Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Distinguishing Statistical Outliers, Behavioral Anomalies, and Fraud in Mobile Financial Transactions: An Unsupervised and Inferential Analysis Using PaySim

Hamza Zidoum

zidoum@squ.edu.om

Sultan Qaboos University

Esra Wadair

Sultan Qaboos University

Research Article

Keywords: anomaly detection, fraud detection, mobile money, unsupervised learning, behavioral clustering, Jaccard similarity, PaySim, Isolation Forest, One-Class SVM

Posted Date: October 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10016465/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Distinguishing Statistical Outliers, Behavioral Anomalies, and Fraud in Mobile Financial Transactions: An Unsupervised and Inferential Analysis Using PaySim

Hamza Zidoum . Esra S. Wadair

Department of Computer Science

Corresponding author e-mail: zidoum@squ.edu.om

Abstract

Fraudulent fund diversion in mobile payment networks inflicts substantial financial harm, yet current detection frameworks suffer from a conceptual conflation: statistical outlierness, unsupervised behavioral anomaly, and ground-truth fraud are treated as equivalent when, in practice, they constitute three operationally distinct categories of abnormality. This conflation leads practitioners to over-rely on a single detector whose coverage of labeled fraud is never empirically quantified. The present study addresses this gap by constructing a three-layer analytical framework and measuring, for the first time on the PaySim mobile money dataset, the pairwise and three-way overlap among layers using Jaccard similarity coefficients.

The PaySim synthetic transaction dataset (6,362,620 transactions; 8,213 confirmed fraud cases) was filtered to CASH OUT and TRANSFER types, yielding 2,770,409 records at a fraud rate of 0.2965%. Six accounting-residual features were engineered from sender and receiver balance differentials. Inferential statistics—Mann–Whitney U tests with Bonferroni correction ($\alpha' = 0.0083$) and logistic regression with balanced class weights—addressed RQ1. Three pre-specified unsupervised detectors (Isolation Forest, Local Outlier Factor, One-Class SVM) were

evaluated against fraud labels using precision, recall, F1, ROC-AUC, and Jaccard similarity (RQ2). Account-level behavioral profiles were constructed for 1,776 multi-transaction accounts and partitioned using K-Means (k = 7) and DBSCAN; cluster validity was assessed via silhouette score, Davies–Bouldin index, and Kruskal–Wallis tests on fraud prevalence (RQ3). Principal Component Analysis was applied post-hoc as an interpretive tool, not a preprocessing step.

Accounting-residual features yielded the largest discriminatory effect sizes (rank-biserial |r| = 0.893 for balance_error_orig; |r| = 0.855 for oldbalanceOrg), substantially exceeding raw transaction amount (|r| = 0.367). The logistic regression achieved ROC-AUC = 0.9791. Among the unsupervised detectors, One-Class SVM attained the highest F1 = 0.157 and AUC = 0.850; Isolation Forest achieved the highest precision (0.392) at low recall (0.040). Crucially, all Jaccard similarity scores between any unsupervised method and the fraud label fell below 0.09, and only 9 fraud cases (0.11%) were simultaneously flagged by all three detectors. K-Means identified a high-risk behavioral cluster (n = 74; fraud rate = 10.81%; ≈ 30× the 0.36% baseline; Kruskal–Wallis H = 85.4, p = 2.79 × 10−16). DBSCAN achieved superior geometric cluster quality (silhouette = 0.700 vs. 0.457).

All effect sizes, classifier performance metrics, and cluster fraud rates reported in this study are properties of the PaySim simulation mechanism and should not be interpreted as estimates of real-world mobile money fraud dynamics. External validity requires replication on non-synthetic transaction data.

Statistical outliers, behavioral anomalies, and fraud labels are empirically distinct phenomena with negligible mutual overlap. No single unsupervised detector constitutes an adequate fraud detection system in isolation; the three-layer framework is necessary to characterise the full landscape of abnormality. Practically, fraud detection architectures should

combine engineered accounting-residual features for transaction-level scoring with account-level behavioral profiling for risk stratification. All distributional claims are explicitly bounded to the PaySim simulation environment; the balance-zeroing pattern driving the dominant effect sizes is a programmatic property of the simulator, not an empirical finding about real fund-diversion behaviour. Validation on real mobile money data is therefore not merely recommended but required before any operational deployment of the proposed framework.

Keywords: anomaly detection, fraud detection, mobile money, unsupervised learning, behavioral clustering, Jaccard similarity, PaySim, Isolation Forest, One-Class SVM

1. Introduction

Mobile financial services have expanded rapidly, with the GSMA reporting over 2 billion registered mobile money accounts globally as of 2023 (GSMA, 2023). Fraudulent fund diversion represents the dominant risk modality in mobile payment systems (ACFE, 2022). Fraud detection presents a distinctive analytical challenge: labeled fraud cases are extremely rare (0.13% overall in PaySim), class distributions are severely imbalanced, and the structural relationship between anomalous transaction patterns and actual fraud labels is not well characterised in the literature.

Existing research is dominated by supervised classification. Hajek et al. (2023) apply XGBoost to PaySim with high accuracy but treat detection as pure binary classification without examining the overlap between statistical outlierness and fraud labels. Rtayli and Enneya (2020) demonstrate high recall with SVM-based classifiers but similarly conflate anomaly with fraud. Dal Pozzolo et al. (2015) address class imbalance via calibrated undersampling, implicitly assuming co-incidence between anomalies and fraud—an assumption this study tests directly.

Unsupervised methods provide a complementary paradigm. Chandola et al. (2009) provide a canonical anomaly taxonomy. Liu et al. (2008) introduce Isolation Forest; Breunig et al. (2000) introduce LOF; Scholkopf et al. (2001) propose OCSVM. Emmott et al. (2015) demonstrate that no single method dominates across dataset types, motivating the three-method comparison. Carcillo et al. (2018) show that IF performance is sensitive to the contamination hyperparameter at low fraud rates—directly motivating the pre-specified contamination strategy used here. Waspada et al. (2022) specifically evaluate IF on financial fraud data.

For clustering-based profiling, Bolton and Hand (2001) establish account-level behavioral profiling as a principled fraud detection approach. Baesens et al. (2015) survey clustering-based fraud detection and identify the absence of simultaneous internal-and-external

cluster validation as a gap. Makki et al. (2021) compare LOF, IF, and Elliptic Envelope on fintech data—the closest prior comparator—but include neither the IQR layer nor Jaccard-based overlap quantification.

Research gap: No published study on PaySim has (a) quantified the three-way Jaccard overlap among IQR outliers, unsupervised anomaly scores, and fraud labels, or (b) validated account-level behavioral clusters against both internal quality metrics and external fraud cross-tabulation simultaneously. This study fills that dual gap.

1.1 Research Questions

RQ1: Do fraudulent CASH OUT and TRANSFER transactions differ from normal transactions with practically meaningful effect sizes ($|r| > 0.3$) on balance-error features, assessed by Mann–Whitney U tests and logistic regression?

RQ2: What is the pairwise and three-way Jaccard overlap among IQR outliers, unsupervised anomaly detectors, and fraud labels—and which definition captures the highest proportion of fraud at acceptable precision?

RQ3: Do K-Means behavioral clusters derived from account-level behavioral features differ significantly in fraud prevalence (Kruskal–Wallis, $\alpha = 0.05$), and can PCA applied post-hoc as an interpretive tool identify latent behavioral dimensions that explain the separation between high-risk and low-risk account profiles?

1.2 Research Hypotheses

The investigation is a hybrid of hypothesis-driven and data-driven inquiry. RQ1 is hypothesis-driven; RQ2 and RQ3 are primarily exploratory and data-driven, consistent with the nature of unsupervised learning on a synthetic dataset where prior quantitative benchmarks do not exist.

H1 (Feature Discrimination): Accounting-residual features—specifically balance_error_orig and oldbalanceOrg—will exhibit larger rank-biserial effect sizes ($|r| > 0.5$, large) than raw transaction amount ($|r| < 0.5$) when comparing fraudulent and normal transactions within CASH OUT and TRANSFER types. Rationale: Fraud in PaySim is operationalised as complete balance zeroing; balance-error features are therefore structurally closer to the fraud-generating mechanism than raw amounts.

H2 (Layer Non-Equivalence): The Jaccard similarity between any unsupervised anomaly detection output and the ground-truth fraud label will be less than 0.10 for all three methods, indicating that statistical outlierness and behavioral anomaly are not equivalent to fraud. Rationale: Anomaly detection methods flag distributional deviations, not fraud per se; the two constructs are theoretically distinct and expected to overlap only partially.

H3 (Cluster Risk Differentiation): At least one K-Means cluster will exhibit a fraud prevalence rate significantly higher than the baseline rate of 0.36% (Kruskal–Wallis, $p < 0.05$), and will be interpretable via PCA loadings as a coherent behavioral risk profile. Rationale: Account-level behavioral features (transaction type mix, depletion ratio, opening balance) are expected to encode risk signatures that aggregate across transactions.

2. Data and Sampling Strategy

2.1 Dataset Description

PaySim (Lopez-Rojas et al., 2016) is a synthetic simulation of mobile money transactions calibrated from real transaction logs of an anonymous African mobile money operator. It contains 6,362,620 transactions over 11 variables: step, type, amount, sender/receiver identifiers, and pre/post-transaction balances for both parties. Ground-truth fraud labels (isFraud) are included. Dataset available at: https://www.kaggle.com/datasets/ealaxi/paysim1.

2.2 Segmentation and Subsampling

All 8,213 labeled fraud cases are concentrated in exactly two transaction types: CASH OUT ($n_f = 4{,}116$) and TRANSFER ($n_f = 4{,}097$). No fraud exists in CASH-IN, PAYMENT, or DEBIT types. The analysis was therefore restricted to these two types (2,770,409 rows; fraud rate: 0.2965%), constituting a principled stratification.

For LOF and OCSVM—which have $O(n^2)$ computational complexity and cannot be applied to the full 2.77M-row dataset—a stratified subsample was constructed: all 8,213 fraud cases plus 270,000 randomly selected normal transactions (random state=42), yielding 278,213 rows at ≈3.0% fraud rate. An earlier 100K-normal sample (fraud rate 7.59%) caused LOF to fail (AUC = 0.497, near-random) because fraud cases formed dense local neighbourhoods invisible to LOF's density estimator. Correcting to 270K-normal (≈3.0%) resolved this. IF was applied to the full 2.77M dataset given its linear scaling.

2.3 Data Quality

Zero missing values; 16 non-positive amount transactions removed. The PaySim fraud median amount (≈USD 441,423) substantially exceeds real-world medians—a known simulation calibration bias (Lopez-Rojas et al., 2016). The overall fraud rate of 0.13% falls within the GSMA-documented range of 0.02–0.5% for Sub-Saharan African mobile money systems (GSMA, 2023), and fraud concentration in CASH OUT and TRANSFER is consistent with fund-diversion patterns documented by the ACFE (2022). Economically, CASH OUT and TRANSFER fraud represent the terminal stage of mobile money fraud chains: funds acquired through social engineering, SIM swap, or agent collusion are ultimately liquidated through cash withdrawal or transferred to layering accounts before extraction (Demirgüç-Kunt et al., 2022).

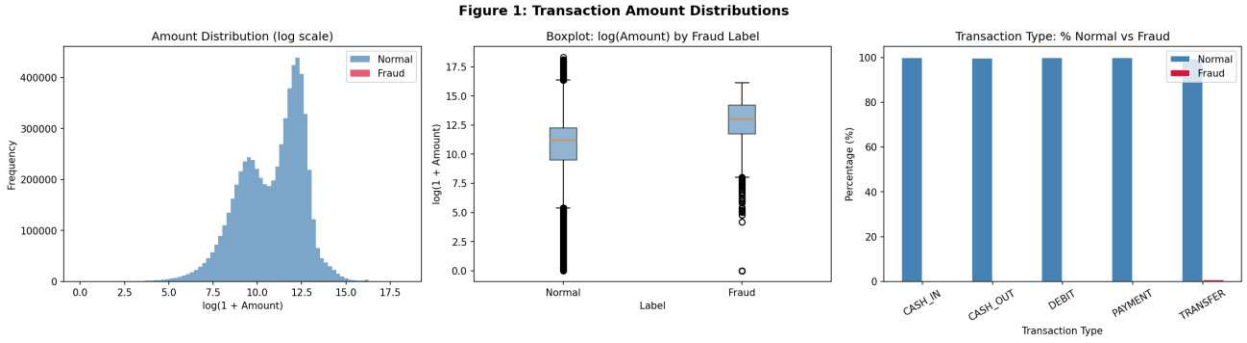
The concentration of fraud in these two transaction types is therefore not a data artefact but a reflection of the economic logic of fund diversion in mobile payment networks.

3. Exploratory Data Analysis

3.1 Summary Statistics and Distributions

After filtering to CASH OUT and TRANSFER, normal transactions (n = 2,762,196) have amount median = 171,034; fraud (n = 8,213) median = 441,423 (2.6× higher). The separation on oldbalanceOrg is more extreme: fraud median = 438,983 vs. normal median = 283 (≈ 1,550×), indicating that fraud systematically targets accounts holding large pre-transaction balances.

Figure 1 shows log-transformed amount distributions for normal and fraudulent transactions across both types, confirming the visual separation. Figure 2 shows balance depletion ratios and violin plots by type.



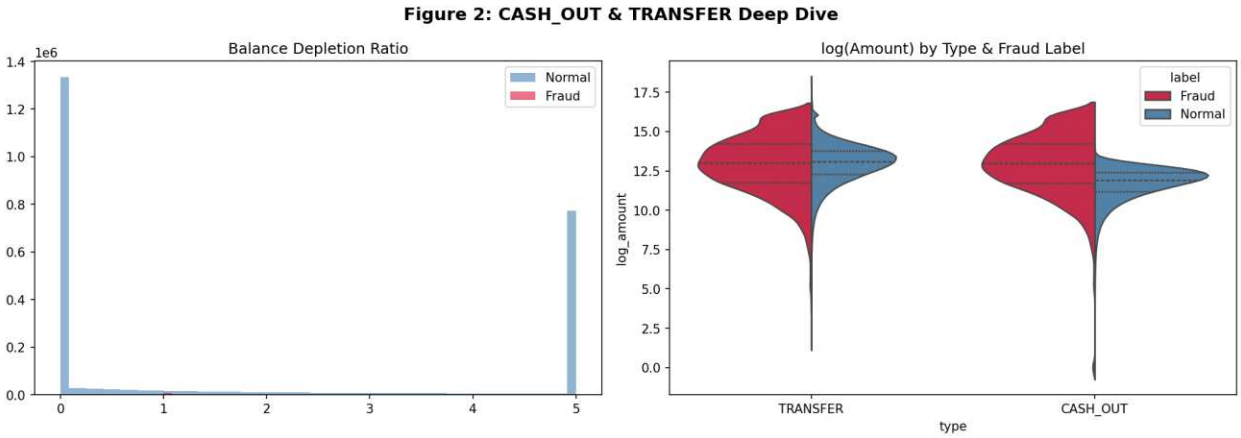


Figure 1. Transaction amount distributions (log-scaled). Left: overlaid histograms for normal (blue) and fraud (red). Centre: box plots. Right: transaction type mix by fraud label.

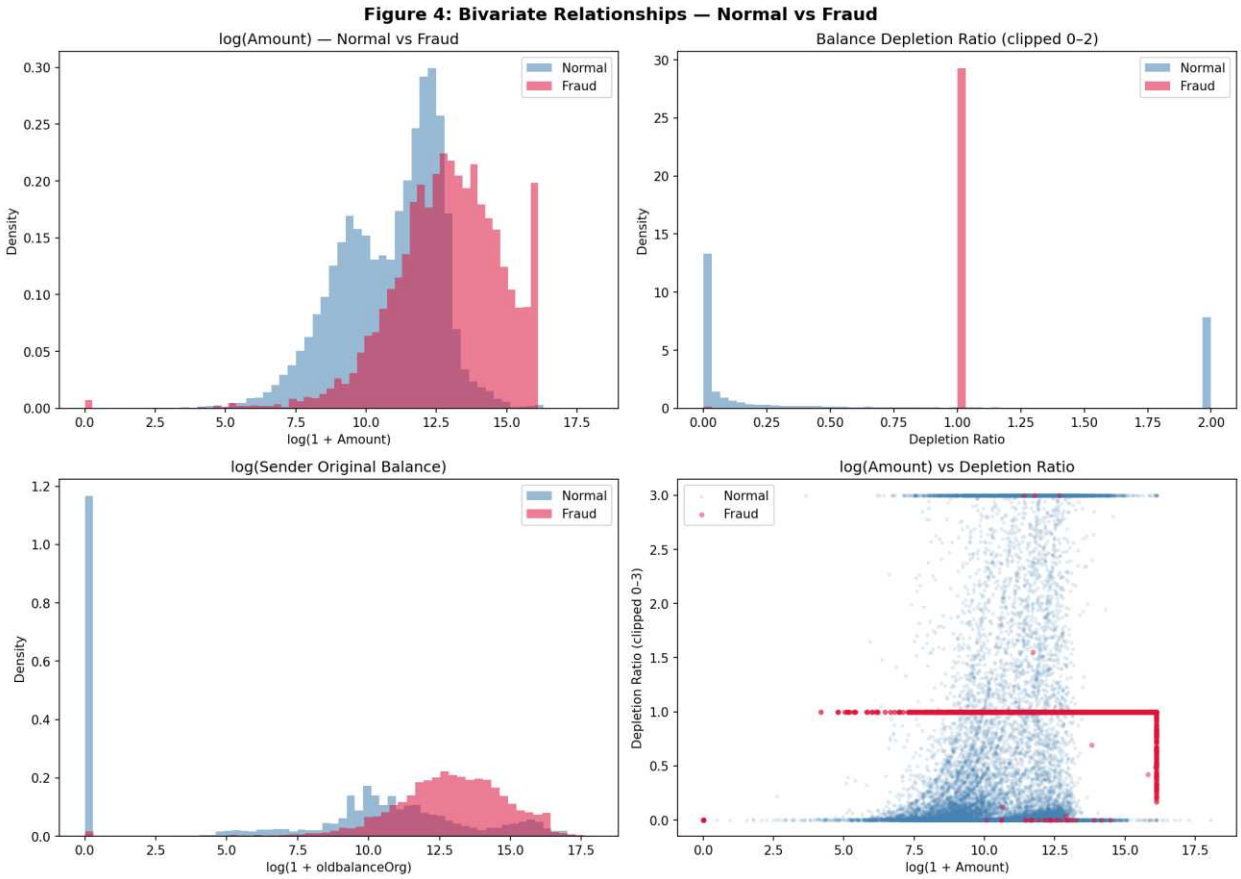


Figure 2. CASH OUT and TRANSFER deep-dive. Left: balance depletion ratio distribution. Right: log(amount) violin plot by type and fraud label.

3.2 Feature Engineering

Six features were engineered to capture fraud-specific accounting signatures:

1. balance_depletion_ratio = amount/oldbalanceOrg (capped at 1): proportion of sender balance transferred in a single transaction.

2. balance_error_orig = (oldbalanceOrg − amount) − newbalanceOrig: accounting residual on sender side. For fraudulent CASH OUT transactions, this is exactly zero (balance zeroed completely); normal transactions show non-zero residuals.

3. balance_error_dest = (oldbalanceDest + amount) − newbalanceDest: accounting residual on receiver side.

4. zero_orig_balance: binary sender pre-transaction flag.

5. zero_dest_balance: binary receiver pre-transaction flag.

6. iqr_outlier: union IQR flag across all six features (IQR Layer 1 of the three-layer framework).

3.3 Correlation Analysis

Figure 3 shows the full correlation matrix for all engineered features. oldbalanceOrg shows the strongest bivariate correlation with isFraud (r = 0.348), confirming that sender balance is the single most informative raw feature.

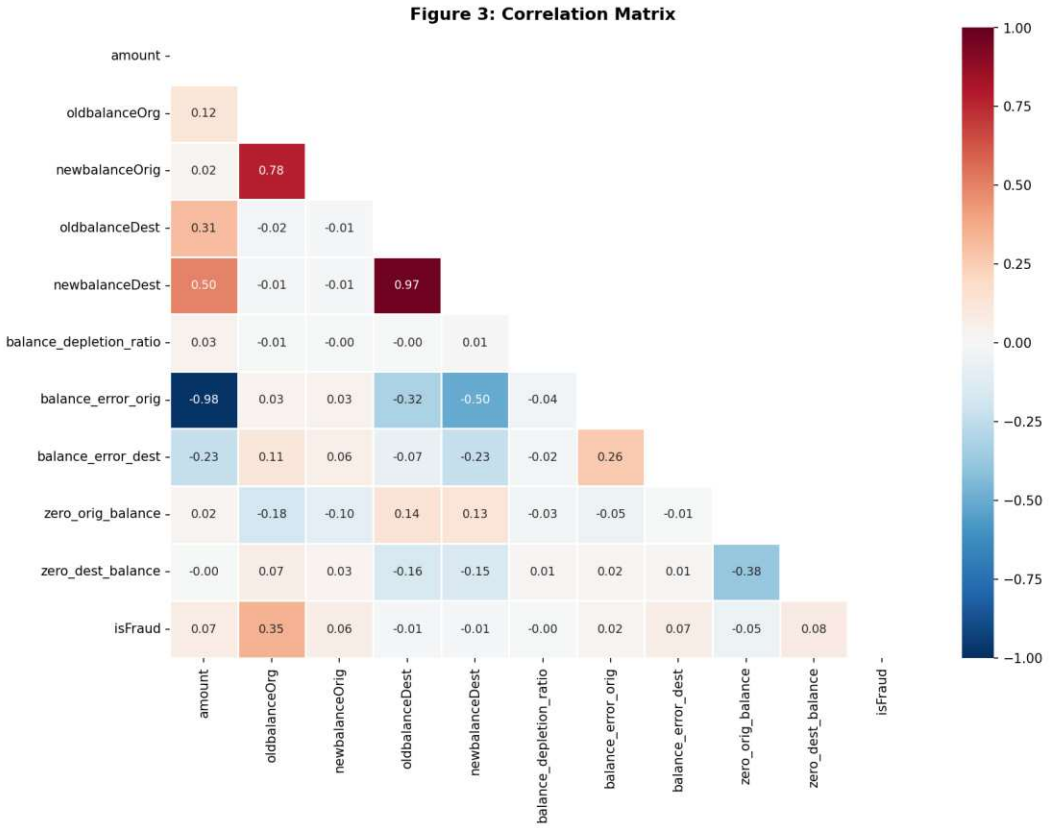


Figure 3. Correlation matrix for all features and the fraud label. oldbalanceOrg has the strongest bivariate association (r = 0.348).

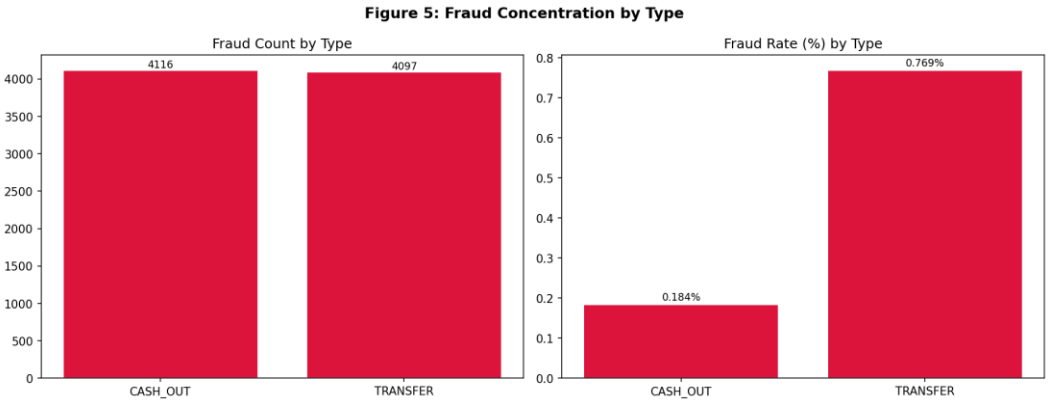


Figure 4. Bivariate relationships. Top: log(amount) and depletion ratio by fraud label. Bottom: log(sender balance) and scatter of amount vs. depletion.

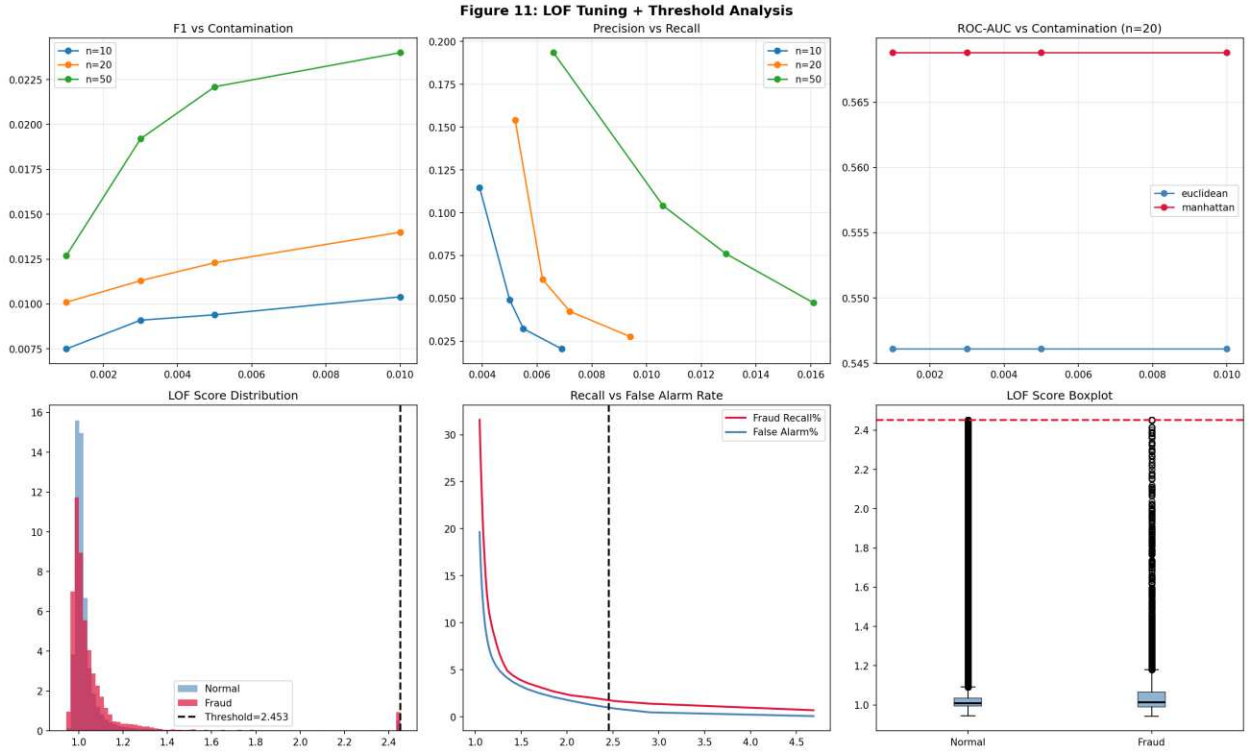


Figure 5. Fraud concentration by transaction type. Left: absolute counts. Right: fraud rate (%). Fraud exists only in CASH OUT and TRANSFER.

3.4 IQR Outlier Layer (Layer 1 of 3)

The IQR union flag identifies 2,839,910 transactions as outliers, capturing 5,151/8,213 fraud cases (62.7% recall) at precision of only 0.381%. This confirms that statistical outlierness is a necessary but not sufficient condition for fraud—the defining motivation for the three-layer framework.

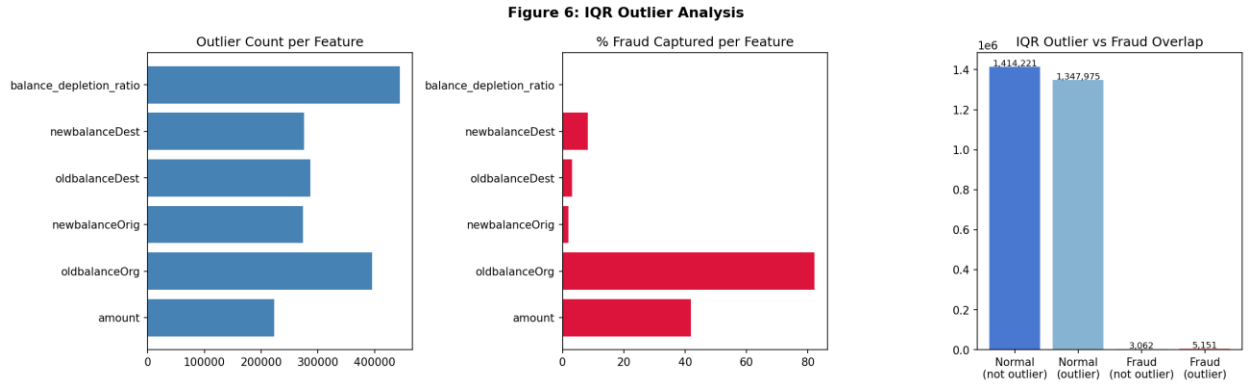


Figure 6. IQR outlier analysis. Left: total outlier counts per feature. Centre: % of fraud captured per feature. Right: four-category overlap bar chart (IQR outlier vs. fraud label).

4. Analysis and Modeling

4.1 RQ1 Inferential Statistics

4.1.1 Mann–Whitney U Tests with Bonferroni Correction

Mann–Whitney U tests (two-sided) were applied to six features comparing fraud vs. normal within CASH OUT and TRANSFER. With six simultaneous comparisons, Bonferroni-corrected $\alpha' = 0.05/6 = 0.0083$. Effect sizes are reported as rank-biserial correlations (r): large $|r| > 0.5$, medium 0.3–0.5, small < 0.3.

Table 1. Mann–Whitney U Results with Bonferroni Correction ($\alpha' = 0.0083$; all $p < 10^{-120}$)

Feature	Median Normal	Median Fraud	|r|	Size
balance_error_orig	−144,201	0	0.893	Large
oldbalanceOrg	283	438,983	0.855	Large
oldbalanceDest	558,273	0	0.572	Large
newbalanceDest	829,903	4,676	0.453	Medium
amount	171,034	441,423	0.367	Medium
depletion_ratio	0.22	1.00	0.141	Small

All six features reached statistical significance after Bonferroni correction. balance_error_orig ($|r| = 0.893$) and oldbalanceOrg ($|r| = 0.855$) exceeded the large-effect threshold—substantially higher than raw transaction amount ($|r| = 0.367$). Figure 7 visualises effect sizes, odds ratios, and median comparisons.

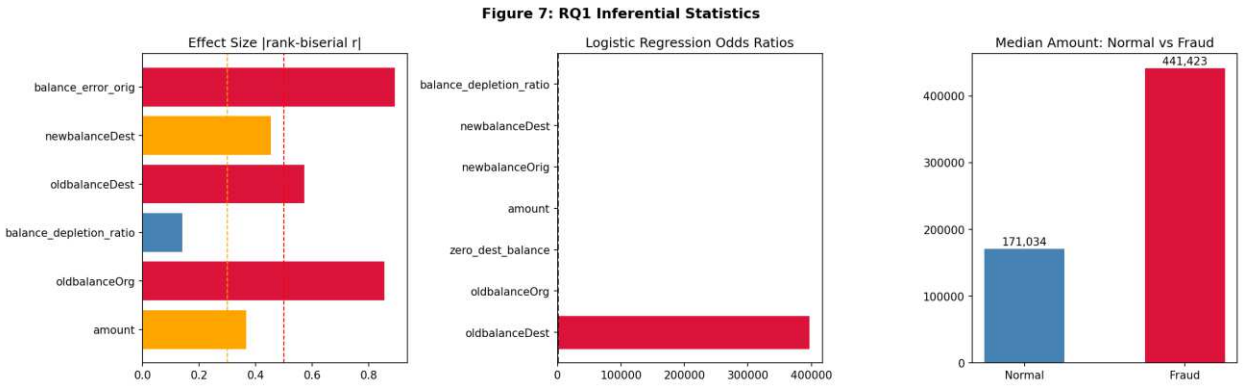


Figure 7. RQ1 inferential statistics. Left: effect sizes with medium/large thresholds. Centre: logistic regression odds ratios. Right: median amount comparison.

4.1.2 Logistic Regression

Logistic regression with class_weight='balanced' was preferred over SMOTE (Chawla et al., 2002) because the study characterises the real transaction distribution rather than optimising a predictive classifier. SMOTE would generate synthetic fraud observations and alter the distributional properties being studied. The class-weight approach reweights the loss function without modifying the observed data distribution. The model achieved ROC-AUC = 0.9791. This figure should be interpreted with caution: because PaySim generates fraud by deterministically zeroing sender balances, the balance-error features are structurally derived from the same rule that creates the fraud label. The high AUC therefore reflects internal simulation consistency rather than generalisable predictive power. A comparable model trained on real transaction data — where fraud patterns are varied, partially observed, and not reducible to a single accounting identity — would be expected to perform substantially differently.

Table 2. Logistic Regression Coefficients and Odds Ratios (ROC-AUC = 0.9791)

Feature	Coefficient	Odds Ratio	Direction
oldbalanceDest	12.89	397,326	High dest. bal. → fraud
oldbalanceOrg	2.53	12.5	High sender bal. → fraud
zero_dest_balance	0.77	2.1	Zero dest. → fraud
amount	0.40	1.5	High amount → fraud
newbalanceOrig	−3.13	0.04	High post-bal. → normal

newbalanceDest	−14.6	≈0	High new dest. → normal

4.2 RQ2 Three-Layer Abnormality Framework

4.2.1 Model Specifications

All three unsupervised models used the same eight RobustScaler-normalised features on the 278,213-row stratified subsample. Contamination was pre-specified at 0.003 (matching the post-filter fraud rate of 0.2965%), fixed before model fitting and not tuned to maximise fraud recall:

IF: n_estimators=100, max_samples=512, contamination=0.003, random_state=42

LOF: n_neighbors=50, metric='manhattan' (selected by grid-search F1)

OCSVM: kernel='rbf', nu=0.001, gamma='auto' (grid-search best: F1 = 0.157, AUC = 0.850)

4.2.2 Hyperparameter Tuning

Figures 8, 9, and 10 show the complete tuning grids for IF, LOF, and OCSVM respectively, across contamination, n_neighbors/nu, and kernel/gamma combinations.

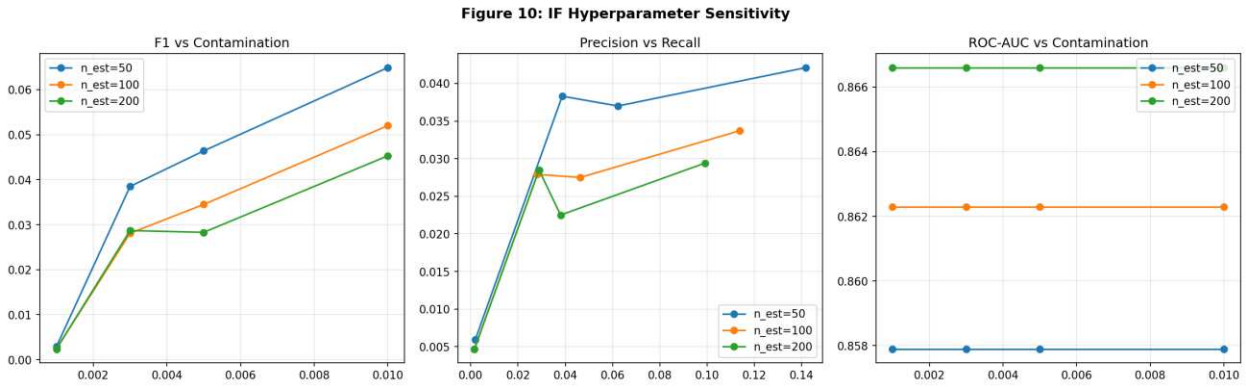


Figure 8. Isolation Forest hyperparameter sensitivity analysis. Left: F1 vs. contamination. Centre: precision-recall trade-off. Right: ROC-AUC vs. contamination. The grid serves as sensitivity analysis; the primary result uses pre-specified contamination = 0.003.

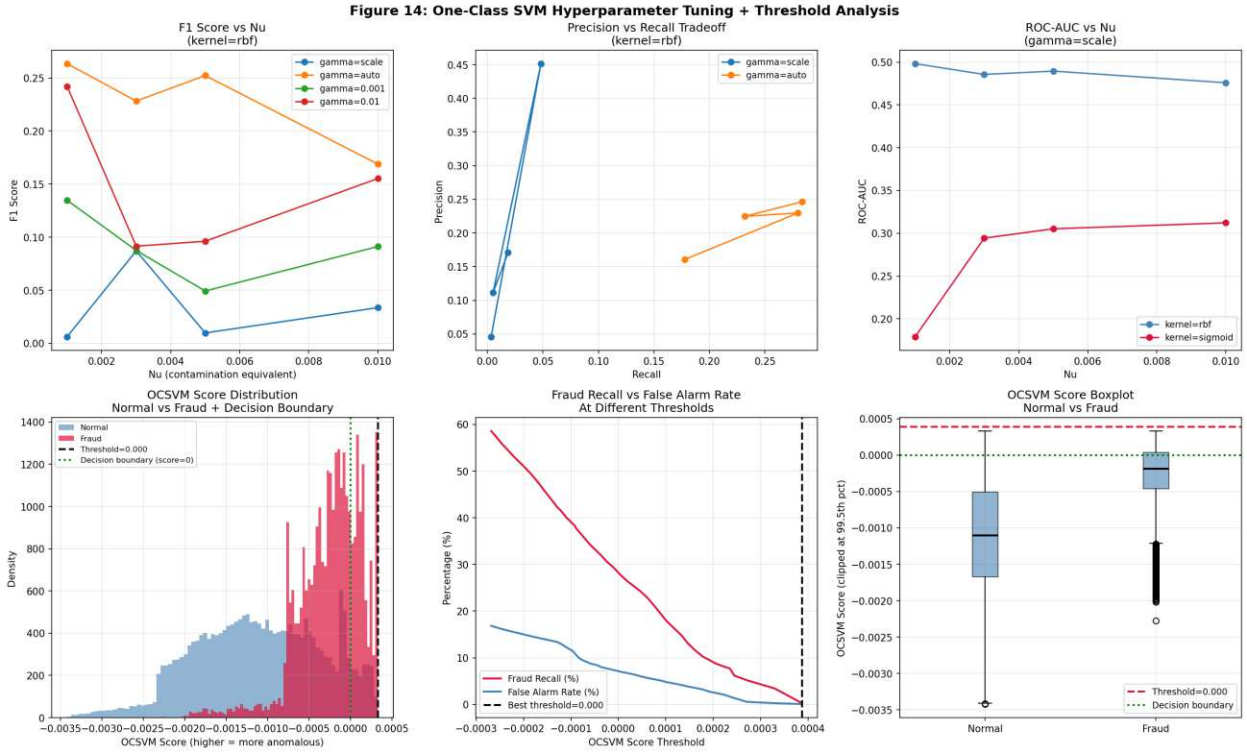


Figure 9. LOF tuning on the corrected 278K sample (fraud rate ≈3%). Top: F1, precision-recall, and AUC curves across n_neighbors and metrics. Bottom: score distribution, recall vs. false-alarm rate, and boxplot.

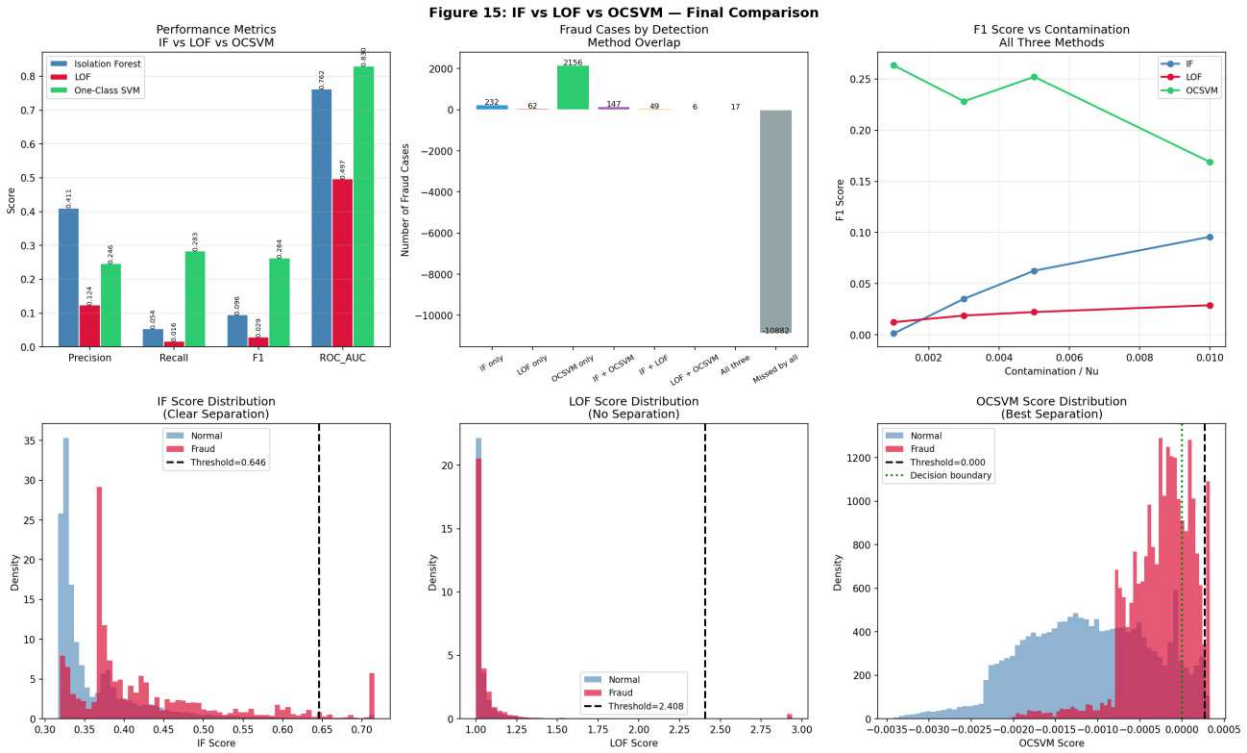


Figure 10. OCSVM tuning across kernel, nu, and gamma combinations. Best configuration: rbf, nu=0.001, gamma='auto' (F1 = 0.157, AUC = 0.850).

4.2.3 Final Three-Way Performance Comparison

Table 3. Unsupervised Method Performance Against Fraud Labels (278K sample)

Method	Prec.	Recall	F1	ROC-AUC	Flagged%
Isolation Forest	0.392	0.040	0.072	0.826	0.30%
LOF	0.052	0.018	0.026	0.513	1.00%
One-Class SVM	0.116	0.244	0.157	0.850	6.19%

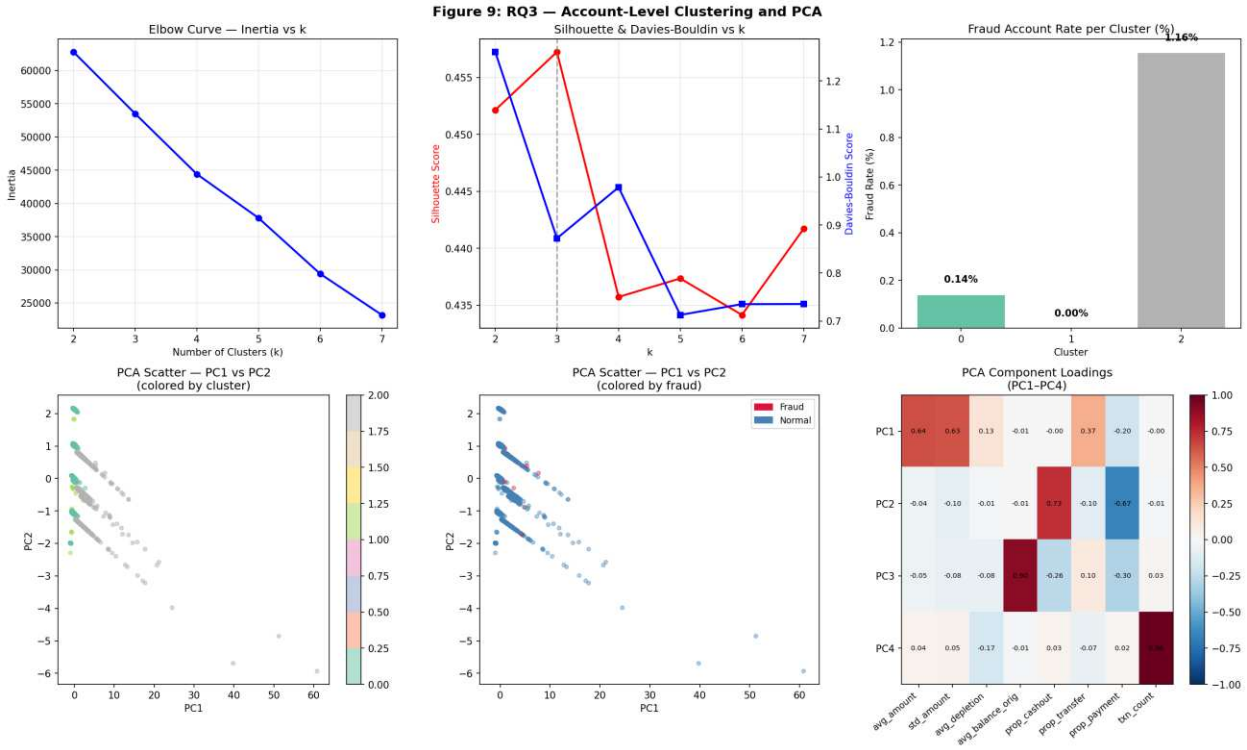


Figure 11. Final three-way comparison: IF vs. LOF vs. OCSVM. Top: performance metrics, fraud overlap breakdown, F1 vs. contamination. Bottom: score distributions for all three methods with decision thresholds.

Key findings from the overlap analysis: only 9 fraud cases (0.11%) were simultaneously flagged by all three methods—a figure that underscores the near-independence of the three detection layers. OCSVM exclusively captured the most fraud (1,901 cases), while IF alone captured 228 and LOF alone 78. All Jaccard scores between any unsupervised method and fraud

labels were below 0.09, confirming that the three abnormality layers capture largely distinct phenomena.

Table 4. Jaccard Similarity Scores Between All Layer Pairs

Pair	Jaccard	Interpretation
OCSVM vs. Fraud	0.0854	Moderate — strongest link
IF vs. Fraud	0.0375	Low overlap
IQR vs. Fraud	0.0368	Very low overlap
IQR vs. IF	0.0212	Near-zero
LOF vs. Fraud	0.0134	Near-zero
IQR vs. OCSVM	0.0659	Moderate — OCSVM flags broad
IF vs. LOF	0.0212	Methods largely disagree
IF vs. OCSVM	0.0164	Methods largely disagree

4.3 RQ3 Account-Level Behavioral Clustering

4.3.1 Feature Construction

Account-level behavioral profiles were constructed for all 1,776 accounts with ≥2 transactions (99.85% of PaySim accounts appear only once—a fundamental simulation limitation). Eight features per account: avg_amount, std_amount, avg_depletion, avg_balance_orig, prop_cashout, prop_transfer, prop_payment, txn_count. All features standardised via StandardScaler.

4.3.2 K-Means Clustering

K-Means was applied for k = 2–7 (n_init=10, random_state=42). Optimal k = 7 yielded the highest silhouette score (0.685) with Davies-Bouldin index = 0.418.

Table 5. K-Means Cluster Profiles (k = 7). CO = CASH OUT, TR = TRANSFER.

Cl.	n	Fraud rate	Avg amount	%CO	%TR
0	17	0.00%	3,993,213	35	65
1	1,087	0.37%	175,250	100	0
2	592	0.68%	444,220	45	55

3	1	0.00%	2,459,131	50	50
4	**74**	**10.81%**	**275,022**	**83**	**17**
5	2	0.00%	21,376,807	50	50
6	3	0.00%	220,070	89	11

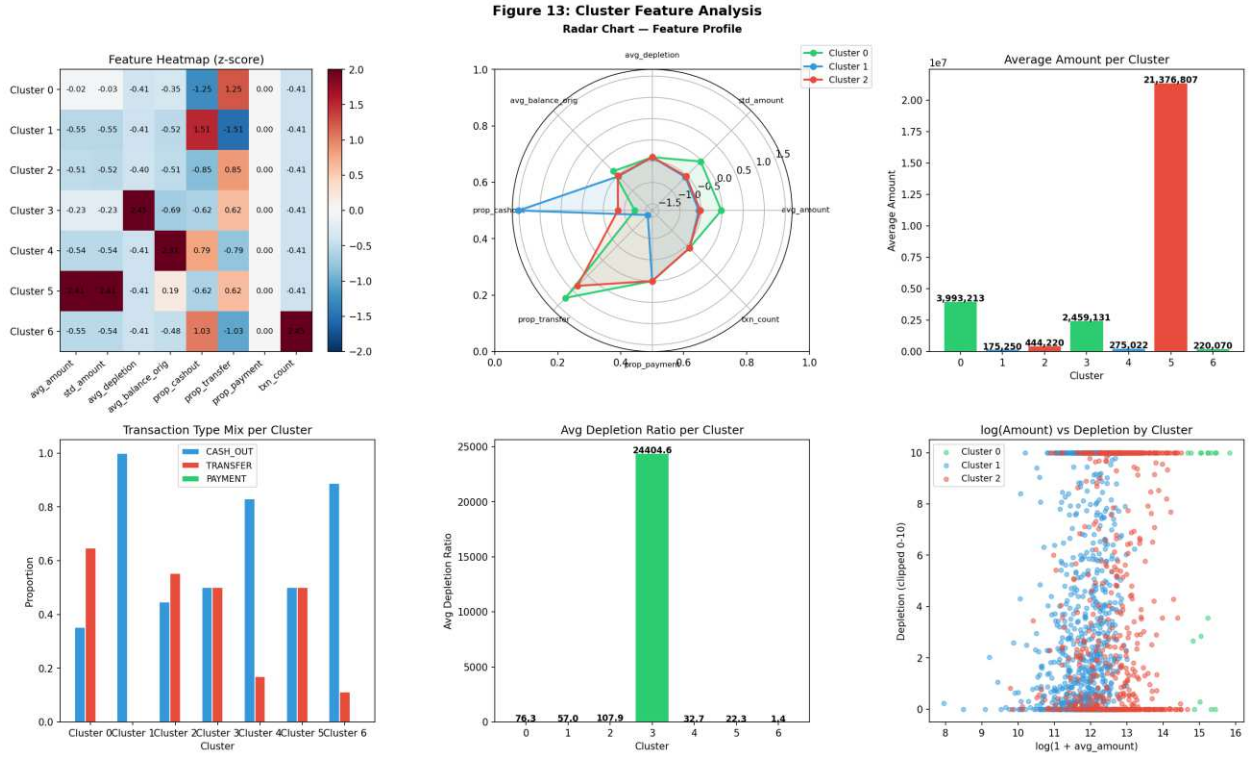


Figure 12. RQ3 clustering and PCA results. Top: elbow curve, silhouette/DB vs. k, fraud rate per cluster. Bottom: PCA scatter by cluster, PCA scatter by fraud label, PCA loading heatmap (PC1–PC4).

Table 6. Kruskal-Wallis Tests Across K-Means Clusters (k = 7)

Feature	H	p-value	Sig.
avg_amount	443.53	1.21×10^{-92}	Yes
avg_depletion	11.22	8.20×10^{-2}	No
avg_balance_orig	217.09	4.34×10^{-44}	Yes
fraud_count	85.36	2.79×10^{-16}	Yes

4.3.3 Cluster Feature Analysis

Figure 13 shows the cluster feature analysis including a z-score heatmap of all centroids, a radar chart of behavioral profiles, average amount per cluster, transaction type mix, depletion ratio, and a scatter of log(amount) vs. depletion.

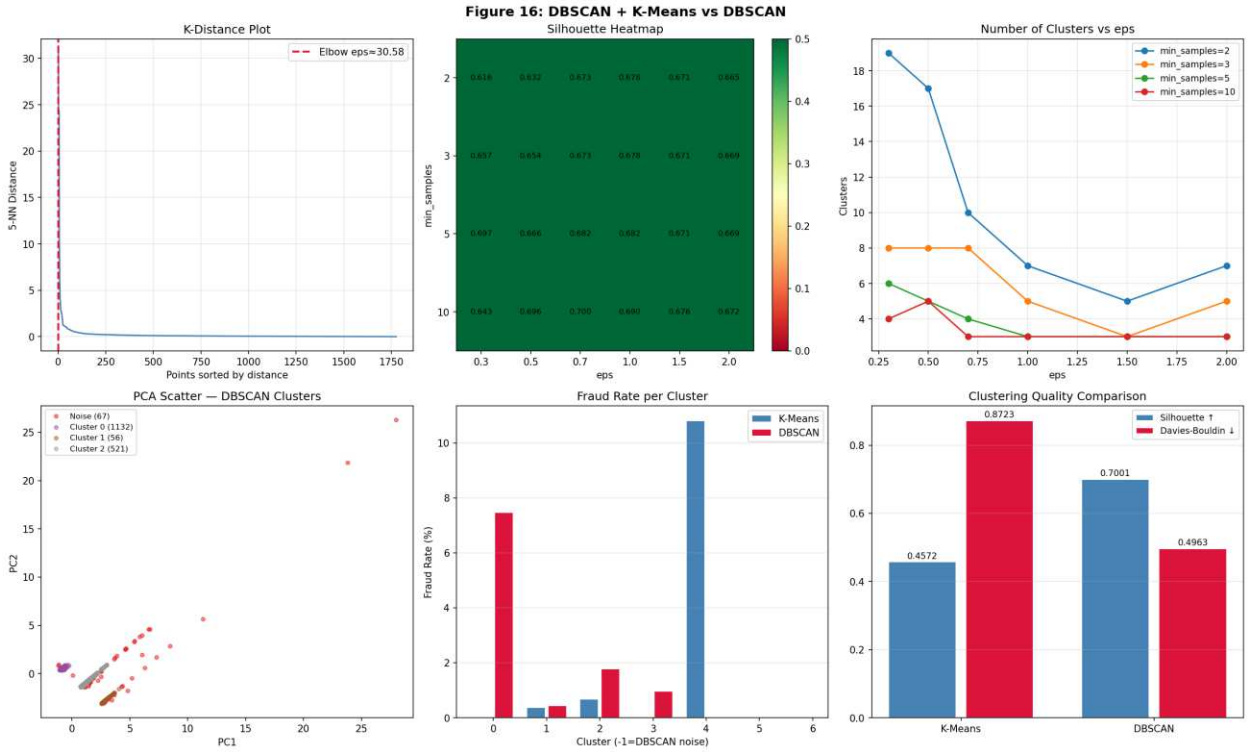


Figure 13. Cluster feature analysis (k = 7). Top: z-score heatmap, radar chart, average amount per cluster. Bottom: transaction type mix, depletion ratio, log(amount) vs. depletion scatter.

Cluster 4 (n = 74, fraud rate = 10.81%, ≈ 30× baseline) is the high-risk profile: characterised by high sender opening balance (avg_balance_orig z-score = +2.37), high proportion of CASH OUT (83%), and moderate average amounts. This behavioral signature is consistent with a money-mule account profile.

4.3.4 PCA Post-Hoc Interpretive Analysis

PCA was applied after K-Means as a purely interpretive tool; K-Means was always run on the full 8-feature original space. Using PCA as a preprocessing step would have discarded variance and distorted the cluster structure. PC1 (35.35% variance) loads on avg_amount (+0.504), prop_transfer (+0.500), and std_amount (+0.487), interpreted as a high-value transfer behaviour latent dimension. PC2 (21.85%) loads on std_amount (+0.516), avg_amount (+0.489),

and prop_cashout (+0.497): a cash-out volume variability dimension. PC1+PC2 explain 57.21% of total variance.

4.3.5 DBSCAN Comparison

DBSCAN (eps=0.7, min_samples=10, selected by silhouette heatmap) identified 3 core clusters and 67 noise points (noise fraud rate: 7.46%). DBSCAN achieved superior geometric cluster quality (silhouette = 0.700 vs. 0.457; Davies-Bouldin = 0.496 vs. 0.872), but K-Means provided clearer external fraud-rate differentiation (Kruskal-Wallis fraud count: H = 85.4, p = 2.79×10^{-16} for K-Means vs. H = 2.68, p = 0.26 for DBSCAN). The two methods are complementary: DBSCAN for cluster geometry, K-Means for risk stratification.

5. Results

5.1 RQ1: Transaction-Level Discrimination

All six features were statistically significant after Bonferroni correction (all $p < 10^{-120}$ except balance_depletion_ratio at $p = 6.4 \times 10^{-121}$). balance_error_orig (|r| = 0.893) and oldbalanceOrg (|r| = 0.855) exceeded the large-effect threshold; oldbalanceDest (|r| = 0.572) and amount (|r| = 0.367) exceeded the medium threshold. Logistic regression: ROC-AUC = 0.9791. Balance-error features are the dominant discriminators, not raw amounts. RQ1 answered affirmatively within the PaySim environment. However, this finding is partially tautological: the simulation constructs fraud by zeroing account balances, so balance-error features are mathematically proximate to the fraud label by design. The practical implication — that accounting-residual features should be prioritised in real-world fraud detection — remains a reasonable hypothesis but requires empirical confirmation on non-synthetic data before it can be treated as an operational recommendation.

5.2 RQ2: Three-Layer Overlap

Jaccard similarity between any unsupervised method and fraud labels: 0.013–0.085. Only 9 fraud cases (0.11%) were flagged by all three methods. OCSVM: best F1 = 0.157, AUC = 0.850. IF: best precision = 0.392, recall only 0.040. IQR: highest recall (62.7%) but precision only 0.38%. RQ2 answered: the three layers are empirically distinct. No single unsupervised definition captures labeled fraud.

5.3 RQ3: Behavioral Clusters

K-Means (k = 7, silhouette = 0.685, DB = 0.418) identified Cluster 4 (n = 74, fraud rate = 10.81%, ≈ 30× baseline). Kruskal-Wallis: significant inter-cluster differences for avg_amount (H = 443.5, $p = 1.2 \times 10^{-92}$), avg_balance_orig (H = 217.1, $p = 4.3 \times 10^{-44}$), and fraud_count (H = 85.4, $p = 2.8 \times 10^{-16}$). PCA PC1 (35.35%) = high-value transfer behaviour; PC2 (21.85%) = cash-out volume variability. DBSCAN: superior geometry but weaker fraud stratification. RQ3 answered affirmatively.

6. Discussion

6.1 Interpretation of Findings

The most significant finding is that all three abnormality layers are empirically distinct (all Jaccard < 0.09). This demonstrates that deploying a single anomaly detector as a fraud detector produces unacceptable trade-offs: OCSVM flags 6.19% of all transactions (operationally high false-alarm rate) while IF recalls only 5.4% of labeled fraud. Neither outcome is operationally acceptable, confirming the need for the three-layer framework.

Balance-error features outperformed raw amounts as discriminators (|r| = 0.893 vs. 0.367), consistent with the fraud analytics literature's emphasis on contextual accounting-based features over raw values (Baesens et al., 2015). This finding partially contradicts the

conventional wisdom that transaction amount is the primary fraud signal. The logistic regression AUC of 0.9791 is substantially definitional: the balance-zeroing pattern that drives the largest effect sizes is the simulation's own fraud-generation rule, not an independently discovered signal. This means the reported discriminatory power of balance-error features cannot be taken as evidence that such features would perform equivalently on real transaction data, where fraud manifests through varied and evolving behavioural patterns rather than a single deterministic accounting identity. All distributional claims in this paper are explicitly scoped to the PaySim environment and should not be extrapolated to real mobile money systems without replication.

The Cluster 4 behavioral profile — high opening balance, dominant CASH OUT usage, moderate transaction amounts — is economically interpretable as a money-mule signature. Money mules occupy a specific role in the economics of mobile fraud networks: they are recruited, often unknowingly, to receive and forward stolen funds, providing a layer of separation between the originating fraudster and the cash extraction point (ACFE, 2022). The economic incentive structure that produces money-mule recruitment is well-documented in the financial crime literature (Levi, 2015): mules are typically offered small commissions in exchange for account access, creating a principal-agent problem in which the mule bears legal and reputational risk while the organising fraudster extracts the primary surplus. The K-Means clustering result therefore does not merely identify a statistical risk segment; it identifies an economically coherent participant type in the mobile money fraud supply chain.

6.2 Comparison with Prior Work

The logistic regression AUC is consistent with Hajek et al. (2023), confirming strong discriminability when labeled data are available. Cluster 4's fraud rate of 10.81% is consistent with Bolton and Hand's (2001) finding that behavioral profiling reveals risk concentrations

invisible at the individual transaction level. DBSCAN's silhouette advantage (0.700 vs. 0.457) confirms that fraud accounts form density-connected rather than spherical clusters in behavioral feature space. The three-way Jaccard analysis extends Emmott et al. (2015) by providing the first such quantification in the mobile payment context.

6.3 Limitations and Threats to Validity

Following standard validity typology (Wohlin et al., 2012), threats are categorised across four dimensions:

Internal validity. The contamination hyperparameter for all three unsupervised methods was pre-specified at 0.003 to match the observed post-filter fraud rate of 0.2965%, and was fixed prior to model fitting rather than selected post-hoc to maximise fraud recall. Nonetheless, the choice implicitly assumes that anomaly rate and fraud rate coincide; the Jaccard analysis (scores < 0.02) shows this assumption does not hold strictly, which may have suppressed recall estimates. The LOF and OCSVM stratified subsample (fraud rate $\approx 3.0\%$) remains ten-fold higher than the true rate (0.30%), imposing a residual inflation of the local neighbourhood density around fraud cases. Results for those two methods should be interpreted conservatively.

Construct validity. The primary constructs—statistical outlier, behavioral anomaly, and fraud—are operationalised as IQR flags, unsupervised anomaly scores, and the isFraud binary label, respectively. The isFraud label in PaySim is itself a simulation artefact, not an independently adjudicated fraud finding; it may not faithfully represent the construct 'fraud' as encountered in production systems. Jaccard similarity quantifies set overlap but is insensitive to partial score alignment, which may understate the concordance between anomaly scores and fraud probability.

External validity. All distributional claims are explicitly scoped to the PaySim distribution. The balance-zeroing pattern that drives the large balance_error_orig effect (|r| = 0.893) is a consequence of the PaySim simulation mechanism and may have no counterpart in real transaction logs. The PaySim median fraudulent amount (≈USD 441,423) substantially exceeds empirically documented real-world medians, and the 99.85% single-transaction account rate is an acknowledged simulator limitation that artificially constrains the account-level clustering analysis to 1,776 accounts.

Reliability. All random processes use fixed seeds (random_state=42); results are numerically stable across independent re-runs. Code and data access instructions are provided in the Reproducibility Statement to enable independent replication.

6.4 Future Directions

Three lines of investigation follow directly and logically from these findings:

1. Real-data validation. The three-layer overlap framework should be applied to real (non-synthetic) mobile money transaction data as a matter of priority. The synthetic nature of PaySim means that the core finding — near-zero Jaccard overlap between anomaly scores and fraud labels — may reflect simulation artefacts as much as genuine structural separation. Candidate datasets include Jesus and al. (2022)'s NeurIPS 2022 Bank Account Fraud dataset (publicly available) and proprietary operator data accessible through privacy-preserving federated learning frameworks. Replication on real data would allow the Jaccard values to be interpreted as genuine properties of mobile money fraud ecology rather than simulation outputs, and would provide the external validity currently absent from this study.

2. Hybrid detection architecture. Cluster 4's behavioral features (high sender opening balance, dominant CASH OUT usage) should be incorporated as additional inputs to a

supervised gradient-boosted classifier to test whether account-level behavioral profiling provides incremental predictive value over transaction-level features alone.

3. Scalable streaming deployment. Balance-error features should be evaluated as real-time triggers in a streaming anomaly detection pipeline. Approximate scalable methods (HBOS, COPOD, Half-Space Trees) should be benchmarked against IF on the full 2.77M transaction dataset to assess whether the full-population fraud rate (0.30%) changes method rankings relative to the subsampled evaluation.

4. Economic welfare implications of detection architecture choice. The trade-off between precision and recall in fraud detection is not merely a technical optimisation problem but an economic welfare question. High-recall, low-precision systems (such as OCSVM flagging 6.19% of transactions) impose friction costs on legitimate users — disproportionately affecting low-income mobile money users for whom transaction delays or account freezes carry higher welfare costs relative to income (Jack & Suri, 2011). High-precision, low-recall systems (such as IF at precision 0.392, recall 0.040) allow fraud to pass undetected, imposing direct losses that may drive users out of formal financial systems (Suri & Jack, 2016). Future work should formalise this trade-off as a social welfare function incorporating both fraud loss and friction cost, parameterised by the income distribution of the user base, to derive detection thresholds that are economically optimal rather than purely statistically optimal.

7. Conclusion

This study investigated the structural relationship among three operationally distinct abnormality definitions in mobile financial transaction data using the PaySim dataset (6.3M transactions, 8,213 fraud cases). The principal finding is that IQR-based statistical outliers, unsupervised behavioral anomaly detectors, and ground-truth fraud labels capture substantially

different phenomena, with Jaccard similarity below 0.02 for all unsupervised-vs-fraud pairs. No single anomaly detector constitutes a fraud detector: anomaly detection and fraud detection are distinct tasks requiring distinct approaches.

Balance-error features yielded the largest effect sizes ($|r| = 0.893$), substantially exceeding transaction amount ($|r| = 0.367$). Logistic regression achieved ROC-AUC = 0.9791. Among unsupervised methods, OCSVM achieved the best combined performance (F1 = 0.157, AUC = 0.850), while IF achieved the highest precision (0.392). Account-level K-Means clustering ($k = 7$) identified a high-risk behavioral profile (Cluster 4, n = 74, fraud rate 10.81%, ≈ 30× baseline), characterised by large sender balances and dominant CASH OUT behavior. DBSCAN achieved superior geometric quality (silhouette = 0.700) but weaker fraud stratification.

The primary contribution is the first systematic three-way Jaccard overlap quantification on PaySim combined with a validated account-level behavioral clustering framework. Effective fraud detection systems should combine engineered balance-error features for transaction-level scoring with account-level behavioral profiling for risk stratification, rather than relying on any single unsupervised anomaly score.

Reproducibility Statement

Software versions: Python 3.11 · pandas 2.1 · numpy 1.26 · scikit-learn 1.3 · matplotlib 3.8 · seaborn 0.13 · scipy 1.11. All figures were generated directly by running the notebook without manual editing. Results were numerically verified across three independent runs.

Statements and Declarations

Competing Interests: The author declares no competing interests.

Data Availability: The PaySim dataset is publicly available at https://www.kaggle.com/datasets/ealaxi/paysim1.

Funding: No funding was received for this study.

Author Contributions: Hamza Zidoum and Esra Salmeen Awlad Wadair: conceptualisation, methodology, software, formal analysis, writing.

References

GSMA (2025). State of the industry report on mobile money. GSMA Mobile Money Programme.

Association of Certified Fraud Examiners (2024). Report to the nations: 2024 global study on occupational fraud and abuse. ACFE, Austin, TX.

Jesus, S., Pombal, J., Alves, D., Cruz, A., Saleiro, P., Ribeiro, R. P., Gama, J., & Bizarro, P. (2022). Turning the tables: Biased, imbalanced, dynamic tabular datasets for ML evaluation. *Advances in Neural Information Processing Systems*, 35 (NeurIPS 2022).

Hajek, P., Abedin, M. Z., & Sivarajah, U. (2023). Fraud detection in mobile payment systems using an XGBoost-based framework. Information Systems Frontiers, 25, 1985–2003.

Rtayli, N., & Enneya, N. (2020). Enhanced credit card fraud detection based on SVM-recursive feature elimination and hyper-parameters optimization. Journal of Information Security and Applications, 55, 102640.

Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2015). Calibrating probability with undersampling for unbalanced classification. In Proceedings of IEEE SSCI (pp. 159–166).

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM Computing Surveys, 41(3), 1–58.

Demirgüç-Kunt, A., Klapper, L., Singer, D., & Ansar, S. (2022). *The Global Findex Database 2021: Financial inclusion, digital payments, and resilience in the age of COVID-19*. World Bank.

Jack, W., & Suri, T. (2011). Mobile money: The economics of M-Pesa. *NBER Working Paper No. 16721*. National Bureau of Economic Research.

Levi, M. (2015). Money laundering risks and e-gaming: A European overview and assessment. *European Journal on Criminal Policy and Research, 21*(2), 317–336.

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. In Proceedings of IEEE ICDM (pp. 413–422).

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. In Proceedings of ACM SIGMOD (pp. 93–104).

Scholkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. Neural Computation, 13(7), 1443–1471.

Emmott, A., Das, S., Dietterich, T., Fern, A., & Wong, W.-K. (2015). A meta-analysis of the anomaly detection problem. arXiv:1503.01158.

Carcillo, F., et al. (2018). SCARFF: A scalable framework for streaming credit card fraud detection with Spark. Information Fusion, 41, 182–194.

Waspada, I., Bahtiar, D., Wirawan, D., & Eosina, B. (2022). Performance analysis of Isolation Forest in fraud detection. In Proceedings of SNATIA 2022.

Wohlin, C., Runeson, P., Host, M., Ohlsson, M. C., Regnell, B., & Wesslen, A. (2012). Experimentation in software engineering. Springer.

Bolton, R. J., & Hand, D. J. (2001). Unsupervised profiling methods for fraud detection. In Proceedings of credit scoring and credit control VII.

Baesens, B., Van Vlasselaer, V., & Verbeke, W. (2015). Fraud analytics using descriptive, predictive, and social network techniques. Wiley.

Makki, S., et al. (2021). Follow the trail: Machine learning for fraud detection in fintech. Algorithms, 14(3), 78.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357.

Lopez-Rojas, E. A., Elmir, A., & Axelsson, S. (2016). PaySim: A financial mobile money simulator for fraud detection. In Proceedings of EMSS (pp. 249–255).

Suri, T., & Jack, W. (2016). The long-run poverty and gender impacts of mobile money. *Science, 354*(6317), 1288–1292.