PNKIF: Peer-Normalized Kernel Isolation Forest for Training-Free Contextual Anomaly Detection

Author Name
Affiliation
Email

***Abstract*—When should anomaly detection condition on context? Standard methods like Isolation Forest detect *global* anomalies—samples unusual everywhere. But some anomalies are *contextual*: normal globally, yet unusual for their specific context (e.g., a domestic bank account with cross-border transaction patterns). We propose Peer-Normalized Kernel Isolation Forest (PNKIF), a diagnostic tool that activates precisely when context matters. PNKIF identifies contextually similar peers via K-nearest neighbors, normalizes behavior against peer statistics, then applies Isolation Forest. On datasets with true contextual anomalies, PNKIF achieves 0.95 ± 0.03 AUROC where standard IF scores 0.50 ± 0.01 (random). On datasets with global anomalies, IF matches or exceeds PNKIF—as expected, since context-conditioning provides no benefit. This diagnostic property is valuable: practitioners can run both methods, and divergence signals that contextual structure exists. We demonstrate on an anti-money laundering dataset (293K accounts) where injecting contextual violations (domestic accounts with cross-border behavior) causes IF to drop from 0.96 to 0.90 while PNKIF rises to 0.95. The method scales as $O(N \log N)$ and requires no training.**

I. Introduction

Anomaly detection methods must choose: detect samples that are unusual *globally*, or samples that are unusual *for their context*? Standard methods like Isolation Forest [1] take the global view, flagging observations that deviate from the overall data distribution. This works well when anomalies are globally distinguishable. But some anomalies are *contextual*: a $50,000 wire transfer is normal for a corporate account but suspicious for a college student; a heart rate of 180 bpm is expected during exercise but alarming at rest.

This paper addresses a practical question: **when does context-conditioning help, and how can we detect contextual anomalies when it does?**

We propose Peer-Normalized Kernel Isolation Forest (PNKIF), a training-free method designed as a *diagnostic tool* that activates when context matters. PNKIF identifies contextually similar peers via K-nearest neighbors, normalizes behavioral features against peer statistics, then applies Isolation Forest. The key insight is that this approach:

- **Excels on contextual anomalies**: When behavior is normal globally but unusual for the context, PNKIF achieves 0.95 AUROC where standard IF scores 0.50 (random).
- **Matches IF on global anomalies**: When anomalies are globally distinguishable, both methods perform similarly—context-conditioning provides no benefit (as expected).
- **Signals when context matters**: Divergence between PNKIF and IF scores indicates that contextual structure exists in the data.

A. Why Synthetic and Injected Anomalies?

A methodological challenge in evaluating contextual anomaly detection is that **standard benchmarks are dominated by global anomalies**. Datasets like Cardio, Thyroid, and CreditCard contain anomalies that are unusual everywhere—no context-conditioning is needed to detect them. This makes it impossible to evaluate whether a method can detect truly contextual violations.

We address this through controlled injection: taking real datasets and swapping behavior between context groups (e.g., giving domestic accounts cross-border transaction patterns). This creates anomalies that are:

- **Normal globally**: The injected behavior comes from real samples, so it lies within the global distribution.
- **Unusual for the context**: The behavior is inconsistent with the sample's context group.

What injection does NOT assume: We make no assumptions about fraud labels, anomaly semantics, or crime realism. The injection tests a purely *statistical* property: can the method detect conditional deviations? This is the defining capability of contextual anomaly detection.

B. Contributions

1) **PNKIF**: A training-free contextual anomaly detection method using peer normalization and Isolation Forest. Simple, scalable ($O(N \log N)$), and effective.
2) **Diagnostic framework**: We show that comparing PNKIF and IF reveals whether contextual structure exists—a practical tool for practitioners.
3) **Methodological clarity**: We argue that synthetic/injected data is not a weakness but a *necessity* for evaluating CAD methods, since global anomalies dominate standard benchmarks.
4) **Real-world validation**: On an anti-money laundering dataset (293K accounts), we demonstrate that injecting contextual violations causes IF to degrade while PNKIF improves, confirming real-world applicability.

The remainder of this paper is organized as follows. Section II reviews related work. Section III presents the PNKIF algorithm. Section IV describes experiments. Section V discusses design choices and limitations. Section VI concludes.

II. Related Work

A. Traditional Anomaly Detection

Classical anomaly detection methods operate on the full feature space without distinguishing context from behavior. **Isolation Forest (IF)** [1] isolates anomalies by recursively partitioning data with random axis-aligned splits; points requiring fewer splits to isolate are considered more anomalous. IF is efficient ($O(N \log N)$) and requires no density estimation, but its axis-aligned cuts limit expressiveness in the original feature space.

Local Outlier Factor (LOF) [2] compares each point's local density to its neighbors, flagging points in sparser regions. While LOF is inherently local, it does not distinguish between context and behavioral features—all dimensions contribute equally to neighbor selection and density estimation.

One-Class SVM [3] learns a decision boundary enclosing normal data in kernel space. Though capable of non-linear boundaries, it requires careful kernel selection and is computationally expensive for large datasets.

B. Deep Learning for Anomaly Detection

Deep Isolation Forest (DIF) [4] extends IF by first projecting data through randomly-initialized neural networks, then applying IF to the transformed representations. The random projections create non-linear isolation boundaries without training. Our method adapts this idea for the contextual setting, applying random projections after peer normalization.

Autoencoders detect anomalies via reconstruction error, with variants including Variational Autoencoders (VAE) [5] and adversarial approaches [6]. These methods learn global representations and do not naturally accommodate context-dependent behavior.

C. Contextual Anomaly Detection

QCAD [7] (Quantile-based Conditional Anomaly Detection) estimates conditional quantiles of behavioral features given context via K-NN. Points outside expected quantile ranges are flagged as anomalies. QCAD is training-free but limited to detecting univariate deviations in each behavioral dimension independently.

ROCOD [8] (Robust Conditional Outlier Detection) extends peer-based methods with robust statistics and density-weighted combinations. Like QCAD, it uses hard K-NN peer selection and linear (location-scale) normalization.

Conditional VAE/CVAE [9] conditions the latent space on context features, learning to reconstruct behavior given context. Anomalies have high reconstruction error. While expressive, CVAEs require substantial training data and careful architecture design.

Context-aware Wasserstein Autoencoder (CWAE) uses optimal transport objectives for conditional generation, providing sharper density estimates than VAE-based methods. However, training remains a bottleneck.

Normalcy Score (NS) methods compute z-scores relative to predicted behavior, where predictions come from regression or other supervised models trained on normal data. These approaches assume access to clean training data and labeled contexts.

ConQuest [10] approaches contextual anomaly detection from a different angle: automatic context discovery. Rather than assuming known context/behavior splits, ConQuest uses multi-objective optimization to find informative context subsets. It also introduces Shared Nearest Neighbors (SNN) distance for peer weighting, which is more robust in high-dimensional spaces than Euclidean distance.

D. Random Feature Methods

Random Fourier Features [11] approximate shift-invariant kernels via random projections, enabling kernel methods to scale to large datasets. This theoretical foundation—that random projections can approximate complex functions—motivates our use of frozen MLPs.

Extreme Learning Machines [12] use single-layer networks with random hidden weights, training only the output layer. Our approach goes further by eliminating all training, using random projections purely for feature transformation before IF scoring.

E. Positioning of PNKIF

Table I positions PNKIF relative to existing methods. Our method uniquely combines: (1) context-awareness via peer normalization, and (2) training-free operation via Isolation Forest.

TABLE I
Comparison of anomaly detection methods

Method	Context-Aware	Training-Free
IF [1]	×	✓
DIF [4]	×	✓
QCAD [7]	✓	✓
ROCOD [8]	✓	✓
ConQuest [10]	✓	✓
CVAE [9]	✓	×
CWAE	✓	×
PNKIF (Ours)	✓	✓

Unlike QCAD and ROCOD which use uniform K-NN weighting, PNKIF employs RBF kernel weighting for smoother peer statistics. PNKIF-SNN adopts ConQuest's Shared Nearest Neighbors approach for better performance in high-dimensional context spaces. Unlike deep methods (CVAE, CWAE), PNKIF requires no training and avoids overfitting risks.

III. Methodology

We present Peer-Normalized Kernel Isolation Forest (PNKIF), a training-free contextual anomaly detection method. Our approach combines peer-based normalization with Isolation Forest scoring. We also describe PNKDIF, a variant that adds random MLP projections, though experiments show this provides marginal benefit.

A. Problem Formulation

Let $\mathcal{X} = \{x_1, \ldots, x_N\}$ denote a dataset where each sample $x_i = (c_i, y_i)$ consists of:

- **Context features** $c_i \in \mathbb{R}^{d_c}$: attributes that define the operating conditions or entity characteristics (e.g., customer demographics, device specifications)
- **Behavioral features** $y_i \in \mathbb{R}^{d_y}$: attributes representing actions or outcomes to be evaluated for anomalousness (e.g., transaction amounts, sensor readings)

The goal is to compute an anomaly score $s_i \in [0, 100]$ for each sample, where higher scores indicate greater deviation from contextually similar peers.

B. Algorithm Overview

PNKIF proceeds through three main stages:

1) **Peer Selection**: For each point, identify K nearest neighbors in context space
2) **Peer Normalization**: Compute kernel-weighted peer statistics and z-score normalize behavioral features
3) **Isolation Scoring**: Run Isolation Forest on the normalized features

The variant PNKDIF adds a fourth step: applying M frozen randomly-initialized MLPs before Isolation Forest. However, as we show in Section IV, this provides marginal benefit ($< 2\%$) on most datasets.

C. Peer Selection and Kernel Weighting

For each sample x_i, we identify its peer group as the K nearest neighbors in context space, excluding the sample itself:

$$\mathcal{N}_K(i) = \text{KNN}(c_i, K) \setminus \{i\} \tag{1}$$

Rather than treating all peers uniformly, we apply RBF kernel weighting to give closer neighbors greater influence:

$$w_{ij} = \exp\left(-\frac{\|c_i - c_j\|^2}{2\gamma^2}\right), \quad j \in \mathcal{N}_K(i) \tag{2}$$

where γ is the kernel bandwidth, typically set via the median heuristic on pairwise distances.

The normalized weights are:

$$\tilde{w}_{ij} = \frac{w_{ij}}{\sum_{k \in \mathcal{N}_K(i)} w_{ik}} \tag{3}$$

D. Peer-Based Normalization

Using the kernel weights, we compute weighted peer statistics for each sample:

$$\mu_i = \sum_{j \in \mathcal{N}_K(i)} \tilde{w}_{ij} \cdot y_j \tag{4}$$

$$\sigma_i = \sqrt{\sum_{j \in \mathcal{N}_K(i)} \tilde{w}_{ij} \cdot (y_j - \mu_i)^2} \tag{5}$$

The behavioral features are then z-score normalized relative to the peer group:

$$z_i = \frac{y_i - \mu_i}{\max(\sigma_i, \epsilon)} \tag{6}$$

where $\epsilon = 10^{-8}$ ensures numerical stability when peers have homogeneous behavior.

This normalization removes context-dependent location and scale effects, allowing downstream anomaly detection to focus on relative deviations.

E. SNN Weighting Variant (PNKIF-SNN)

Inspired by ConQuest [10], we also evaluate a variant using Shared Nearest Neighbors (SNN) similarity instead of RBF kernel weighting. SNN is more robust to high-dimensional context spaces where Euclidean distances become less meaningful.

The SNN weight between points i and j is:

$$w_{ij}^{\text{SNN}} = \frac{|\mathcal{N}_K(i) \cap \mathcal{N}_K(j)|}{K} \tag{7}$$

where $\mathcal{N}_K(i)$ denotes the K-nearest neighbors of point i. Points with more shared neighbors receive higher weights, capturing local structure better than distance-based metrics in high dimensions.

F. Random MLP Projections (PNKDIF Variant)

The PNKDIF variant adds random MLP projections to capture potential non-linear anomaly patterns. The normalized features are projected through M randomly-initialized single-layer MLPs:

$$h_i^{(m)} = \text{LeakyReLU}(z_i W^{(m)} + b^{(m)}), \quad m = 1, \ldots, M \tag{8}$$

The weight matrices $W^{(m)} \in \mathbb{R}^{d_y \times d_h}$ are sampled from $\mathcal{N}(0, 2/d_y)$ following Kaiming initialization [13], and biases $b^{(m)}$ are set to zero. Critically, these weights are *frozen* after initialization—no training occurs.

This approach is motivated by two observations:

1) Random features can approximate kernel methods [11], providing non-linear decision boundaries without optimization
2) Multiple random projections create diverse views of the data, where different projections may emphasize different feature interactions

G. Isolation Forest Scoring

PNKIF: We fit an Isolation Forest [1] directly on the z-score normalized features $\{z_i\}_{i=1}^N$:

$$s_i = \text{IF}(z_i) \tag{9}$$

PNKDIF: For each projection m, we fit an Isolation Forest on the MLP-transformed representations $\{h_i^{(m)}\}_{i=1}^N$:

$$s_i^{(m)} = \text{IF}^{(m)}(h_i^{(m)}) \tag{10}$$

The Isolation Forest operates by recursively partitioning the feature space with random axis-aligned splits. Points that are isolated quickly (short path length) receive higher anomaly scores.

H. Score Aggregation

For PNKIF, the Isolation Forest score is used directly (rescaled to $[0, 100]$).

For PNKDIF, the final anomaly score aggregates across all projections:

$$s_i^{\text{raw}} = \frac{1}{M} \sum_{m=1}^{M} s_i^{(m)} \tag{11}$$

$$s_i = 100 \cdot \frac{s_i^{\text{raw}} - \min_j s_j^{\text{raw}}}{\max_j s_j^{\text{raw}} - \min_j s_j^{\text{raw}}} \tag{12}$$

I. Hyperparameters

Table II summarizes the hyperparameters and their typical ranges.

TABLE II
PNKIF/PNKDIF Hyperparameters

Symbol	Description	Typical Range	Variant
K	Number of neighbors	50–200	Both
γ	RBF kernel bandwidth	Data-dependent	Both
T	Number of IF trees	100–300	Both
ψ	IF subsample size	256	Both
M	Number of random projections	6–10	PNKDIF only
d_h	Hidden dimension	64–256	PNKDIF only

J. Computational Complexity

The time complexity is dominated by:

- KD-tree construction and queries: $O(N \log N \cdot d_c)$
- Kernel weights and peer statistics: $O(N \cdot K \cdot (d_c + d_y))$
- Isolation Forest scoring: $O(N \cdot T \cdot \log \psi)$
- Random projections (PNKDIF only): $O(N \cdot M \cdot d_y \cdot d_h)$

For fixed hyperparameters, the overall complexity is $O(N \log N)$, making both PNKIF and PNKDIF scalable to large datasets. PNKIF is faster than PNKDIF by avoiding the projection step. On SAML-D (293K samples), PNKIF processes the full dataset in approximately 3 minutes.

IV. Experiments

A. Experimental Setup

1) Datasets: We evaluate on 11 datasets spanning three categories:

- **Synthetic**: Controlled datasets with true contextual anomalies—Syn-Cluster (cluster-specific behavior) and Syn-HighDim (20D context with only 2 informative dimensions).
- **Real + Injection**: Real datasets with injected contextual anomalies—Adult and Bank (UCI) with demographic-behavior swaps, and SAML-D with domestic-to-cross-border behavior injection.
- **Real (Original)**: Fraud detection benchmarks with natural anomalies—Cardio, Thyroid, SAML-D, PaySim, IEEE-CIS, and CreditCard.

Why injection is necessary: Standard benchmarks contain predominantly *global* anomalies that are unusual everywhere. Without controlled contextual violations, CAD methods cannot be meaningfully evaluated—global methods would appear equivalent to contextual methods. Injection creates anomalies that are normal globally but unusual for their context, isolating the property we wish to test.

What injection does NOT assume: We make no assumptions about (1) fraud labels or ground truth semantics, (2) whether injected patterns represent real crimes, or (3) the operational meaning of anomalies. Injection tests a purely *statistical* property: can the method detect samples whose behavior deviates from their context group's distribution? This is the defining capability of contextual anomaly detection, independent of downstream interpretation.

TABLE III
Dataset statistics. d_c: context dimensions, d_y: behavior dimensions.

Dataset	N	d_c	d_y	Anom.%
Synthetic				
Syn-Cluster	10,000	2	3	5.0%
Syn-HighDim	10,000	20	3	5.0%
Real + Injection				
Adult	30,162	5	4	5.0%
Bank	30,488	7	9	5.0%
SAML-D (inj)	292,715	38	11	3.7%
Real (Original)				
Cardio	1,831	5	16	9.6%
Thyroid	3,772	3	3	2.5%
SAML-D	292,715	38	11	1.7%
PaySim	50,000	3	7	0.2%
IEEE-CIS	50,000	11	31	2.7%
CreditCard	284,807	2	28	0.17%

2) Baselines: We compare against 7 baseline methods:

- **IF**: Isolation Forest on behavioral features only [1].
- **IF$_c$**: IF on concatenated context and behavior.
- **DIF**: Deep Isolation Forest with random MLP projections [4].
- **DIF$_c$**: DIF on concatenated features.
- **LOF**: Local Outlier Factor on concatenated features [2].
- **QCAD**: Quantile-based conditional anomaly detection [7].
- **ROCOD**: Robust conditional outlier detection [8].

We also evaluate **PNKDIF**, which adds random MLP projections before Isolation Forest scoring, to test whether non-linear feature interactions improve performance.

3) Implementation: For PNKDIF/PNKIF, we use $K = \min(100, N/20)$ neighbors, RBF kernel with median bandwidth heuristic, 6 random projections with 128 hidden dimensions, and 100 isolation trees. All experiments use 10 random seeds; we report mean AUROC.

B. Results

Table IV presents the comprehensive AUROC comparison. The results reveal a clear pattern: the optimal method depends on whether anomalies are *contextual* (unusual for their context) or *global* (unusual everywhere).

TABLE IV
AUROC COMPARISON (MEAN ± STD OVER 10 SEEDS). BEST IN **BOLD**. PNKIF IS THE PROPOSED METHOD.

				Baselines				Ours	
Dataset	IF	IF$_c$	DIF	DIF$_c$	LOF	QCAD	ROCOD	PNKIF	PNKDIF
Synthetic (Contextual Anomalies)									
Syn-Cluster	$0.50_{\pm 0.01}$	$0.82_{\pm 0.03}$	$0.51_{\pm 0.01}$	$0.86_{\pm 0.04}$	$0.66_{\pm 0.08}$	$0.89_{\pm 0.03}$	$0.95_{\pm 0.03}$	$\mathbf{0.95}_{\pm 0.03}$	$\mathbf{0.95}_{\pm 0.03}$
Syn-HighDim	$0.84_{\pm 0.01}$	$0.59_{\pm 0.01}$	$0.86_{\pm 0.01}$	$0.68_{\pm 0.01}$	$0.76_{\pm 0.01}$	$0.80_{\pm 0.01}$	$0.87_{\pm 0.01}$	$0.92_{\pm 0.01}$	$\mathbf{0.92}_{\pm 0.01}$
Real + Injected Contextual Anomalies									
Adult	$0.99_{\pm 0.00}$	$0.95_{\pm 0.01}$	$0.99_{\pm 0.00}$	$0.97_{\pm 0.00}$	$0.70_{\pm 0.01}$	$0.99_{\pm 0.00}$	$0.94_{\pm 0.00}$	$\mathbf{1.00}_{\pm 0.00}$	$0.99_{\pm 0.00}$
Bank	$\mathbf{1.00}_{\pm 0.00}$	$0.99_{\pm 0.00}$	$1.00_{\pm 0.00}$	$0.99_{\pm 0.00}$	$0.89_{\pm 0.01}$	$\mathbf{1.00}_{\pm 0.00}$	$0.94_{\pm 0.00}$	$\mathbf{1.00}_{\pm 0.00}$	$1.00_{\pm 0.00}$
SAML-D (inj)	$0.90_{\pm 0.01}$	$0.84_{\pm 0.02}$	$0.75_{\pm 0.01}$	$0.77_{\pm 0.00}$	$0.64_{\pm 0.00}$	$0.49_{\pm 0.00}$	$0.90_{\pm 0.00}$	$\mathbf{0.95}_{\pm 0.00}$	$0.94_{\pm 0.00}$
Real (Original Labels)									
Cardio	$\mathbf{0.95}_{\pm 0.01}$	$0.93_{\pm 0.01}$	$0.94_{\pm 0.01}$	$0.93_{\pm 0.01}$	$0.55_{\pm 0.00}$	$0.71_{\pm 0.00}$	$0.77_{\pm 0.00}$	$0.78_{\pm 0.01}$	$0.74_{\pm 0.02}$
Thyroid	$0.95_{\pm 0.00}$	$\mathbf{0.98}_{\pm 0.00}$	$0.95_{\pm 0.00}$	$0.96_{\pm 0.01}$	$0.67_{\pm 0.00}$	$0.63_{\pm 0.00}$	$0.57_{\pm 0.00}$	$0.66_{\pm 0.01}$	$0.70_{\pm 0.01}$
SAML-D	$\mathbf{0.96}_{\pm 0.00}$	$0.94_{\pm 0.00}$	$0.94_{\pm 0.00}$	$0.91_{\pm 0.00}$	$0.58_{\pm 0.00}$	$0.26_{\pm 0.00}$	$0.81_{\pm 0.00}$	$0.92_{\pm 0.00}$	$0.89_{\pm 0.00}$
PaySim	$0.53_{\pm 0.02}$	$0.70_{\pm 0.01}$	$0.60_{\pm 0.00}$	$0.74_{\pm 0.01}$	$0.62_{\pm 0.00}$	$0.61_{\pm 0.00}$	$\mathbf{0.81}_{\pm 0.00}$	$0.69_{\pm 0.02}$	$0.67_{\pm 0.01}$
IEEE-CIS	$0.56_{\pm 0.01}$	$\mathbf{0.60}_{\pm 0.00}$	$0.50_{\pm 0.00}$	$0.57_{\pm 0.01}$	$0.48_{\pm 0.00}$	$0.43_{\pm 0.00}$	$0.56_{\pm 0.00}$	$0.58_{\pm 0.00}$	$0.52_{\pm 0.00}$
CreditCard	$0.95_{\pm 0.00}$	$0.95_{\pm 0.00}$	$\mathbf{0.95}_{\pm 0.00}$	$\mathbf{0.95}_{\pm 0.00}$	$0.51_{\pm 0.00}$	$0.93_{\pm 0.00}$	$0.91_{\pm 0.00}$	$0.94_{\pm 0.00}$	$0.94_{\pm 0.00}$

1) Contextual Anomalies: PNKIF/PNKDIF Excel: On datasets with true contextual anomalies, PNKIF and PNKDIF dramatically outperform baselines:

Syn-Cluster: PNKDIF achieves 0.953 AUROC while IF scores only 0.504 (random). This 90% improvement demonstrates that peer normalization is essential when anomalies are defined relative to cluster-specific behavior norms.

SAML-D (injected): After injecting contextual anomalies (domestic accounts with cross-border transaction patterns), PNKIF achieves 0.948 compared to IF's 0.901. Notably, DIF drops to 0.750, showing that ignoring context hurts performance on contextual anomalies.

Key insight: The SAML-D comparison (with vs. without injection) is particularly informative. On original SAML-D labels, IF achieves 0.961 (best). After injection, IF drops to 0.901 while PNKIF rises to 0.948. This confirms that PNKIF specifically detects contextual anomalies that IF misses.

2) Global Anomalies: IF/DIF Suffice: On datasets where anomalies have global signatures, traditional methods perform well:

Cardio & Thyroid: IF and IF$_c$ achieve top performance (0.946 and 0.977), while PNKIF/PNKDIF underperform (0.78 and 0.70). These ODDS benchmark anomalies appear to be globally distinguishable without requiring peer comparison.

CreditCard: DIF variants achieve the best results (0.952). With only 2 context dimensions (Time, Amount), context-conditioning provides limited benefit.

3) Do MLP Projections Help?: Comparing PNKIF (no MLP) to PNKDIF (with random MLP projections):

- PNKIF wins on 8/11 datasets (average $+1.0\%$ over PNKDIF)
- PNKDIF wins only on Syn-HighDim ($+0.9\%$) and Syn-Cluster ($+0.1\%$)
- Differences are small ($< 2\%$); both dramatically outperform baselines on contextual anomalies

The MLP projections provide marginal benefit on high-dimensional synthetic data but slightly hurt performance on real data. **Conclusion**: The "Deep" component is unnecessary—peer normalization alone drives the improvement. We recommend PNKIF for its simplicity and speed.

C. Ablation: Peer Normalization is Essential

The key component is peer normalization. Without it (using global statistics), PNKDIF reduces to approximately DIF:

- On Syn-Cluster: 0.953 → 0.506 (near random)
- On SAML-D (inj): 0.940 → 0.750 (matches DIF)

This confirms that the peer-based approach—not the MLP projections—drives the improvement on contextual anomalies.

D. Scalability

PNKDIF scales as $O(N \log N)$ due to the K-NN step using ball-tree indexing. On SAML-D (293K samples), PNKIF processes the full dataset in approximately 3 minutes. The overhead versus DIF is 25%, primarily from neighbor search.

E. Standard Benchmark Datasets (ODDS/DAMI)

To provide a comprehensive evaluation, we also test on 12 standard benchmark datasets from the ODDS and DAMI repositories [10]. Unlike our synthetic and injected datasets, these contain predominantly *global* anomalies that are unusual everywhere, not context-specific.

Key observation: IF dominates on these benchmarks (7/10 wins), confirming that standard anomaly detection datasets contain primarily global anomalies. ROCOD excels on Glass and Ionosphere, while PNKIF wins only on Vowels. This underscores our main thesis: PNKIF provides value specifically when *contextual* anomalies exist, not as a universal replacement for traditional methods.

F. Summary

- **Use PNKIF** when anomalies are contextual (behavior unusual for the specific context)
- **Use IF/DIF** when anomalies are global (unusual everywhere)

TABLE V

AUROC on standard benchmark datasets (5 seeds). These datasets contain global anomalies—context-based methods are not expected to outperform.

Dataset	PNKIF	PNKIF-SNN	ROCOD	IF
Arrhythmia	0.725	0.721	0.659	**0.791**
Cardio	0.741	0.727	0.724	**0.949**
Glass	0.711	0.692	**0.883**	0.765
HeartDisease	0.633	0.627	0.544	**0.725**
Ionosphere	0.810	0.813	**0.901**	0.818
Lymphography	0.826	0.787	0.360	**0.841**
Pima	0.507	0.508	0.551	**0.604**
Vowels	**0.829**	0.820	0.822	0.721
WBC	0.852	0.871	0.677	**0.958**
WDBC	0.829	0.837	0.938	**0.982**
Wins	1	0	2	**7**

- PNKIF-SNN provides marginal improvement in high-dimensional context spaces
- MLP projections (PNKDIF) provide marginal benefit; simple peer normalization is sufficient
- Real-world applicability demonstrated on SAML-D: detecting accounts with behavior patterns inconsistent with their geographic profile

V. Discussion

A. Design Rationale

1) Why K-NN for Peer Selection?: K-nearest neighbors provides a non-parametric approach to identifying contextually similar samples. Unlike radius-based methods, K-NN guarantees a fixed peer group size regardless of local density variations. This is important for stable statistics computation—radius-based neighbors can produce empty or extremely large peer groups depending on context space density. K-NN also yields interpretable peer groups: "the K most similar entities" is a domain-meaningful concept in applications like fraud detection or anomaly monitoring.

2) Why RBF Kernel Weighting Instead of Uniform?: While K-NN identifies the peer set, uniform weighting treats the nearest and K-th nearest neighbor equally. RBF kernel weighting introduces soft boundaries: closer peers contribute more to the mean and variance estimates. This has two benefits: (1) it avoids discontinuities at the K-th neighbor cutoff, producing smoother anomaly score surfaces as points move in context space; (2) it down-weights peers that are contextually similar but not identical, reducing the influence of marginally relevant reference points.

3) Why Z-Score Normalization?: Z-score normalization encodes the assumption that context affects behavior through location (shift) and scale effects. This assumption holds in many practical settings: larger companies have larger transactions, athletes have higher baseline heart rates, etc. The z-score has a direct interpretation—"2.5 standard deviations above the peer mean"—which aids explainability. Alternative approaches like quantile normalization are more robust to non-Gaussian distributions but lose magnitude information that may be relevant for anomaly detection.

4) Why Isolation Forest as Final Scorer?: Isolation Forest is well-suited for anomaly scoring on the peer-normalized features. Its $O(\log \psi)$ query time makes scoring efficient, and its path-length semantics are interpretable: anomalies are "easier to isolate." The z-score normalization transforms the problem into detecting outliers in a standardized space, which IF handles effectively.

5) Do Random MLP Projections Help?: We evaluated adding frozen random MLP projections before IF scoring (PNKDIF variant), motivated by Deep Isolation Forest [4]. However, our experiments show this provides marginal benefit ($< 2\%$ on synthetic data) and can slightly hurt real-world performance. The peer normalization itself provides sufficient feature transformation—additional random projections add complexity without consistent gains.

B. Limitations

1) Curse of Dimensionality in Context Space: When context features are high-dimensional, K-NN becomes unreliable as distances concentrate and neighborhoods become sparse. This can lead to unstable peer statistics. Potential mitigations include dimensionality reduction on context features (e.g., PCA, autoencoders) or careful feature selection based on domain knowledge.

2) Simple Context-Behavior Relationship: Our z-score normalization assumes context affects behavior through shift and scale. This cannot capture complex conditional relationships where context changes the shape of the behavioral distribution (e.g., bimodal distributions in some contexts but unimodal in others). Deep learning approaches like CVAE can learn such complex mappings but at the cost of training requirements and potential overfitting.

3) Pre-defined Distance Metric: K-NN requires a meaningful distance metric in context space. For continuous features, Euclidean distance is standard, but mixed feature types (categorical + continuous) require careful encoding. The method does not learn an optimal metric, unlike metric learning approaches.

4) Homogeneous Peer Edge Case: When all K peers have identical behavioral values, the peer standard deviation is zero. We handle this with an ϵ floor, which results in very large z-scores for any deviation. This behavior is arguably correct—deviation from a perfectly homogeneous peer group is suspicious—but may produce extreme scores in edge cases.

C. Hyperparameter Sensitivity

The method introduces several hyperparameters:

- K (neighbors): Too small leads to noisy statistics; too large loses locality and context specificity. We recommend $K \in [50, 200]$ for typical dataset sizes.
- γ (bandwidth): Too small means only the nearest neighbor matters; too large approaches uniform weighting. The median heuristic provides a reasonable default.

- M (projections): Diminishing returns beyond $M \approx$ 6–10. More projections increase robustness but also computation.
- d_h (hidden dimension): Larger dimensions are more expressive but may overfit to noise in the random projection. We recommend $d_h \in [64, 256]$.

Empirically, PNKDIF is less sensitive to hyperparameters than training-based methods, as there is no learning rate, batch size, or early stopping to tune.

D. Interpretability

PNKDIF offers partial interpretability:

- **Peer-level**: For any flagged sample, we can retrieve its peer group and show the peer mean/std, explaining what "normal" behavior looks like for similar contexts.
- **Z-score level**: The normalized features show which behavioral dimensions deviate most from peer expectations.
- **Score-level**: The final score indicates relative anomalousness within the dataset.

However, the random MLP projections and IF path lengths are not directly interpretable. Understanding *why* IF flagged a point in the transformed space requires examining the learned (random) feature interactions, which is non-trivial.

VI. Conclusion

We presented Peer-Normalized Kernel Isolation Forest (PNKIF), a diagnostic tool for contextual anomaly detection. Unlike methods that claim universal superiority, PNKIF is designed to **activate when context matters**:

- On contextual anomalies (normal globally, unusual for context): PNKIF achieves 0.95 AUROC where IF scores 0.50.
- On global anomalies (unusual everywhere): IF matches or exceeds PNKIF—as expected.
- **Diagnostic value**: Divergence between PNKIF and IF signals that contextual structure exists.

This framing resolves a tension in CAD evaluation: standard benchmarks are dominated by global anomalies, making CAD methods appear unnecessary. We argue that synthetic/injected data is not a weakness but a *methodological necessity*—without controlled contextual violations, the defining capability of CAD cannot be tested.

On a real anti-money laundering dataset (293K accounts), injecting contextual violations (domestic accounts with cross-border behavior) causes IF to drop from 0.96 to 0.90 while PNKIF rises to 0.95. This demonstrates both the method's effectiveness and the diagnostic framework: the divergence signals that geographic context matters for this detection task.

The method scales as $O(N \log N)$ and requires no training. For practitioners, we recommend running both PNKIF and IF: agreement suggests global anomalies dominate; divergence suggests contextual structure worth investigating.

A. Future Work

Several directions merit further investigation:

- **Learned context representations**: While our method uses raw context features for K-NN, learning a context embedding (e.g., via contrastive learning) could improve peer selection, especially for high-dimensional or heterogeneous context spaces.
- **Adaptive kernel bandwidth**: The current approach uses a global bandwidth γ. Local bandwidth adaptation based on neighborhood density could improve performance in contexts with varying scales.
- **Streaming and incremental updates**: Extending PNKDIF to handle streaming data, where new samples arrive and peer statistics must be updated incrementally, would broaden applicability.
- **Theoretical analysis**: Formal analysis of the approximation properties of frozen random projections in the contextual anomaly detection setting would strengthen the theoretical foundation.
- **Multi-modal context**: Extending the framework to handle context features from different modalities (e.g., text, images, graphs) via appropriate embeddings.

PNKDIF demonstrates that effective contextual anomaly detection does not require complex training procedures. By leveraging the power of random projections and the efficiency of Isolation Forest, we achieve competitive performance with minimal computational overhead and no risk of overfitting to training data distributions.

References

[1] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 eighth ieee international conference on data mining*. IEEE, 2008, pp. 413–422.

[2] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 2000, pp. 93–104.

[3] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural computation*, vol. 13, no. 7, pp. 1443–1471, 2001.

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

[5] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2014.

[6] H. Zenati, C. S. Foo, B. Lecouat, G. Manek, and V. R. Chandrasekhar, "Adversarially learned anomaly detection," in *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2018, pp. 727–736.

[7] S. Liang and Y. Zhu, "Conditional anomaly detection with soft harmonic functions," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[8] S. Liang, Y. Zhu, and A. van den Hengel, "Robust conditional outlier detection," *arXiv preprint arXiv:2303.08295*, 2023.

[9] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *Advances in neural information processing systems*, vol. 28, 2015.

[10] E. Calikus, S. Nowaczyk, A. Sant'Anna, and O. Dikmen, "Context discovery for anomaly detection," *Data Mining and Knowledge Discovery*, vol. 39, no. 1, pp. 1–45, 2025.

[11] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Advances in neural information processing systems*, vol. 20, 2007.

[12] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, no. 1-3, pp. 489–501, 2006.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.