilistic classifier to present the validation process which is a basic model, not a top of the line production level classifier.

An approximate probability should report an observed frequency in a proper evaluation setting. In the case of the Categorical Naive Bayes study we see that it does not achieve this sort of operational calibration and in addition to that the best performance of its augmented models does not in fact support use of a deployment threshold.

4.3 Behavioral and Contextual Layer

Behaviorial intelligence is the comparison of present action to what is seen in the past for that entity or peer group. As to what is looked at we see time of transaction, difference between which currencies are used, the sender and receiver's geographies, how the transaction took place, what categories of customers are involved, and also any other factors which the bank's internal policies put forth. The goal is to identify what is normal and what is very abnormal which may in turn be put forward as suspicious in that context.

In this regard, it is recommended to adjust behavioral baselines depending on the account type and observation period. For instance, the history of a corporate settlement account suggests a considerably higher number of counterparts than a newly opened consumer account. In this regard, specific activities could be selected for further investigation according to the account's nature, history, and properly established peer groups. Besides, small sample sizes and observation periods reduce the ability to detect outliers due to the increased risk of rare legitimate activities. Nevertheless, the described context layer assists in explaining the outlier's nature, providing the rationale behind each outlier, and differentiating between alerts and ordinary transactions of a similar type. The paper's findings confirm this assumption, as no valid peer-group behavioral baseline can be found in the dataset.

4.4 Graph and Coordination Layer

The framework has a graph layer which in it financial entities are the nodes and the relationships between them are put forward by edges of the graph. Degree measures which allow the researcher to identify actors that stand out, and transaction volume measures can be had which play as indicators of the weight of the edges. Reciprocity is what we see in two way relationships, and cycles which are directed paths that go back to the start point. That kind of structure requires a temporal and transactional setting before it may be put into practice. Also we see more complex solutions like graph neural networks, temporal graph networks, heterogeneous graphs and subgraph classification algorithms which do a better job at getting value out of structures which do not work in a tabular data layer.

The framework has a graph layer which in it financial entities are the nodes and the relationships between them are put forward by edges of the graph. Degree measures which allow the researcher to identify actors that stand out, and transaction volume measures can be had which play as indicators of the weight of the edges. Reciprocity is what we see in two way relationships, and cycles which are directed paths that go back to the start point. That kind of structure requires a temporal and transactional setting before it may be put into practice. Also we see more complex solutions like graph neural networks, temporal graph networks, heterogeneous graphs and subgraph classification algorithms which do a better job at getting value out of structures which do not work in a tabular data layer.

One further distinction is required between unweighted connectivity on the one hand and transaction-weighted relationships on the other. Counting a sender-receiver pair just once is suitable for representing distinct interactions, but hides whether the pair transferred funds just once or repeatedly. Adding the amounts may over-emphasise a few large transfers without indicating how widespread the network is . An integrated graph feature set would ideally retain both unique directed links and repeated transactions, and would make the observation window explicit. Network density, community memberships and the direction of funds could be interpreted against the account's expected activity rather than in isolation. The current graph summaries establish the presence of observable relationships; they do not establish the baseline distribution against which those relationships should be considered suspicious.

4.5 Temporal Layer

Financial-crime networks evolve over time. Thus, a temporal layer encodes the ordering, speed, recurrence, and persistence of interactions. Temporal validation is important in research. Random sampling of transactions can allow future patterns to leak into the training, therefore, the present classification experiment trains on earlier months and tests on later months.

The ordering of transactions matters because two graphs containing the same links may describe substantially different processes. A sequence completed within minutes can have a different investigative significance from similar transfers spread across months. Temporal analysis can therefore consider inter-transaction intervals, repeated bursts of activity, the introduction of new counterparties, and the persistence of particular paths. Such features must be computed using only information available at the moment a decision would have been made. Constructing a training feature with transfers that occurred after the target transaction can inadvertently expose future events and make an offline evaluation overly optimistic. A deployment-oriented study should simulate the time at which the alert was issued and reproduce the evidence then available.

4.6 Risk Fusion, Explainability, and Human Review

The final layer creates an interpretable risk representation from transactional, behavioral, contextual, graph, and temporal signals. A production system should explain what specific transactions and relationships resulted in the selection of the entity or network, what the level of uncertainty around the assessment is, and why an entity or network has been prioritized. Human investigators must review the evidence underlying an algorithmic alert, because the signal alone is not sufficient evidence of criminal activity.

Risk fusion, in turn, requires clear instructions as to how partly dependent evidence should be combined. Transactional irregularities, network centrality and timing peculiarities, however, could come from one piece of evidence, which requires care when combining their scores to avoid counting the same evidence more than once. One could argue that a calibrated model would analyse the increase of risk each layer provides and record instances where one suspicious feature outweighs others in contributing to the score. An investigator reading the case would appreciate an evidence breakdown, distinguishing between observed data, derived statistics and analytical conclusions. It would be important to point out discrepancies in supporting evidence and highlight what other records, if found, could impact upon the evaluation. Such a report would then be useful to guide an investigation, rather than simply suggest it, whilst still working on the fundamental principle of advising further review rather than presenting proof of crime.

5. EXPERIMENTAL METHODOLOGY

5.1 Experimental Design

The we did two separate studies as we did not find a public data set that at the same time had relevant transaction labels and in which that which is called for in repeat account networks played out. In study 1 we looked at classification results of a labeled AML transaction data set, and in study 2 we looked at structural association results in a different transaction network. That which we did was to separate out issues of transaction label accuracy from that of which we were able to successfully reconstruct the criminal networks.

In the two stage evaluation we looked at what is a wide spread issue in research which combines network visualizations with predictive metrics. While networks may present us with clusters, hubs or cycles a true positive rate for criminal network detection is not determined without verified illicit group labels. At the same time high transaction classification accuracy does not in itself prove that the classifier has identified relationships between multiple coordinated accounts. By keeping the elements of the study separate we are able to look at the structural component as an exploratory method and the labeled component as a classification exercise. Thus the present results should be looked at as a proof of concept for validation and graph explanation rather than a head to head comparison of competing AML solutions.

5.2 Dataset I: Labeled AML Transactions

In the 10,000 sender and receiver positions, we identified 9,958 unique accounts, only 42 of which appeared more than once, and in no more than two transactions. For this reason, in this particular case, we used transaction-level classification rather than attempt any graph validation on the dataset.

It is notable that the number of repeated account identifiers is exceedingly low, which is why this consideration is even relevant to the analysis of Dataset I. After all, in 9,958 unique account positions for every 10,000 sender and receiver appearances, the overwhelming majority of identified positions only appear once. From this characteristic, we can surmise that it would be hard to identify any long-term relationships, repeating counterparties, or multiple steps in a financial exchange using this dataset. The graph that would be generated by it would primarily depict an absence of structure or connectivity between nodes. In turn, its

comparison to other graphs that do have connectivity data would be skewed, and it would be easier to mistake a lack of information for actual information about the surveyed network. Therefore, limiting this dataset to transaction-level classification is the most sensible approach to avoid assuming too much about the relationships between financial actors.

5.3 Temporal Train-Test Split

To evaluate the model on transactions from a later period, we used data from January through September 2023 for training and the data from October through December 2023 for testing. Thus, we obtained 3,749 training transactions, 1,379 of which were suspicious, and 1,251 testing transactions, 446 of which were suspicious. In so doing, we created a temporal division to assess whether the model could learn from past transactions and generalize to later observations.

Chronological splitting makes a higher demand on a model than random mixing of records since one must consider how patterns obtained from the mining of previous months are applicable to the observation dates that follow. The division of 3,749 training observations and 1,251 test observations allows one to consistently compare the specified feature groupings according to this order. At the same time, the temporal split does not automatically exclude the possibility of leaks. The predictor that uses the same rule as the target can still carry complete information for both periods, while duplicated entities or deterministic categorical encodings can also overlap the split. Thus, chronological separation is a needed safety measure for this experiment, but the audit of the kind of payment underlines the need for complementing it by an inspection of the definitions of features and construction of labels.

Furthermore, temporal setting accents the need for preserving all the operations of pre-processing within the training limit. Dictionaries of categories, scaling factors, rules for imputations, and thresholds have to be derived from the training period if the estimation procedure uses information about distributions in the later data. The same applies to graph construction: a feature representing an entity on a particular date must not include neighbors or transfers that reach later. This is a different phenomenon from the label leakage due to the payment-type variable that is worrying even with the proper division into periods. Future experiments can explicitly document these safeguards to enable other researchers to identify whether an improvement reported has come from more informative representations or the utilization of unauthorized future information.

5.4 Classification Models

A categorical Naive Bayes benchmark was chosen due to the algorithm's ability to provide class probability estimates and examine the class-conditional feature probabilities. The base model included transaction amount bands and time of day categories as features while the transaction model incorporated the payment type, currency mismatch, and location mismatch features in addition to the base features. The multi-layer contextual model consisted of payment and received currencies, sender and receiver locations, and month as features. The Laundering_type field was excluded from the analysis since it explicitly states suspicious or normal transaction typology that would be considered label leakage.

5.5 Evaluation Metrics

For evaluation, we considered accuracy, precision, recall, F1 score, specificity, the confusion matrix, and the area under the ROC curve. We focused on F1, which is explained here, and the overall accuracy.

The other metrics differ from what we need in the particular case. Recall, for instance, is the proportion of suspicious transactions identified as such, while precision is the ratio of suspicious transactions to all transactions identified as suspicious. Specificity is the proportion of non-suspicious transactions that are correctly identified as such. F1 score is the harmonic mean of precision and recall, and AUC-ROC is the integral of true-positive rates over false-positive rates, which shows the efficiency of the threshold selection. Finally, an AML analyst would also be interested in the number of alerts that require investigation, the personnel and time costs associated with investigations, and whether the costs and benefits of unwisely escalated cases are acceptable. This information is not provided by any of the metrics discussed above. The performance of the classification model is thus evaluated based on a few characteristics of interest.

6. EXPERIMENT 1 RESULTS: TRANSACTION-LEVEL CLASSIFICATION

The baseline amount-and-time model did not detect any suspicious transactions at the default probability threshold. The dataset contained 1,251 records in the temporal test set, and the solution reported 805 true negatives, 446 false negatives, 0 true

positives, and 0 false positives. Thus, its accuracy was 64.35% which reflected the majority-class result. Besides, the F1-score was 0.000 and the ROC-AUC was 0.494 which was close to random classification.

After adding payment type and simple contextual variables, the transaction model had correctly classified all test examples. The same was true for the multi-layer contextual model which reported the same statistics. In both instances, the accuracy, precision, recall, F1-score, specificity, and ROC-AUC were 100% which was unrealistically high for an actual AML detection scenario. Therefore, this result prompted a feature-leakage investigation rather than supporting model deployment.

The comparison between the baseline and the augmented models demonstrates that model evaluation must go beyond summary performance metrics. The baseline predicted the majority class at the selected threshold and did not catch any suspicious cases, despite having accuracy above sixty percent. The augmented models had perfect scores, but the jump in performance is not an indicator of a better ability to uncover coordination. The additional variables contained a categorical field which repeated the label, and thus, the observed performance gain does not confirm learning of transferable fraud mechanisms. A credible follow-up experiment would involve removing the deterministic proxy, verifying the meaning of each remaining feature, and investigating whether performance remains similar with a genuinely independent labeling and data-generation process.

Table 1. Temporal holdout classification results

Model	Accuracy	Precision	Recall	F1	Specificity	ROC-AUC
Baseline (amount + time)	0.643	0.000	0.000	0.000	1.000	0.494
Transaction layer	1.000	1.000	1.000	1.000	1.000	1.000
Multi-layer contextual model	1.000	1.000	1.000	1.000	1.000	1.000

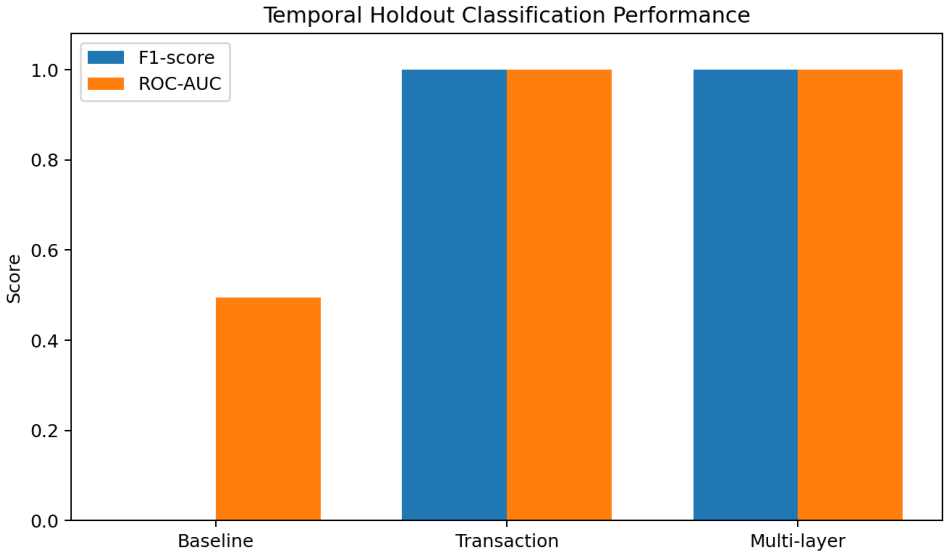


Figure 1. F1-score and ROC-AUC on the October-December 2023 temporal holdout set.

7. LEAKAGE AUDIT AND INTERPRETATION

The feature audit revealed that the type of payment deterministically encoded the laundering label in Dataset I. All Cash transactions in the dataset were labeled as suspicious, as were all Cheque and Cross-Border entries. Meanwhile, Credit Card, Debit Card, E-Wallet, Mobile Payment, and ACH Transfer transactions were all labeled normal. In other words, the data contained 584 suspicious Cash, 654 suspicious Cheque and 587 suspicious Cross-Border records, with no normal examples of

these payment types, while the other payment types contained only normal transactions. By introducing payment type as a feature, the problem was whittled down to a lookup in a synthetic data-generation recipe.

This is an important conclusion, as it reveals the limits of high accuracy for the problem of fraud classification. The perfect result does not show that the model captured sophisticated laundering patterns, but rather that the algorithms identified a feature that deterministically encoded the label in the training set. We posit that this finding is valuable, as it highlights the need to entertain alternative explanations in AI-assisted investigations. Indeed, our results suggest that the artificial intelligence designed for financial-crime detection should pursue an audit of potential target leakage, deterministic proxies, duplicated clients, temporal overlaps, and synthetic generation in the experimental pipeline.

A leakage audit can be performed as an investigation of the data, not a mere technical preprocessing check. Investigators of the model validity should inquire as to how each target label was assigned, when each input field became available, if a field comes from an outcome, and if unusual predictive categories may be artifacts of the dataset. Cross-tabulation of categorical predictors against the label is especially valuable when in a category there are only positive or only negative examples. In this study, the payment-type counts represent direct evidence of such a problem, as for each listed payment channel there is a single class. The same logic applies to numerical cutoffs, identifiers, and post-investigation flags that have more information about the labeling rules than the underlying suspicious conduct.

One valuable diagnostic would be to compare the full feature classifier against variants that exclude potentially deterministic or post-outcome variables, with the chronological holdout left intact. Such an ablation would not, by itself, demonstrate that remaining predictors generalize to actual banking environments, but it would quantify how much each suspected proxy contributes to the measured result. A second check could compare label distributions within categories across independent datasets or data-generation procedures. The present article does not report these additional numerical tests, and so no ablation metrics are asserted here. Instead, the observed payment-channel separation is taken as sufficient evidence that the 100% classification scores should not be understood as robust detection of coordinated wrongdoing.

Table 2. Payment-type label distribution demonstrating deterministic leakage

Payment type	Normal	Suspicious
Cash	0	584
Cheque	0	654
Cross-Border	0	587
Credit Card	625	0
Debit Card	635	0
E-Wallet	588	0
Mobile Payment	628	0
ACH Transfer	699	0

8. EXPERIMENT 2: COORDINATED NETWORK ANALYSIS

8.1 Dataset II and Graph Construction

In the second study, we used the publicly available Fraud-Network-Analysis set for the AML link analysis task. We selected 200 customers (nodes) and 2,000 of their transactions (edges). We constructed a directed graph which contains customers as nodes

and their transactions as edges; we defined the transaction's sender as its source and its receiver as its target. The same sender-receiver pair could appear in multiple transactions; however, we aggregated repeated pairs into unique directed edges for network statistics. The monetary value of transactions was reported for the purposes of transaction-volume calculations.

The network contained our 200 customers and 1,953 unique directed edges. In this network, accounts had approximately 20 total transaction links on average, with a median of 20, a 95th percentile of 27, and a maximum of 31. Furthermore, there were 45 pairs of accounts which had transactions in both directions between them, i.e. the network contained 45 such reciprocated pairs of edges. While it is true that this data does not necessarily represent financial crimes, these are structural properties which should be subject to further research before any conclusions are drawn.

The graph statistics here are properties of the supplied transaction network, not estimates of financial crime rates. The discrepancy between 2,000 transaction records and 1,953 unique directed edges is indicative of some ordered customer pairs transacting multiple times. The mean and median total degree of approximately 20 is suggestive of substantial connectivity between the 200 observed nodes, and a single account having a maximum degree of 31 is singled out. The 45 reciprocal pairs and 290 directed three-node cycles found in this constructed network suggests that there is a degree of diversity of return connectivity patterns among them. These numbers may be used to identify cases for further investigation; however, their interpretation requires context, including expected network structure, common business practices, and the availability of supporting evidence not considered here.

8.2 Exploratory Coordination Score

An exploratory coordination score was calculated to indicate a graph layer's prioritization of accounts. The scores of accounts' total degree, total transaction volume, the number of reciprocal links, and participation in the directed three-node cycles were standardized in z-scores and summarized in sum. This score is proposed as an exploratory ranking, not a criminality label, since the Dataset II does not contain verified laundering labels. Of special interest is the discrimination between the prominent accounts within the network and the accounts suspected to participate in suspicious activity.

The standardization of constituent indicators makes the proposed ranking easily inspectable as it puts the degree, the transaction volume, the number of reciprocal links, and cycle participating of accounts in comparable numerical scale. The sum, however, embeds a substantial design choice: each standardized signal is assumed to contribute the ranking without any estimates of its relation to confirmed unlawful activity. Features which have high degree of association with each other may amplify one structural aspect. We see that for CUST0042 and CUST0011 there is a difference between the rank based on coordination and that which we have for the metadata risk category which in turn puts into relief the need for multiple points of view in the review. Careful application of the put forth score will present out the four component indicators along with the total, will allow the investigator to study the background transaction picture and also will put to rest the practice of classifying a high structural rank as a verified suspicious report.

The issue of stable ranking is a key research issue at hand. That a minor change in the observation period, treatment of duplicate elements, or relative weight of indicators produces a different set of results is what investigators will look to report. In the future a sensitivity analysis which reports on rank agreement across different weight sets and which compares which accounts which maintain rank prominence will be done. Also we will look at which account's score is dominated by high cycle count or large transaction volume. In the absence of ground truth labels we will be testing for ranking consistency and transparency of the used heuristics which we are to present as against the issue of the heuristics' success in identifying criminal groups.

Table 3. Structural characteristics of Dataset II

Network characteristic	Observed value
Transactions	2,000
Customers / active nodes	200
Unique directed edges	1,953
Mean total degree	20

Median total degree	20
95th percentile degree	27
Maximum degree	31
Reciprocal account pairs	45
Directed three-node cycles	290

Table 4. Highest exploratory coordination scores

Account	Degree	Volume	Reciprocal links	3-cycles	Score	Metadata risk
CUST0042	27	15,338.74	3	14	10.678	low
CUST0011	29	16,125.37	3	11	10.345	high
CUST0165	30	18,995.65	1	11	8.696	medium
CUST0137	29	12,229.21	3	10	8.665	medium
CUST0080	28	16,032.03	2	6	6.821	medium
CUST0044	25	13,648.73	2	9	6.37	low
CUST0144	31	11,533.17	1	10	6.055	low
CUST0067	28	10,394.02	2	8	5.641	medium
CUST0012	25	15,004.97	2	5	5.376	medium
CUST0035	25	13,451.84	1	10	5.234	medium

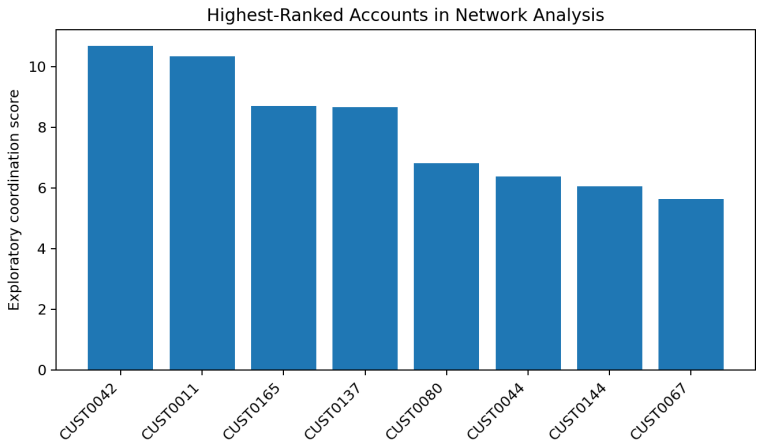


Figure 2. Exploratory coordination score for the eight highest-ranked accounts. The ranking is structural and is not a verified criminality label.

9. DISCUSSION

The two studies present a case for a multi-layered approach. In Dataset I, we had labeled transaction types, but this did not include much in terms of repeated-account structure, which in turn did not allow us to evaluate coordinated networks. In

Dataset II, we had a dense transaction graph but did not have confirmed money-laundering labels to determine true-positive and false-positive network detection rates. To put forth either of these as validation of the entire framework is methodologically inappropriate.

Graph analysis may identify key players which may not be apparent in single transaction reports. Degree is a measure of connection, weighted volume of trade reflects level of exposure, reciprocity indicates two-way traffic and cycles point to closed loops. These we use in combination with separate risk assessment, customer profile analysis, entity resolution accuracy, and study of temporal trends. We did not determine that any one graph metric proves crime; instead, we put forth a model which presents a framework for aggregating data. We put forth a multi-layer approach which does not claim single metrics to be the solution to crime identification but instead we present a method to put the pieces together.

The exploratory results present which accounts may require manual review. CUST0042 had the highest structural coordination score out of all that we looked at, although CUST0042's associated metadata risk was low, while that of CUST0011 was high. We should note that these discrepancies are to be used as a reason to investigate, not as a reason to come to a conclusion. A risk system should put in place what allows investigators to see the reason behind which account is made a priority, which may help distinguish established hubs from possible coordination.

The joint analysis of the experiments results in a less effusive, and thus more useful, conclusion than either result alone. Dataset I highlights the vulnerability of model evaluation in the presence of a deterministic label proxy, and Dataset II highlights the availability of relationship-based evidence in spite of the absence of labels. These observations support the architectural logic of combining transaction, graphical, and temporal information in a single review but do not provide evidence that combining the two datasets would improve detection accuracy. An end-to-end test of the multi-layer framework would involve linked accounts, comparable timestamps across accounts, independent labels and known illicit substructures within the same evaluation population. Before that evidence can be presented, the multi-layer framework should be viewed as a design backed by component level findings.

Practically, a coordinated-crime alert would be most useful when binding a broad risk signal to a focused evidence trail. Network ranking can guide a reviewer as to where they should start looking, but the investigation record should identify the transactions involved, the dates and directions of the flows, the counterparties, and any remaining inconsistencies. In this way, an evidence trail enables efficient review, and reduces the risk that the model's confidence equates to certainty about misconduct. It also provides a foundation for exploring false trails and refining future workflows. No investigator study or case level outcome assessment was undertaken as part of the present research; these points address the practical implications of the proposed architecture rather than the benefits of its adoption.

10. IMPLICATIONS FOR A PRODUCTION MULTI-LAYER SYSTEM

A production setting needs higher quality entity and transaction data than is showcased in the public samples here. Account, customer, device, corporate ownership, IP, beneficiary, merchant, and wallet identifiers could be embedded into a heterogeneous graph within the constraints of legal authorization and data minimization. Relationships should also be tagged with timestamps and transaction attributes to differentiate between established business relationships and transient coordinated activity.

The transaction layer is to leverage calibrated supervised or semi-supervised models which have been trained on leakage-audited labels. The graph layer could make use of engineered structural features, graph neural networks, subgraph classifiers, or temporal graph learning. The temporal layer should focus on fast-moving events, recurring patterns, account creation timing, counterparty rotation, and emerging communities. A risk fusion layer should fuse these various signals together into a calibrated ensemble or meta-model.

Explainability should facilitate in-depth investigation rather than simply model interpretation. For each high-risk network, the system should be able to identify key transactions, associated entities, cycles, shared attributes, and deviations from normal behavior. Investigators will be able to go from a network-level alert to the underlying documents which generated it.

Transaction classification metrics should be differentiated from entity-level and network-level metrics. At the network level, we should consider whether known criminal groups are reconstructed, whether key players are identified, how quickly coordinated

patterns are recognized, and the rate of false-positive investigative leads. Temporal out-of-sample testing is of utmost importance since financial crime typologies evolve over time.

Implementations should translate analytical scores into a clear case-management process. A screening service could take the transaction or network which triggered an alert, identify source records to reconstruct it, and record the model version and feature values at decision time. Investigators could then record whether the apparent anomaly represents expected account behavior, a data-quality issue, an unresolved business relationship, or a pattern worth further inquiry. Feedback from reviewed cases would be valuable to improve prioritization, assuming that investigative outcomes are not assumed to be perfectly complete ground-truth labels. Clear audit trails are especially critical when an alert impacts individuals or organizations, because the supporting evidence must be clearly distinguishable from an automated inference.

11. LIMITATIONS

A significant limitation of the present research is that both experiments are performed on non-confidential, synthetic demonstration datasets, which are potentially unrepresentative of the distribution of records in a production financial system. The excellent classification performance resulting from the inclusion of payment type is a data-generation artifact and must not be extrapolated to operational anti-money laundering performance.

The second limitation is that Dataset I does not include repeated-account network structure, while Dataset II does not include laundering labels. The current work thus evaluates two parts of the framework separately, instead of a single end-to-end detector. A stronger future study would train a detector on a dataset that included repeated entities, transaction-level labels, network-level patterns of laundering, and temporal ground truth.

A third limitation is that the exploratory coordination score is a transparency-inducing structural heuristic, rather than a learned graph model. We applied the method to illustrate risk fusion across multiple network indicators, but did not seek to assert that the highest-ranked accounts are criminals. Future work would explore graph neural networks, temporal graph models, community-based methods and subgraph classifiers on labeled network benchmarks.

A fourth limitation concerns generalizability. Financial institutions differ in their products, populations, jurisdictions, and procedures. Any operational model would require institution-specific calibration, fairness evaluations, privacy controls, governance approvals, and ongoing monitoring.

There are also limitations to the interpretation of the experiments as a combined empirical study. The data sources were independent, which means that the classification outcomes cannot be attributed to the network scores and the graph rankings cannot be validated against Dataset I labels. The exploratory score was not calibrated to a propensity for wrongdoing or compared to an external investigative benchmark. Additionally, the current analysis does not provide an ablation study of each proposed layer, processing costs at production scale, or robustness tests to changing criminal behavior. Notably, we think it is important to acknowledge these limitations: the framework provides a defensible pipeline of analytical checks, but further empirical work is needed to understand its predictive, operational, and institutional utility.

12. FUTURE RESEARCH

Future work should evaluate the architecture on more expressive AML benchmarks, such as the IBM AML data, the AMLSim-synthesized networks, the Elliptic set or institutionally governed transaction graphs, with a specific emphasis on those which contain explicit laundering subgraphs, in order to directly evaluate the network reconstruction rather than rely on general suspicious labels.

A second direction of future work will be the investigation of temporal graph learning. Coordinated financial crimes involve accounts, which are created, abandoned, replaced and attached to new intermediaries. Learning the representations of entities in a coordinated manner over time can be preferable to a static graph, so long as the chronological train/test split is respected.

A third direction of future work concerns privacy-preserving cooperation. Financial crime networks can span multiple institutions, but each institution observes only a subgraph. Federated learning, secure computation, privacy-enhancing techniques, and controlled-release mechanisms may enable cooperative risk identification without unwarranted sharing of customer information between institutions.

Future work should establish auditing and benchmarking protocols for target-leakage detection. Automated audits could identify features that almost determine the label, duplicated records between the train and test sets, test identifiers for leakage, quantify the degree of temporal overlap and compare the performance if the suspiciously predictive variables are removed.

A practical research agenda could present a staged approach to the evaluation of the framework which puts forth full network detection as a single benchmark result. In the first stage we see entity resolution and transaction history reconstruction. In the second stage we put to test a leakage audited transaction baseline against behaviors and structural features which also reports the added value of each component on the same hold out. With verified group labels and laundering sub graphs we could report network level recall, false positive rate which becomes a investigative lead and time to expose a coordinated pattern. The final stage may look at explanation quality, analyst workload, robustness to distribution shift, and privacy or fairness issues. This progression will clarify which improvements are from the use of richer evidence and which from more complex models.

A also has that in the future we require a reproducible report which ties each numerical result to its specific data set release, pre processing version, train test split and model setup. As synthetic benchmarks grow and their generators, random seeds, or sampling methods change we should see that this information is included in reports. That we include in the archival what exact rules were applied to reduce repeated sender receiver and the number of directed cycles in graph analysis, which also includes details on how risk scores were normalized. We should note in what categories our features fall in and which from the trained model set we excluded from the analysis before training. Also if we do this we will enable other researchers to disentangle the model's architecture's performance from the results of the which we use as an example in this study.

13. CONCLUSION

The present study proposed and evaluated a multi-layered AI-based approach to detecting coordinated financial crime networks. The design included transaction, contextual, graph, temporal, explanation, and human review layers that analyzed relational information, rather than isolated transactions.The empirical research reported two complementary findings. In the temporal transaction-classification experiment, the augmented models achieved 100% scores on a synthetic dataset because the payment type deterministically encoded the labels. These results did not provide proof of advanced laundering detection. We also identified an independent 2,000-transaction network which contained observable network structure, including 1,953 distinct directed connections, 45 reciprocal account pairs, and 290 directed three-node cycles. We put forth a transparent coordination score to show how such structural elements can be used to prioritize accounts for review, without treating network prominence as a proof of criminal activity.Therefore, the central issue was not of finding one algorithm that solves financial crime; rather, reliable detection requires multiple layers of evidence and careful validation. Transaction models, behavioral context, graph structure, temporal evolution, and explainability should be made to work together. Also, it is up to the researchers and practitioners to audit data for leakage, maintain temporal separation, distinguish risk signals from criminal findings, and include human oversight. These principles provide a stronger basis for empirical research on AI systems for coordinated financial-crime detection.

REFERENCES

[1] IBM. (2026). AML-Data: Synthetic financial transactions for anti-money-laundering research. GitHub repository. https://github.com/IBM/AML-Data
[2] Suzumura, T., & Kanezashi, H. (2021). AMLSim: Anti-money-laundering datasets and simulator. GitHub repository. https://github.com/IBM/AMLSim
[3] VisionLangAI. (2026). AML Transaction Dataset. GitHub repository. https://github.com/VisionLangAI/AML-Transaction-Dataset
[4] Ric-Royal. (2026). Fraud-Network-Analysis: Fraud and AML link analysis with graph analytics and anomaly detection. GitHub repository. https://github.com/Ric-Royal/Fraud-Network-Analysis
[5] Weber, M., Domeniconi, G., Chen, J., Weidele, D. K. I., Bellei, C., Robinson, T., & Leiserson, C. E. (2019). Anti-money laundering in Bitcoin: Experimenting with graph convolutional networks for financial forensics. KDD Workshop on Anomaly Detection in Finance.
[6] Financial Action Task Force. (2025). Horizon scan: Artificial intelligence and deepfakes. FATF.
[7] Financial Crimes Enforcement Network. (2025). Frequently asked questions regarding suspicious activity reporting requirements. U.S. Department of the Treasury.