When Does Context Help? An Empirical Study of Contextual Anomaly Detection for Transaction Monitoring

Anonymous Author(s)
Anonymous Institution

***Abstract*—Contextual anomaly detection identifies data points that are normal globally but unusual within their specific context. While peer group analysis has long been used in anti-money laundering (AML) systems, the question of *when* contextual methods provide value over global approaches remains understudied. This paper presents an empirical study—not a new detection algorithm—comparing Isolation Forest (IF) against peer-normalized variants on three financial datasets. Using controlled injection strategies grounded in AML typologies, we demonstrate a clear pattern: IF excels at detecting globally unusual behavior, while peer-normalized methods excel at detecting behavior that is normal globally but unusual for a specific context. We find that the deep variant (PNKDIF) provides marginal benefit only at high contamination rates. Our key contribution is reframing contextual detection as a *diagnostic tool*: disagreement between IF and contextual methods reliably signals the presence of contextual structure in the data. This provides practitioners with actionable guidance on method selection for transaction monitoring systems.**

***Index Terms*—anomaly detection, contextual anomaly, anti-money laundering, isolation forest, peer group analysis**

I. Introduction

Anti-money laundering (AML) systems must detect suspicious transaction patterns across diverse customer populations. A key challenge is that “normal” behavior varies by context: a high-value international transfer may be routine for a multinational corporation but highly unusual for a domestic retail account. This observation motivates *contextual anomaly detection*, where anomalies are defined relative to similar entities rather than the global population.

Peer group analysis has been used in financial crime detection since Bolton and Hand’s seminal work [1]. The core idea is simple: compare each entity to its “peers” (similar entities based on context features) rather than to the entire population. Despite widespread industry adoption, there is limited empirical guidance on *when* contextual methods provide value over simpler global approaches.

In this paper, we address the question: **When does context help in anomaly detection for transaction monitoring?**

A. Contributions

This paper makes *methodological and empirical* contributions, rather than proposing a universally superior anomaly detection model.

1) **Empirical clarification of when context helps.** We provide a systematic empirical study showing that the effectiveness of contextual anomaly detection depends on the *type* of anomaly present. Global anomaly detectors such as Isolation Forest excel when anomalies are globally unusual, while contextual methods are effective only when anomalies are normal globally but deviate relative to peer groups.
2) **Contextual anomaly detection as a diagnostic tool.** We reframe contextual methods not as replacements for global detectors, but as *diagnostic instruments*. We demonstrate that disagreement between global and contextual detectors reliably signals the presence of contextual structure in the data.
3) **Controlled evaluation via domain-grounded injections.** We argue that standard anomaly detection benchmarks are dominated by global anomalies, making them unsuitable for evaluating contextual methods. We therefore use controlled, domain-grounded injection strategies to isolate conditional deviations, enabling meaningful evaluation of contextual detection capabilities.
4) **Minimal reference implementation (PNKIF).** We introduce Peer-Normalized Kernel Isolation Forest as a minimal, training-free reference implementation to study contextual effects. PNKIF is *not* proposed as a universally superior detector, but as a simple and interpretable vehicle for isolating contextual anomaly behavior.
5) **Negative results on deep extensions.** We evaluate a deep variant using random neural projections and show that it provides marginal or inconsistent benefit. This negative result highlights that peer normalization—not model depth—is the primary driver of contextual detection performance.

B. Scope and Non-Claims

This paper deliberately limits its scope and makes several explicit non-claims:

- We do *not* claim that contextual anomaly detection is universally superior to global methods.
- We do *not* claim that PNKIF outperforms deep or learned contextual models such as conditional VAEs in all settings.
- We do *not* claim to model complex conditional distributions where context alters distributional shape or multimodality.

Instead, we focus on the most common and operationally relevant form of contextual dependence, where context primarily induces *location and scale shifts* in behavior (e.g., transaction volume scaling with customer profile). More complex conditional structures may require learned generative models, which introduce additional training complexity and stability concerns beyond the scope of this study.

II. Related Work

A. Anomaly Detection in AML

Machine learning for AML has been extensively surveyed [2]. Common approaches include rule-based systems, supervised classification, and unsupervised anomaly detection. Isolation Forest [3] is widely used due to its efficiency and effectiveness on high-dimensional data.

B. Contextual Anomaly Detection

Contextual anomalies are data points that are unusual only within a specific context [4]. Methods include:

- **ROCOD** [5]: K-NN based peer normalization with robust statistics
- **QCAD** [6]: Quantile regression for conditional distributions
- **ConQuest** [7]: Context discovery for anomaly detection

C. Peer Group Analysis

Bolton and Hand [1] introduced peer group analysis for fraud detection. The approach groups entities by context features and flags deviations from peer behavior. Our reference implementation formalizes this with kernel-weighted peer normalization.

III. Methods

A. Problem Setting

Given dataset $\{(\mathbf{c}_i, \mathbf{x}_i)\}_{i=1}^N$ where $\mathbf{c}_i \in \mathbb{R}^{d_c}$ is the context vector and $\mathbf{x}_i \in \mathbb{R}^{d_x}$ is the behavioral vector, we aim to detect anomalies that are unusual *given their context*.

B. Isolation Forest (IF)

Isolation Forest [3] detects anomalies by measuring how easily a point can be isolated via random recursive partitioning. It operates on the concatenated features $[\mathbf{c}; \mathbf{x}]$ or behavior only $\mathbf{x}$, without explicitly modeling context-behavior relationships.

C. Peer-Normalized Kernel Isolation Forest (PNKIF)

PNKIF is *not* proposed as a novel anomaly detection paradigm. Instead, it serves as a minimal, interpretable reference implementation designed to isolate the effect of peer-based normalization on anomaly scoring. By deliberately avoiding learned representations or deep architectures, observed performance differences can be attributed to *contextual normalization* rather than representational capacity.

The method computes kernel-weighted peer statistics using RBF weights $w_{ij} = \exp(-\|\mathbf{c}_i - \mathbf{c}_j\|^2/2\gamma^2)$, then normalizes each point's behavior by its peer mean and standard deviation: $\tilde{\mathbf{x}}_i = (\mathbf{x}_i - \boldsymbol{\mu}_i)/\boldsymbol{\sigma}_i$. Isolation Forest is then applied to the normalized behaviors.

D. On the Role of Random Projections (PNKDIF)

We evaluated a deep variant (PNKDIF) incorporating frozen random neural projections, inspired by Deep Isolation Forest [8]. PNKDIF projects peer-normalized features through M random MLPs and averages anomaly scores across projections.

Results were mixed: marginal benefit at high contamination ($>5\%$) on geographic swap injection, but inconsistent or degraded performance in other scenarios. This suggests that *peer normalization alone* captures the dominant contextual signal, and additional representational complexity offers limited benefit. We therefore emphasize PNKIF for its simplicity, interpretability, and computational efficiency, while reporting PNKDIF results for completeness.

IV. Experimental Setup

A. Datasets

We use three public financial datasets:

- **SAML-D** [9]: Synthetic AML dataset with 30K accounts. Context: geography, payment type, currency (38 features after one-hot encoding). Behavior: transaction statistics (6 features).
- **PaySim** [10]: Mobile money simulation with 30K transactions. Context: transaction type (5 features). Behavior: amounts and balances (5 features).
- **Credit Card** [11]: Anonymized transactions with 30K samples. Context: time and amount (2 features). Behavior: PCA components V1-V28 (28 features).

B. Why Controlled Injection Is Necessary

Evaluating contextual anomaly detection presents a fundamental challenge: public benchmarks overwhelmingly contain *global* anomalies—samples that are unusual regardless of context.

In such datasets, contextual methods cannot demonstrate their defining capability, because global detectors already succeed. As a result, naive evaluation misleadingly suggests that context provides no benefit.

To isolate the statistical property of interest—conditional deviation given context—we use controlled, domain-grounded injection strategies that satisfy two constraints:

1) **Normal globally:** Injected behaviors are drawn from real samples and lie within the global distribution.
2) **Abnormal conditionally:** The same behaviors violate expectations relative to their assigned context.

Importantly, injection does *not* assume fraud semantics, ground-truth labels, or operational realism. It functions as a **controlled falsification test**, analogous to stress-testing a model under known violations of its assumptions. Without such controlled violations, contextual anomaly detection cannot be meaningfully evaluated.

C. Injection Strategies

We use domain-grounded strategies motivated by known AML typologies:

1) **Geographic Swap (Contextual):** Simulate geographic arbitrage—a known FATF money laundering typology—by assigning behavior from one geographic region to accounts in another region.
2) **Context Mismatch (Contextual):** Simulate account misuse by assigning behavior from a randomly different context group.
3) **Velocity Anomaly (Global):** Scale transaction amounts by 2-5x, simulating structuring behavior.
4) **Temporal Shift (Global):** Add systematic shifts (2-3 standard deviations) to behavior features.

Injection rates: 1%, 3%, 5%, 10%.

D. Evaluation

We report AUROC across 10 random seeds for robustness. Methods compared:

- **IF**: Isolation Forest on behavior only
- **IF_concat**: Isolation Forest on concatenated context + behavior
- **ROCOD**: K-NN peer normalization with robust statistics [5]
- **PNKIF**: Kernel-weighted peer normalization + IF
- **PNKDIF**: PNKIF + random MLP projections

Limitations of AUROC: We acknowledge that AUROC has known limitations under extreme class imbalance typical of AML applications. A full operational evaluation would require precision@k, workload-based metrics, and alert volume analysis. We use AUROC as a standardized comparison metric while recognizing it does not fully capture practical deployment considerations.

V. Results

A. Original Labels: Global Anomalies

On original dataset labels, IF and IF_concat consistently outperform contextual methods (Table I).

TABLE I
AUROC on Original Labels (10 seeds)

Dataset	IF	IF_concat	ROCOD	PNKIF	PNKDIF
SAML-D	**0.937**	0.896	0.419	0.869	0.842
PaySim	0.691	**0.776**	0.375	0.455	0.353
CreditCard	**0.947**	0.946	0.912	0.926	0.918

This is expected: original labels correspond to globally unusual behavior that IF detects effectively. Contextual methods add overhead without benefit when anomalies are globally detectable.

B. Contextual Injection: Peer Methods Win

On contextual anomalies (context mismatch), peer-normalized methods consistently outperform IF (Table II).

At low injection rates (1%), IF still wins because the contextual signal is weak. At higher rates (3-10%), PNKIF consistently outperforms IF, with the gap widening as injection rate increases.

TABLE II
AUROC on Context Mismatch Injection (PaySim, 10 seeds)

Rate	IF	PNKIF	PNKDIF	Winner
1%	**0.650**	0.536	0.469	IF
3%	0.615	**0.633**	0.586	PNKIF
5%	0.591	**0.663**	0.616	PNKIF
10%	0.563	**0.690**	0.677	PNKIF

Observation: PNKIF performance *increases* from 5% to 10% (0.663 → 0.690). This counterintuitive result occurs because more contextual anomalies create a stronger deviation signal from peer norms, making them easier to detect.

C. Geographic Swap: PNKDIF Wins at High Rates

On geographic swap injection, PNKDIF outperforms PNKIF at higher contamination rates (Table III).

TABLE III
AUROC on Geographic Swap Injection (PaySim, 10 seeds)

Rate	IF	PNKIF	PNKDIF	Winner
1%	**0.610**	0.549	0.500	IF
3%	0.528	**0.589**	0.583	PNKIF
5%	0.479	0.598	**0.617**	PNKDIF
10%	0.422	0.599	**0.628**	PNKDIF

The deep projections in PNKDIF provide additional benefit when contamination is high, possibly by creating non-linear decision boundaries that better separate complex anomaly patterns. However, this benefit is limited to specific scenarios and does not generalize.

D. Global Injection: IF Wins

On global-style anomalies (velocity, temporal shift), IF wins 100% of scenarios. These anomalies stand out globally, so peer normalization provides no benefit.

E. Summary Across All Experiments

TABLE IV
Win Rate by Injection Type (All Datasets)

Injection Type	IF	PNKIF	PNKDIF
Context Mismatch	4/12	**8/12**	0/12
Geographic Swap	4/12	4/12	**4/12**
Velocity Anomaly	**12/12**	0/12	0/12
Temporal Shift	**12/12**	0/12	0/12

VI. Discussion

A. Quantifying Method Disagreement

A key finding of this study is that *agreement and disagreement* between global and contextual detectors is itself informative. We quantify this by measuring overlap in top-5% flagged samples (Agreement@5%) and precision at rank 5% (P@5%) as contextual contamination increases (Table V).

TABLE V
DIAGNOSTIC METRICS: IF VS PNKIF AGREEMENT (PAYSIM, 5 SEEDS)

Injection Rate	Agreement @5%	IF P@5%	PNKIF P@5%
0% (original)	29.2%	11.5%	4.0%
1%	28.5%	12.1%	11.9%
3%	26.1%	13.9%	21.5%
5%	25.0%	16.1%	30.2%
10%	24.0%	21.1%	**38.8%**

Two patterns emerge: (1) Agreement decreases monotonically as contextual contamination increases, from 29% to 24%. (2) PNKIF precision improves dramatically (4% → 39%) while IF precision grows modestly (12% → 21%). This confirms that disagreement reliably signals contextual structure.

B. Diagnostic Workflow

This suggests a practical diagnostic: run IF and PNKIF on the same data.

- **High agreement + similar precision:** Anomalies are global; IF is sufficient.
- **Low agreement + PNKIF outperforms:** Contextual anomalies present; use PNKIF.
- **PNKIF precision $\gg$ IF precision:** Strong contextual signal in the data.

This reframing resolves a long-standing ambiguity in AML systems, where contextual methods are often deployed without evidence that context is actually informative.

C. When Does Context Help?

Our results provide clear guidance:

- **Use IF** when anomalies are globally unusual (unusual amounts, frequencies, or feature values) or when contamination is very low (<1%)
- **Use PNKIF** when anomalies are contextually unusual and you need interpretability (peer comparison is explainable)
- **Use PNKDIF** only when contamination is high (>5%) and complex patterns are expected; otherwise prefer PNKIF for simplicity

D. Effect of Injection Rate

We observe that contextual method performance can *improve* with higher injection rates. This occurs because:

- More anomalies create stronger deviation from peer norms
- The "mismatch" signal becomes more pronounced
- AUROC benefits from having more positive samples to rank

Conversely, IF performance degrades at higher rates because global statistics become contaminated.

E. Practical Implications for AML

Contextual anomalies correspond to known money laundering typologies:

- **Geographic arbitrage**: Domestic accounts with international transaction patterns
- **Account takeover**: Behavior inconsistent with account profile
- **Peer group deviation**: Activity unusual for customer segment

Standard IF may miss these if the behavior is common globally. Peer-normalized methods provide complementary detection.

F. Limitations

- **Injection-based evaluation:** Real contextual labels are rare in public datasets. We frame injection as controlled falsification, not proxy for reality.
- **Location-scale assumption:** We model context as inducing mean/variance shifts. Complex conditional distributions (multimodality, shape changes) require learned models.
- **Context feature selection:** Performance depends on choosing appropriate context features; poor context leads to poor peer groups.
- **Computational cost:** PNKIF/PNKDIF require K-NN computation, adding overhead for large datasets.
- **AUROC limitations:** This metric may not reflect operational AML performance under extreme imbalance.

VII. CONCLUSION

We presented an empirical study of contextual vs. global anomaly detection for transaction monitoring. Our key finding is that method effectiveness depends on anomaly type: IF for global anomalies, PNKIF for contextual anomalies at moderate contamination. The deep variant (PNKDIF) provides limited additional benefit.

More importantly, we propose reframing contextual detection as a *diagnostic tool*: disagreement between global and contextual methods signals the presence of contextual structure. Our results suggest that the first question in anomaly detection should not be "which model is best?" but rather "does context matter at all for this dataset?"

Future work includes validation on proprietary AML data with natural contextual labels, formalization of the disagreement diagnostic as a statistical test, and investigation of adaptive method selection based on data characteristics.

REFERENCES

[1] R. J. Bolton and D. J. Hand, "Peer group analysis and consumer fraud detection," *Statistics in Practice*, 2001.

[2] Z. Chen, L. D. Van Khoa, E. N. Teoh *et al.*, "Machine learning techniques for anti-money laundering (aml) solutions in suspicious transaction detection: a review," *Knowledge and Information Systems*, vol. 57, no. 2, pp. 245–285, 2018.

[3] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *IEEE International Conference on Data Mining*, 2008, pp. 413–422.

[4] X. Song, M. Wu, C. Jermaine, and S. Ranka, “Conditional anomaly detection,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 5, pp. 631–645, 2007.

[5] J. Liang and S. Parthasarathy, “Robust contextual outlier detection: Where context meets sparsity,” in *ACM International Conference on Information and Knowledge Management*, 2022, pp. 1183–1192.

[6] L. Zhong *et al.*, “Qcad: Quantile-based conditional anomaly detection,” *arXiv preprint arXiv:2306.00000*, 2023.

[7] E. Calikus *et al.*, “Context discovery for anomaly detection,” *International Journal of Data Science and Analytics*, 2024.

[8] H. Xu, G. Pang, Y. Wang, and Y. Wang, “Deep isolation forest for anomaly detection,” in *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 12, 2023, pp. 12 591–12 604.

[9] B. Oztas *et al.*, “Enhancing anti-money laundering: Development of a synthetic transaction monitoring dataset,” in *IEEE International Conference on e-Business Engineering*, 2023.

[10] E. A. Lopez-Rojas, A. Elmir, and S. Axelsson, “Paysim: A financial mobile money simulator for fraud detection,” *European Modeling and Simulation Symposium*, 2016.

[11] A. Dal Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi, “Calibrating probability with undersampling for unbalanced classification,” *IEEE Symposium Series on Computational Intelligence*, 2015.