Journal of Advanced College of Engineering and Management

Interpretable Model For Anomaly Detection In P2P Financial Transactions

[1*]Mausam Adhikari, [2]Ranju Kumari Shiwakoti, [3]Aditi Khanal
[1,2]Department of Electronics and Computer Engineering, Pulchowk Campus, IOE, TU
[3]Softwarica College of IT and E-Commerce, Coventry Univeristy
Email: [2]ranju.shiwakoti@ioe.edu.np, [3]240260@softwarica.edu.np
Corresponding email: [1]080mscsk007.mausam@pcampus.edu.np*

DOI: 10.3126/jacem.v12i01.93911

Abstract
The rapid growth of peer-to-peer (P2P) digital transactions has increased the risk of transaction errors and anomalous activities, which presents a requirement for a robust and understandable system for detecting anomalies. Although machine learning models have shown promising results in anomaly detection, their integration in real world setting is limited due to lack of transparency. This study focuses on developing an anomaly detection framework for P2P financial transactions utilizing a set of contextual, behavioral, network, and temporal features using financial fields. Additionally, several preprocessing techniques are applied to deal with class imbalances. This study utilizes Machine Learning models such as Logistic Regression, Decision Trees, Random Forest, and XGBoost. XGBoost achieves the best overall performance with a ROC-AUC of 0.968 and a PR-AUC of 0.076, representing a 69-fold improvement over the random baseline. A recall of 0.943 indicates that 94.3% of anomalous transactions are correctly identified. Furthermore, to enhance model interpretability, Explainable AI (XAI) techniques such as SHAP, LIME, and CIU are applied to the best performing model (XGBoost). Spearman rank correlation analysis confirms strong inter-method agreement between SHAP and CIU ($\rho = 0.745$, $p < 0.001$), providing quantitative evidence of explanation consistency.

Keywords—*P2P Digital Transactions, Anomaly Detection, Machine Learning, Explainable AI, SHAP, LIME, CIU, XGBoost*

1. INTRODUCTION

The rapid digitization of the financial sector has led to a tremendous shift towards monetary transactions. In addition, the widespread adoption of mobile wallets and payment platforms has enabled peer-to-peer(P2P) financial transactions that improve accessibility. However, reliability on these systems has increased the risk of transactional anomalies. This poses a serious challenge for major financial institutions and individuals. [1] [2]. Identifying anomalous transactions can be challenging as peer-to-peer(P2P) financial transactions deal with large amounts of dynamic data. Furthermore, traditional rule-based and statistical methods are not appropriate for detecting anomalous activity in such dynamic environments as they lack the necessary scalability and adaptability.

This study presents the use of machine learning-based approaches for effective and accurate anomaly detection identifying complex, non-linear patterns within large volumes of financial data. [3] [4]

Models such as Logistic Regression, Decision Trees, Random Forest, and gradient boosting methods are widely utilized in accurately detecting anomalous activities. But these models often operate as black boxes, limiting the user's ability to understand the system. Hence, their adoption in regulated financial institutions is restricted as explainability and transparency are essential in these environments. As a result, explainable artificial intelligence (XAI) has been adopted as a critical research area as it offers an understanding of how model decisions are derived. [5] and LIME [6] are widely used to interpret model predictions. Moreover, recent approaches such as Contextual Importance and Utility (CIU) [7] offer user centered explanations grounded in utility theory [8] [7].

To solve the challenges, this study presents an interpretable framework for P2P financial transactions combining machine learning models with multiple XAI techniques. The aim of this study is to develop a transparent machine learning framework that increases the trustworthiness of stakeholders and is applicable in real world settings.

2. LITERATURE REVIEW

Anomaly detection in financial systems has been extensively studied due to its critical role in anomaly prevention, risk management, and regulatory compliance. Early research primarily relied on rule-based systems and statistical methods, such as thresholding and probabilistic modeling, to identify irregular transaction patterns. While these approaches offered transparency and ease of implementation, they struggled to adapt to evolving anomaly strategies and high-dimensional transaction data.

With advances in machine learning, supervised and unsupervised learning techniques have become prominent in financial anomaly detection. Models such as logistic regression, decision trees, support vector machines (SVM), and random forests have been widely applied to credit card anomaly detection, money laundering detection, and transaction monitoring tasks. Although these models improve detection accuracy, their performance is often affected by severe class imbalance and concept drift in financial datasets.

More recent studies have explored deep learning techniques, including autoencoders, recurrent neural networks (RNN), and long short-term memory (LSTM) models, to capture complex temporal and behavioral patterns in transaction data. These methods have demonstrated strong predictive performance, particularly in sequential transaction analysis. However, their black-box nature and high computational cost limit their applicability in regulated financial environments, where transparency and explainability are essential.

Tree-based ensemble models, especially gradient boosting methods such as XGBoost, have gained significant attention due to their ability to model non-linear relationships while maintaining strong generalization performance. Several studies report that XGBoost outperforms traditional classifiers in anomaly detection tasks under extreme class imbalance. Nevertheless, despite their effectiveness, ensemble models are still considered opaque, necessitating the use of explainability techniques to support trust and regulatory acceptance.

To address interpretability concerns, explainable artificial intelligence (XAI) methods have been increasingly integrated into financial anomaly detection systems. Post-hoc explanation techniques such as LIME [6] and SHAP [9] are among the most widely used tools for explaining complex model predictions. These methods aim to provide feature-level attributions that help users understand why a transaction was flagged as anomalous. However, prior research highlights limitations related to explanation stability, baseline dependency, and limited alignment with business logic, particularly in high-stakes regulatory contexts.

In parallel, research has revisited inherently interpretable models such as decision trees and rule-based systems. Studies in the financial domain emphasize the importance of incorporating domain-specific knowledge and business rules into anomaly detection models to enhance interpretability and reduce false positives. Hybrid approaches that combine machine learning with expert-defined rules have shown promise, yet they often sacrifice predictive performance or scalability.

Despite these advancements, relatively few studies focus specifically on anomaly detection in peer-to-peer (P2P) financial transactions. P2P systems differ from traditional banking environments in terms of transaction frequency, decentralization, and user behavior, introducing unique detection challenges. Moreover, limited work addresses how anomaly detection systems can provide explanations that are not only technically accurate but also actionable for business users and regulators.

This study builds upon existing research by combining a high-performance gradient boosting model with multiple explainability techniques, with particular emphasis on Contextual Importance and Utility (CIU). Unlike prior work that relies primarily on post-hoc feature attribution, this research evaluates explanation of quality in terms of faithfulness, context sensitivity, and business relevance. By focusing on P2P transactions and regulatory-aligned interpretability, the study contributes a novel perspective to the financial anomaly detection literature.

3. METHODOLOGY

This study adopts a structured, multi-phase methodology to detect and interpret anomalous behavior in peer-to-peer (P2P) financial transactions. The proposed framework integrates domain-driven feature engineering, supervised anomaly detection models, and explainable artificial intelligence (XAI) techniques to balance predictive performance with

transparency and regulatory relevance. All experiments were conducted in a Linux-based environment.

A. Problem Formulation and Hypothesis Development

Anomaly detection is formulated as a binary classification problem, where each transaction is labeled as either normal or anomalous. Given a transaction feature vector $x \in R^n$, the objective is to learn a function:

$$f(x) \rightarrow y, \quad y \in \{0, 1\}$$

where $y = 1$ indicates an anomalous transaction.

The study is guided by the following hypotheses:

- H1: Behavioral and temporal features significantly enhance anomaly detection performance.
- H2: Interpretable models improve stakeholder trust and regulatory usability compared to black-box models.

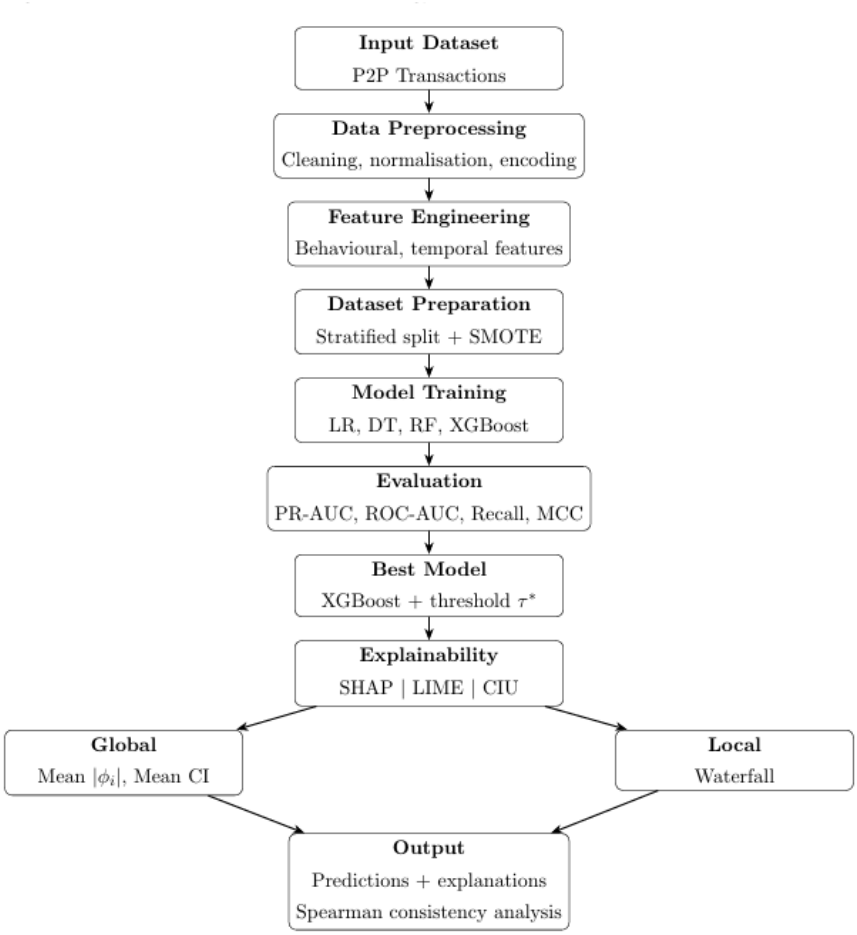

Figure 1: Methodology for interpretable anomaly detection in P2P financial transactions

B. Data Collection and Processing

The dataset [10] consists of structured P2P transaction records obtained from publicly available sources. The overview of the dataset is given below:

Table 1: Dataset Overview

Attribute	Value
Total Transactions	4,547,564
Legitimate (class 0)	4,542,411 (99.89%)
Anomalous (class 1)	5,153 (0.11%)
Class Imbalance Ratio	≈900 : 1
Raw Features	11
Engineered Features	8
Features after Encoding	91
Training Samples (70%)	3,183,294
Test Samples (30%)	1,364,270

The sample data is presented below:

Table 2: Transaction Data Sample

Timestamp	From Bank	Account	To Bank	Account.1	Amount Rec.	Rec. Curr.	Amount Paid	Pay. Curr.	Pay. Format	Is_Laundering
2022-09-01 00:20:00	10	8000EBD30	10	8000EBD30	3697.34	US Dollar	3697.34	US Dollar	Reinvestment	0
2022-09-01 00:20:00	3208	8000F4580	1	8000F5340	0.01	US Dollar	0.01	US Dollar	Cheque	0
2022-09-01 00:00:00	3209	8000F4670	3209	8000F4670	14675.57	US Dollar	14675.57	US Dollar	Reinvestment	0
2022-09-01 00:02:00	12	8000F5030	12	8000F5030	2806.97	US Dollar	2806.97	US Dollar	Reinvestment	0
2022-09-01 00:06:00	10	8000F5200	10	8000F5200	36682.97	US Dollar	36682.97	US Dollar	Reinvestment	0

Preprocessing ensures data quality and modeling readiness through:

- Removed Banks with no recorded Anomaly Activity to reduce noise without discarding positive class signal
- Removal of duplicate and inconsistent records
- Handling missing values via imputation or exclusion
- Feature normalization and log transformation for skewed monetary attributes
- Encoding categorical variables using one-hot or target encoding
- Preliminary outlier screening to distinguish noise from meaningful anomalies

The dataset was split into training (70%) and test (30%) subsets using stratified sampling(random_state=42) to preserve the original class ratio in both partitions. Stratification is essential for imbalanced datasets as random splits may produce training or test folds with disproportionate anomaly rates, leading to bias evaluation. The resulting split contains 3,183,294 training samples and 1,364,270 test samples.

Table 3: Bank and Entity Information Sample

Bank Name	Bank ID	Account Number	Entity ID	Entity Name
Portugal Bank #4507	331579	80B779D80	80062E240	Sole Proprietorship #50438
Canada Bank #27	210	809D86900	800C998A0	Corporation #33520
UK Bank #33	21884	80812BE00	800C47F50	Partnership #35397
Germany Bank #4815	32742	81047F300	80096F0B0	Corporation #48813
National Bank of Harrisburg	127390	80BD8CF00	800FB8760	Corporation #889

In this study, the term anomaly and money laundering are used interchangeably to refer to transactions labeled Is_Laundering=1 in the dataset.

C. Feature Engineering

Effective feature engineering is critical to capturing meaningful patterns. This involves creating transaction-level and aggregated features to capture user behavior, transaction patterns, and potential anomaly signals. The following feature categories were designed:

Let each transaction be denoted as:

$$T_i = (s_i, r_i, t_i)$$

where s_i is the sender, r_i is the receiver, and t_i is the timestamp. Transactions are chronologically ordered such that $t_1 \leq t_2 \leq \cdots \leq t_n$. Let $\mathbf{1}(\cdot)$ denote the indicator function.

a. Transaction Type ($Type_i$) Binary indicator of whether the transaction is a self-transfer.

$$Type_i = \{1 \quad if\ s_i = r_i\ 0 \quad otherwise$$

b. Hour ($Hour_i$) Hour of the day extracted from the timestamp.

$$Hour_i = hour(t_i), \quad Hour_i \in \{0, \ldots, 23\}$$

c. Weekday ($Weekday_i$) Day of the week extracted from the timestamp.

$$Weekday_i = weekday(t_i), \quad Weekday_i \in \{0, \ldots, 6\}$$

d. Sender Overall Transaction Count ($C_i^{(S)}$) Cumulative number of previous transactions by the sender.

$$C_i^{(S)} = \sum_{j<i} 1\left(s_j = s_i\right)$$

e. Receiver Overall Transaction Count ($C_i^{(R)}$) Cumulative number of previous transactions received.

$$C_i^{(R)} = \sum_{j<i} 1\left(r_j = r_i\right)$$

f. Sender Daily Transaction Count ($D_i^{(S)}$) Number of previous transactions by the sender on the same day.

$$D_i^{(S)} = \sum_{j<i} 1\left(s_j = s_i \wedge d_j = d_i\right)$$

g. Receiver Daily Transaction Count ($D_i^{(R)}$) Number of previous transactions received on the same day.

$$D_i^{(R)} = \sum_{j<i} 1\left(r_j = r_i \wedge d_j = d_i\right)$$

These engineered features aim to improve both detection accuracy and explanation clarity.

h. Target Variable

Is_Laundering: Binary indicator of anomaly.

D. Features Description

Anomalous behavior is often characterized by subtle, contextual, and temporal patterns rather than explicit rule violations. Prior studies emphasize that domain-driven features significantly enhance both detection performance and interpretability in financial systems [11]

In this study, features are grouped into transaction-level, behavioral/network, temporal, and entity-level categories to capture diverse aspects of user activity in peer-to-peer (P2P) financial systems.

E. Class Imbalance Handling

Three complementary strategies are applied to address the 900:1 class imbalance:

a. SMOTE(Synthetic Minority Over Sampling Technique): SMOTE generates synthetic minority class samples by interpolating existing anomalous transactions in

feature space, using the k=10 nearest neighbors. Critically, SMOTE is applied exclusively to the training set after the train-test split, and the test set retains the original class distribution. Applying SMOTE before splitting would constitute data leakage, as synthetic test instances derived from training data would inflate performance estimates. After resampling, the training set contains 6346288 samples with a balanced 1:1 class ratio.

b. **Class weighting:** Logistic Regression and Random Forest are trained with class_weight = {0:1,1:3}, applying three times the penalty to misclassified anomalies. This further biases the decision boundary towards the minority class.

c. **scale_pos_weight(XGBoost):**
XGBoost receives scale_pos_weight=n_negative/n_positive = 900, which weights the gradient updates for positive class samples proportionally to the imbalance ratio. This is XGBoost's built-in equivalent of class weighing and is recommended over external oversampling alone for gradient boosted trees.

F. Anomaly Detection Modeling

Considering the requirements of interpretability, computational efficiency, and suitability for business and regulatory contexts, this study emphasizes lightweight and explainable machine learning models for anomaly detection in P2P financial transactions. Prior research highlights that highly complex models, although accurate, often lack transparency, limiting their adoption in regulated financial environments [11] [12]. Therefore, the selected models aim to balance predictive performance with explainability and operational feasibility. This study focused on the following models: Random Forest, Logistic Regression, Decision Tree, XGBoost.

Although deep learning approaches such as Autoencoders and Long Short-Term Memory (LSTM) networks have shown promise in capturing complex temporal and sequential patterns in financial data, their black-box nature and higher computational requirements pose challenges for explainability and regulatory compliance [13]. Given the study's focus on transparency, auditability, and practical deployment in business environments, such models are not prioritized.

Overall, model selection in this study is guided by a balanced consideration of detection performance, interpretability, and computational efficiency. This ensures that anomaly detection outputs are not only accurate but also explainable, actionable, and aligned with financial domain and regulatory expectations.

G. Explainability and Interpretation

Global feature importance is computed for each method by averaging individual instance attributions across 200 test transactions. For SHAP, this is the mean absolute SHAP value, For CIU, the mean Contextual Importance (CI), For LIME, the mean

absolute local weight. All three methods consistently rank Payment Format (ACH), sender transaction daily count and Amount Received as top three predictors of anomalous behavior. This convergence across methodologically distinct approaches. SHAP's exact game theoretic decomposition, CIU's range based contextual measure, and LIME's linear local approximation, provides strong convergent validity for the identified feature.

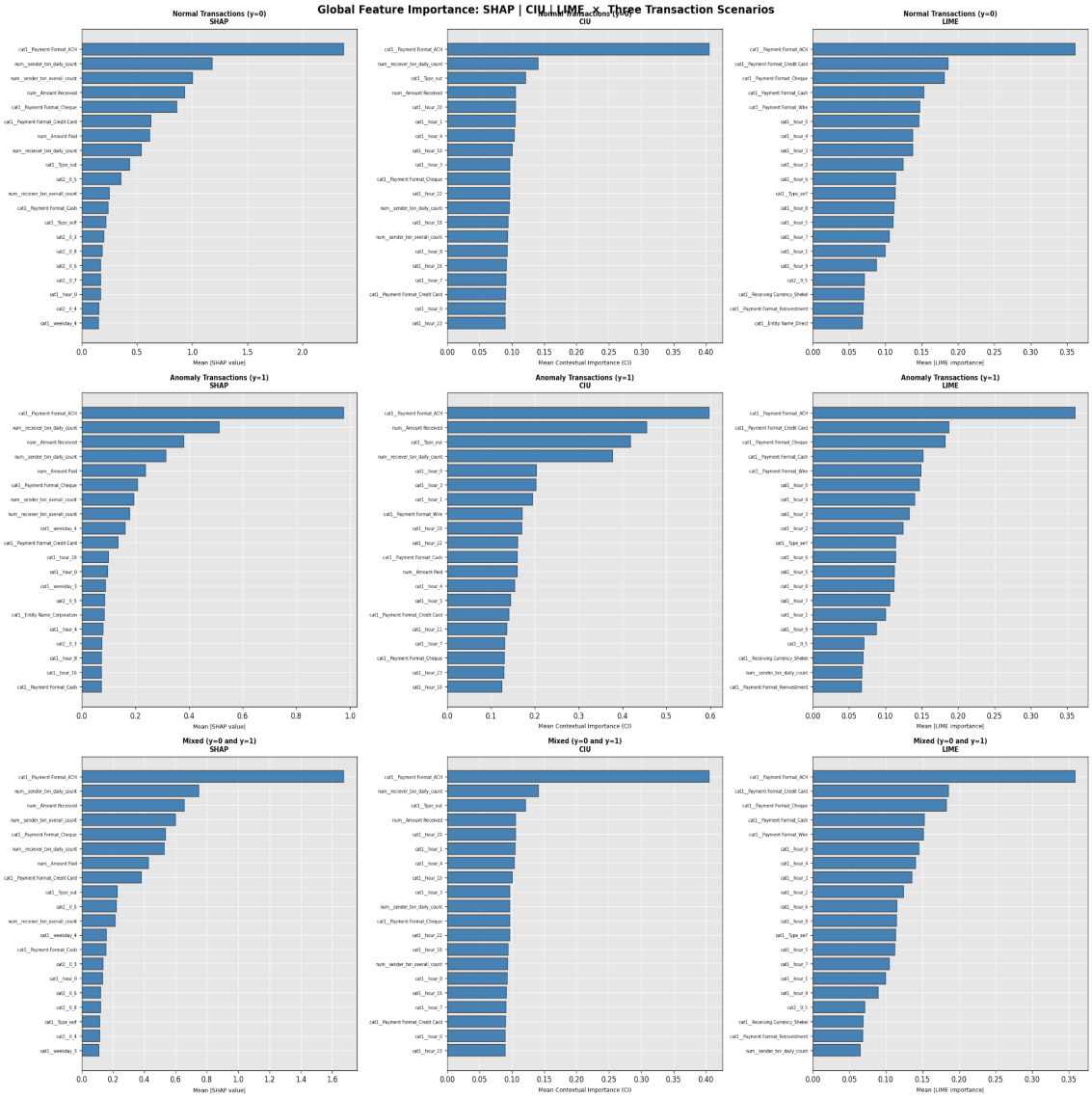

Figure 2: Global feature importance comparison

To illustrate the instance level behavior, two representative transactions are selected, the test sample with no predicted anomaly(0,Is_Laundering=0) and sample with predicted anomaly(1,Is_Laundering=1). For the anomalous transaction, SHAP assigns the largest positive contribution to Payment Format ACH(phi=+1.24), a finding corroborated by LIME(+0.361) and CIU(CI=0.371, the highest contextual feature

importance of all features) For the normal transaction, Wire Transfer (SHAP phi=-2.80) and high transaction amounts strongly suppress the anomaly score, consistent with legitimate high value interbank activity. Figure presents the three-way local comparison across both instances

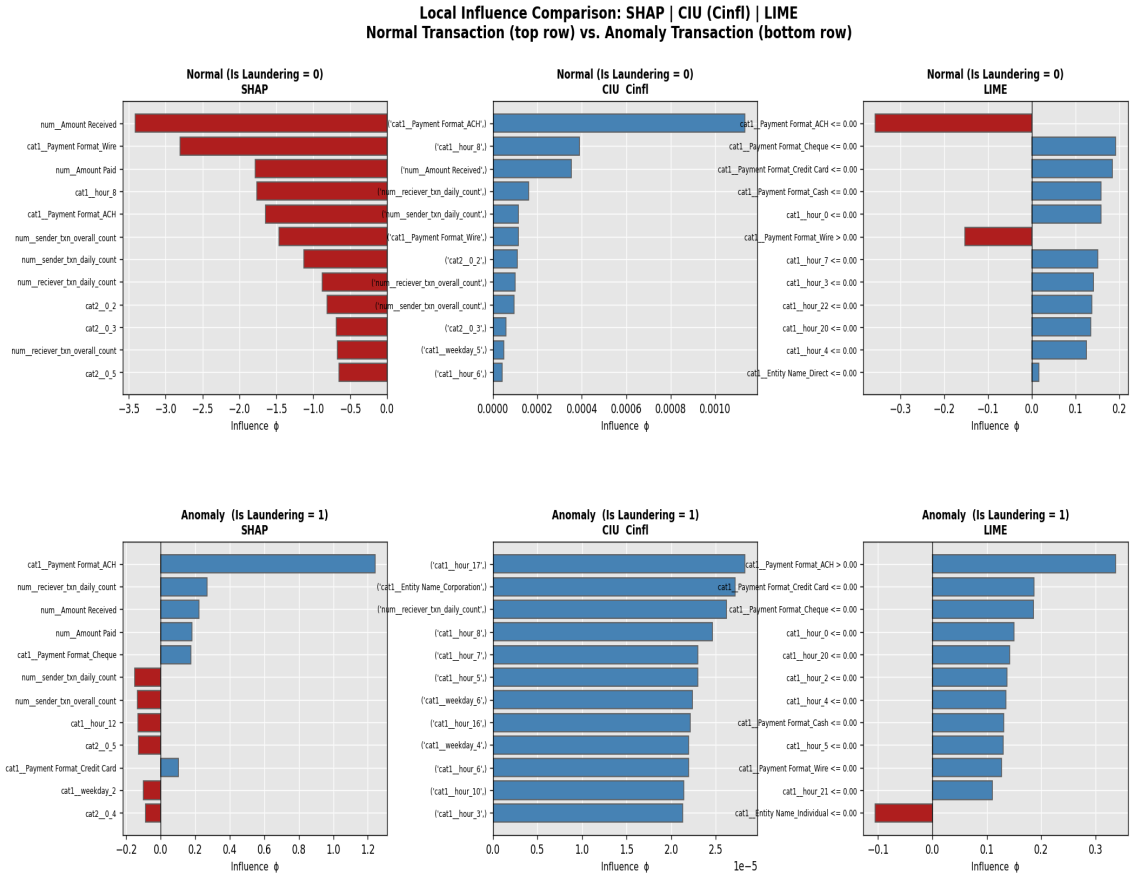

Figure 3: Three-way local influence comparison

To quantify the consistency of the three explainability methods, Spearman rank correlation(rho) of feature importance ranking is computed across 200 test samples for each transaction scenario (Table below). SHAP and CIU exhibit the strongest agreement (rho=0.745, p<0.001 for normal only and mixed scenarios, rho=0.668, p<0.001 for anomaly only), confirming that both methods consistently identify the same features as most influential. SHAP and LIME show moderate agreement (rho-0.371, p<0.001), while CIU and LIME show weaker but statistically significant correlation rho=0.240-0.300, p<0.05). The lower LIME correlation is expected. LIME fits a linear surrogate locally, and weights features differently from SHPA's exact Shapley values or CIU's range based on contextual importance. Crucially, all three methods agree that Payment Format and Transaction Velocity are the primary drivers of Anomaly predictions, lending convergent validity of the framework.

Table 4: Spearman Rank Correlation of Feature Importance Rankings Between Explainability Methods (n=200 samples per scenario)

Method Pair	**Normal (y = 0)**	**Anomalous (y = 1)**	**Mixed**
ρ (SHAP–CIU)	0.745	0.668	0.745
p-value	< 0.001	< 0.001	<0.001
ρ (SHAP–LIME)	0.371	0.176	0.376
p-value	< 0.001	0.096	<0.001
ρ (CIU–LIME)	0.240	0.300	0.228
p-value	0.022	0.004	0.030

SHAP and CIU exhibit the strongest agreement ($\rho = 0.745$, $p < 0.001$ for normal and mixed scenarios; $\rho = 0.668$, $p < 0.001$ for anomalous-only), confirming that both methods consistently identify the same features as most influential. SHAP and LIME show moderate agreement ($\rho = 0.371$, $p < 0.001$), while CIU and LIME show weaker but statistically significant correlation ($\rho = 0.240$–0.300, $p < 0.05$)

4. RESULT AND DISCUSSION

Table 5: Model Performance Comparison

Metric	**Random Forest**	**Logistic Regression**	**Decision Tree**	**XGBoost**
Accuracy	0.861	0.811	0.946	0.868
Precision	0.008	0.006	0.017	0.008
Recall (Sensitivity)	0.946	0.969	0.790	0.943
F1 Score	0.015	0.012	0.032	0.016
ROC AUC	0.965	0.942	0.952	**0.968**
PR AUC	0.047	0.017	0.033	**0.076**
MCC	0.079	0.067	0.110	0.081
Balanced Accuracy	0.904	0.890	0.868	0.905

Due to severe class imbalance in the dataset (0.11% anomaly rate, approximately 900:1 ratio), aggregate accuracy is not a reliable performance indicator. A trivial classifier that predicts "legitimate" for every transaction achieves 99.89% accuracy while detecting Zero anomalies. We therefore designated PR-AUC as primary evaluation Metric, supplemented by ROC-AUC, Recall, Mathews Correlation Coefficient (MCC), and Balanced Accuracy.

XGBoost achieved the highest Roc-AUC of 0.968 and PR-AUC of 0.076. While the absolute PR-AUC value appears low, it represents a 69 times improvement over a random baseline (class rate = 0.0011), which is consistent with results reported in published AML detection literature on similarly imbalanced real-world datasets. A recall of 0.943 indicates that the model successfully identifies 94.3% of all anomalous transactions. The low precision (0.008) reflects the inherent cost of operating under extreme imbalance.

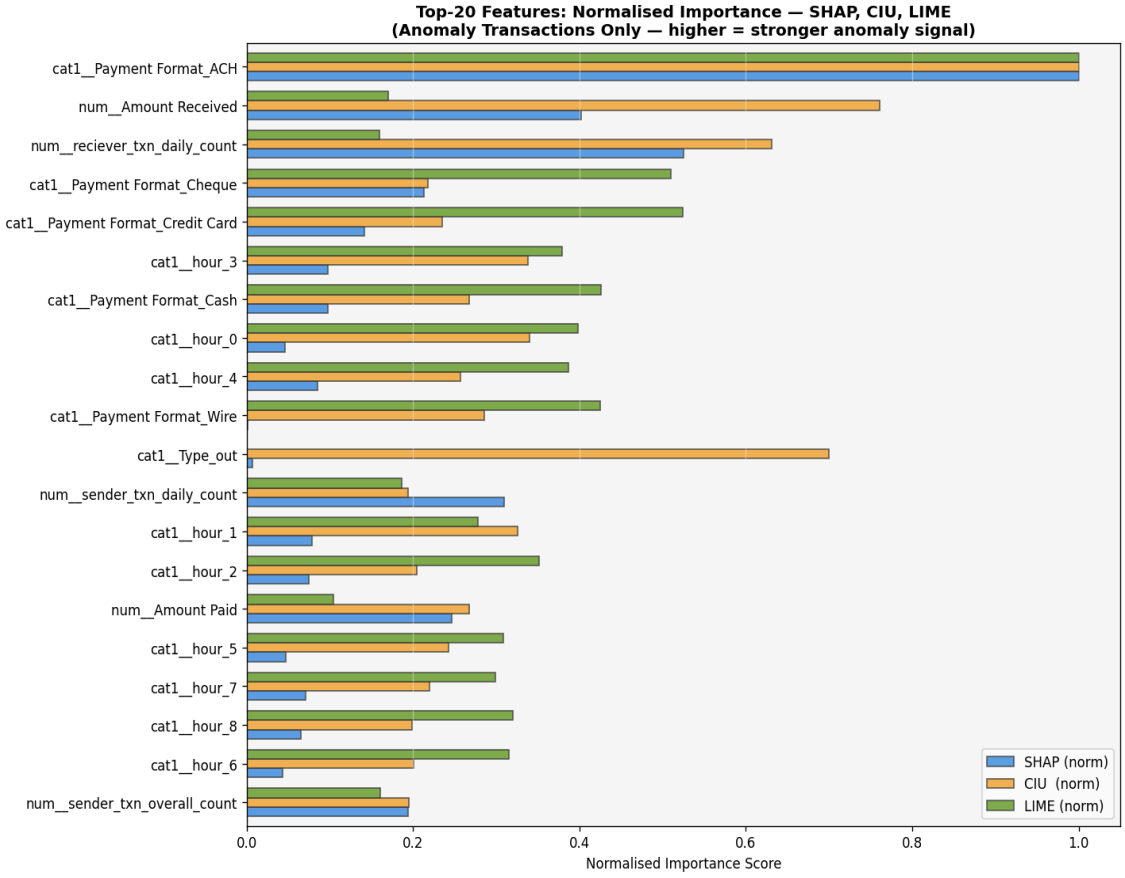

Figure 5: Feature Importance Comparison for SHAP,LIME and CIU with min-max normalization

The Figure shows the dominant feature for anomay is Payment Format ACH, ranked first by three methods with score 1. Beyong ACH, these methods shows partial agreement between Amount Received rank second by scores SHAP(0.42) ,LIME(0.17) and CIU (0.76).LIME's lower importance suggests local linear approximation underestimates global contribution in nonlinear regions. Receiver Transaction Daily Count scores by

SHAP(0.53) , LIME(0.15) and CIU(0.63),shows velocity based features show non linear interaction effects that LIME doesnot fully capture.

The spearman rank correlation analysis quantifies the agreement , SHAP and CIU achieve $\rho = 0.745(p<0.001)$ on anomalous transctions, which confirms strong inter method agreement on most of the important features.

5. CONCLUSION AND RECOMMENDATIONS

The study concludes the interpretable machine learning model for detecting anomalies in peer-to-peer financial systems with a focus on three explanation methods SHAP, LIME, and CIU. With regulatory constraints and operational sensitivity of financial anomaly detection, interpretability is essential for building trust, ensuring accountability, and supporting human decisions.

The results show each method provides complimentary insights on model explainability, SHAP provides consistent, global and local features attributions whereas LIME provides intuitive instance level explanations, CIU captures context dependent feature relevance.

Instead of a single technique for explainability, the combined interpretable framework is suitable for p2p financial anomaly detection.

The study demonstrates that integrating multiple explanation approaches enhances transparency and supports the responsible deployment of machine learning models in financial anomaly detection systems.

The study suggests that effective interpretability in P2P financial anomaly detection is best achieved by integrating Hybrid SHAP, LIME, and CIU as each method provides a complementary explanatory perspective.

Future work should prioritize explanation stability assessment, scalability optimization for high volume transaction environments, and stronger integration of financial domain knowledge.

Also, comparative evaluations across diverse datasets and model architectures would help establish standardized best practices for interpretable anomaly detection in financial systems.

REFERENCES

[1] A. a. o. Ozbayoglu, "Risk-aware fraud detection for mobile payment systems," IEEE Access, vol. 8, pp. 148017-148030, 2020.

[2] M. K. a. o. Lim, "Financial fraud detection using machine learning: A systematic literature review," Applied Soft Computing, vol. 103, 2021.

[3] Y. a. C. Z. a. L. W. Liu, "Imbalanced learning for fraud detection: A survey," IEEE Transactions on Knowledge and Data Engineering, 2022.

[4] J. a. o. Zhang, "Gradient boosting-based fraud detection in large-scale transaction systems," Knowledge-Based Systems, vol. 233, 2021.

[5] L. S. Shapley, "A value for n-person games," Contributions to the Theory of Games, vol. 2, pp. 307-317, 1953.

[6] M. T. a. S. S. a. G. C. Ribeiro, ""“Why Should I Trust You?”: Explaining the Predictions of Any Classifier," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135 - 1144.

[7] A. H. a. A. M. M. a. W. P. B. W. Fra¨mling, "Contextual importance and utility for explainable AI," Artificial Intelligence, vol. 282, 2020.

[8] S. a. o. Lundberg, "From local explanations to global understanding with explainable AI," Nature Machine Intelligence, vol. 2, pp. 252- 260, 2020.

[9] S. M. a. L. S.-I. Lundberg, "A unified approach to interpreting model predictions," Advances in Neural Information Processing Systems, vol. 30, 2017.

[10] E. A. a. J. B. a. L. v. N. a. B. E. a. A. A. a. K. Atasu, "Realistic Synthetic Financial Transactions for Anti-Money Laundering Models," 2024.

[11] N. a. G. P. a. M. D. a. P. J. Bussmann, "Explainable AI in credit risk management," Computational Economics, vol. 57, pp. 203-216, 2021.

[12] C. Molnar, "Interpretable Machine Learning," Lulu.com.

[13] A. a. A. R. S. a. O. S. H. a. E. T. A. E. a. A.-D. A. a. N. M. a. E. T. a. E. H. a. S. A. Ali, "Financial Fraud Detection Based on Machine Learning: A Systematic Literature Review," Applied Sciences, vol. 12, p. 9637, 2022.

[14] N. a. D. J. E. a. G. S. a. T. M. a. M. E. T. a. B. K. Baisholan, "FraudX AI: An Interpretable Machine Learning Framework for Credit Card Fraud Detection on Imbalanced Datasets," Computers, vol. 14, 2025.

[15] M. a. M. A. N. a. H. J. Ahmed, "A survey of network anomaly detection techniques," Journal of Network and Computer Applications, vol. 60, pp. 19-31, 2016.