

Federated Bounded-Memory Anytime-Valid Kernel Independence Testing for Structural Equation Models in Non-Stationary IoT Anomaly Detection

Rahul Misra

Department of Computer Science and Engineering, Jagannath University, Jaipur, Rajasthan, India

Published October 2026

Abstract

Federated Internet of Things deployments require anomaly detectors that respect strict memory limits while remaining statistically valid under non-stationary data streams. Conventional kernel independence tests, such as the Hilbert-Schmidt Independence Criterion, demand the retention of full Gram matrices, and their batch formulation cannot support sequential decision making; consequently, they are ill-suited to edge devices whose distributions drift over time. We propose a federated bounded-memory framework for anytime-valid kernel independence testing within structural equation models. Each node first embeds sensor variables and model residuals into a finite-dimensional Euclidean space through random Fourier features, and then maintains a decayed cross-covariance sketch whose memory footprint scales only with the square of the feature dimension. Anytime-valid e-values are constructed from a martingale wealth process driven by betting fractions, so that the false-alarm rate is controlled at a prescribed level irrespective of the stopping time. Furthermore, only quantized adjacency updates, compressed covariance sketches, and sample counts are transmitted to the server, where a weighted consensus update refines the global causal graph and shared structural parameters. The proposed method replaces the batch-based conditional independence tester in the causal inference engine, and its continuous e-value output feeds the anomaly scoring module and the federated graph coordinator, thereby closing the loop between causal discovery and equation fitting. Experiments on non-stationary IoT benchmarks assess detection accuracy, communication cost and memory consumption. The framework attains constant memory usage and linear per-step time complexity, and it therefore offers a practical route to rigorous, privacy-preserving causal anomaly detection at the network edge.

Keywords: federated learning; kernel independence testing; structural equation models; non-stationary IoT anomaly detection; anytime-valid inference

1 Introduction

Anomaly detection in Internet of Things (IoT) environments has gradually shifted from purely statistical outlier identification toward causal reasoning, because causal structures allow a detector to distinguish genuine system faults from benign distributional shifts [1]. Structural Equation Models (SEMs) provide a natural language for this task, since they encode directed dependencies among sensor variables and permit residual-based diagnosis [2]. Classical causal discovery algorithms, including PC, FCI, GES, and LiNGAM, nevertheless assume linearity or strong distributional regularity, and their batch nature conflicts with the non-stationary streams observed in modern IoT deployments [3].

Kernel-based independence tests such as the Hilbert-Schmidt Independence Criterion (HSIC) and the Kernel Conditional Independence (KCI) test alleviate the linearity restriction and provide rigorous

statistical guarantees [4]. However, these tests require the full Gram matrix, which incurs quadratic time and memory complexity, and they are formulated for a fixed sample size. Both properties are prohibitive on low-power edge devices, and neither supports continuous monitoring under drifting distributions. Random Fourier features offer an explicit finite-dimensional approximation of shift-invariant kernels, reducing kernel computations to inner products in a low-dimensional Euclidean space [4]. Anytime-valid inference, built on e-values and martingale wealth processes, permits sequential testing without a pre-specified sample size and preserves false-alarm control under optional stopping [2].

Federated learning has become the standard paradigm for privacy-preserving collaborative analytics in resource-constrained networks [5]. Secure aggregation and its variants allow nodes to share model updates without exposing raw data, and recent studies have begun to couple federated optimization with causal graph learning [6]. Most of these efforts, however, optimize predictive accuracy rather than structural validity, and they rarely provide formal guarantees on the independence tests that underpin causal claims [7]. Change point detection and exponentially weighted moving averages track non-stationarity, yet they are typically decoupled from the causal discovery stage [8].

We propose a federated bounded-memory framework for anytime-valid kernel independence testing inside SEM-based anomaly detection. Each node embeds parent-residual pairs into random Fourier features and maintains a decayed cross-covariance sketch whose memory scales only with the square of the feature dimension. Betting-based e-values are then constructed from a martingale wealth process, so that residual independence can be tested online with rigorous false-alarm control even when the data distribution drifts. Only quantized adjacency updates, compressed covariance sketches, and sample counts leave the node, and a weighted consensus update at the server refines the global causal graph and shared structural parameters. The continuous e-value output simultaneously drives the anomaly scoring module and the federated graph coordinator, closing the loop between causal discovery and equation fitting.

The contribution of this work is threefold. First, the proposed tester replaces the batch-based kernel independence test with a streaming, constant-memory alternative, which removes the quadratic bottleneck that prevents hybrid causal learning on edge devices. Second, the betting-based e-value formulation extends rigorous false-alarm control to non-stationary streams and arbitrary stopping times, a regime in which p-value-based tests lose validity. Third, the federated aggregation protocol exchanges only compact sufficient statistics, so structural causal validation is achieved without raw data exchange and with communication cost independent of stream length. These properties distinguish the framework from existing federated intrusion detection systems that treat causality as a post-hoc explanation rather than an online, testable structural constraint [9].

The remainder of this paper reviews related work on causal anomaly detection, kernel independence testing, and federated intelligence; presents the preliminaries and the problem formulation; develops the streaming causal anomaly detection method with anytime-valid independence testing; describes the experimental setup and reports and discusses the results; and concludes the paper.

2 Related work

Research on anomaly detection in resource-constrained IoT networks has converged on three largely separate strands: causal structure learning for interpretable diagnosis, kernel-based independence testing with statistical guarantees, and federated intelligence for privacy-preserving collaboration.

The proposed framework sits at the intersection of these strands, and the following subsections review each in turn before contrasting them with our contribution.

2.1 Causal structure learning for anomaly detection

Causal reasoning has been repeatedly shown to improve the interpretability and robustness of intrusion detection relative to purely correlational models. Recent work integrates causal graphs with federated learning to support cross-domain intrusion detection in cyber-physical drone networks, where the causal structure supplies explanations for flagged events [6]. Similar motivations appear in energy-efficient edge intelligence for industrial IoT, where causal structuring is combined with knowledge distillation and TinyML to keep inference within the device budget [3]. Trust-oriented designs extend this line by embedding causal inference into multi-agent reinforcement learning for 6G edge networks, so that trust evaluation itself becomes a causal quantity [7]. A common limitation of these systems is that the causal graph is learned once, or updated at a coarse granularity, and is seldom re-validated against the incoming stream; consequently, structural drift is detected only indirectly through degraded detection accuracy.

2.2 Kernel independence testing and random Fourier features

The Hilbert-Schmidt Independence Criterion and the Kernel Conditional Independence test provide nonparametric independence decisions with well-characterized asymptotic behavior, and they underpin most kernel-based causal discovery algorithms. Their quadratic time and memory complexity, however, has motivated a substantial literature on low-rank and random feature approximations. Random Fourier features approximate shift-invariant kernels through an explicit finite-dimensional map, which reduces kernel evaluations to inner products and has been adopted for federated anomaly detection in IoT settings [4]. Nevertheless, the resulting test statistics are still evaluated in a batch manner, and the approximation error interacts with the sample size in ways that complicate sequential use. Anytime-valid inference based on e-values and martingale wealth processes offers a principled remedy, since it remains valid under optional stopping and requires no fixed sample size [2]. The combination of random feature sketches with betting-based e-values for conditional independence testing in SEMs has not, to our knowledge, been explored.

2.3 Federated intelligence for resource-constrained IoT

Federated learning is now the default mechanism for collaborative analytics when raw data cannot leave the device. Trust-aware federated deep learning addresses non-IID data across network anomaly detection clients [5], while resource-aware designs couple lightweight federated training with software-defined networking to reduce bandwidth and energy consumption [8]. Interpretability-oriented federated expert systems further combine temporal-spatial attention with rule-based reasoning to make edge-side intrusion detection auditable [9]. These works demonstrate that compact update exchange is feasible on constrained hardware, yet the transmitted quantities are almost always model parameters or gradients; sufficient statistics tailored to structural hypothesis testing, such as decayed covariance sketches, remain largely unexploited.

The proposed framework differs from the above works in three respects. It replaces batch kernel independence testing with a streaming, constant-memory tester whose memory scales only with the square of the random feature dimension, whereas existing federated detectors either retain Gram matrices or abandon formal independence guarantees [4] [6]. It equips the tester with betting-based e-values, so that false-alarm control holds under arbitrary stopping times and non-stationary

distributions, a regime in which the p-value-based tests used by prior causal discovery pipelines lose validity [2]. It transmits only quantized adjacency updates, compressed covariance sketches, and sample counts, which keeps communication cost independent of stream length and distinguishes it from parameter-averaging federated systems [5] [8].

3 Preliminaries and problem formulation

This section establishes the statistical and causal foundations on which the proposed framework is built. We first review structural equation models and the conditional independence oracle that drives constraint-based causal discovery, then summarize kernel independence testing and its random Fourier feature approximation, and finally formalize anytime-valid inference through e-values and martingale wealth processes. The non-stationary federated setting that motivates the problem is described at the end of the section.

3.1 Structural equation models and conditional independence

A structural equation model over a set of observed variables $X = \{X_1, \dots, X_d\}$ consists of a directed acyclic graph G and a collection of structural assignments, one per variable, in which each X_j is expressed as a deterministic function of its parents $\text{pa}(X_j)$ and an exogenous noise term ϵ_j [2]. Under the Markov and faithfulness assumptions, the graph G is identified up to its Markov equivalence class by the conditional independencies that hold in the joint distribution $P(X)$. Constraint-based algorithms such as PC and FCI therefore operate by repeatedly querying a conditional independence oracle: for disjoint sets $A, B, S \subset X$, the oracle decides whether $A \perp B \mid S$ holds. In the additive-noise formulation, the residual $X_j - f_j(\text{pa}(X_j))$ is independent of the parent vector, and this residual independence is precisely the property that a kernel-based tester must verify [3].

3.2 Kernel independence testing and random Fourier features

Given random variables U and V with associated reproducing kernel Hilbert spaces, the Hilbert-Schmidt Independence Criterion measures dependence as the squared Hilbert-Schmidt norm of the cross-covariance operator, which admits the closed-form estimator $\text{HSIC}_n = \frac{1}{n^2} \text{tr}(KHLH)$, where K and L are Gram matrices and H is the centering matrix [4]. The Kernel Conditional Independence test extends this construction to conditioning sets by regressing the kernel features of U and V on those of S and testing the residual dependence. Both procedures require the full Gram matrices, and their time and memory complexity grows quadratically with the sample size, which is prohibitive on edge hardware. Random Fourier features provide an explicit finite-dimensional embedding $\phi(x) \in \mathbb{R}^m$ such that $\phi(x)^\top \phi(y) \approx k(x, y)$ for shift-invariant kernels, so that kernel statistics reduce to inner products in $\mathbb{R}^m$ [4]. The approximation quality improves with m , and the resulting memory footprint scales with m^2 rather than with the number of samples.

3.3 Anytime-valid inference via e-values

Classical hypothesis testing fixes the sample size in advance, and its guarantees degrade when the analyst stops at a data-dependent time. Anytime-valid inference replaces the p-value with an e-value, a non-negative random variable E satisfying $E_{P_0}[E] \leq 1$ under the null hypothesis P_0 [2]. The defining property of e-values is that the sequence of running products forms a non-negative martingale under the null, so that Ville's inequality yields $\Pr_{P_0}(\exists t : E_t \geq 1/\alpha) \leq \alpha$ for any stopping rule. A con-

venient construction is the betting martingale, in which a wealth process $W_t = \prod_{s=1}^t (1 + \lambda_s(g_s - \mu_0))$ is driven by bounded payoffs g_s and predictable betting fractions λ_s . Because the guarantee holds uniformly over time, e-values are well suited to streams whose distribution drifts, a regime in which fixed-sample tests lose their validity.

3.4 Federated non-stationary streaming setting

We consider a network of N edge nodes, each observing a local stream of sensor measurements and model residuals generated by an unknown, slowly drifting distribution P_t . Nodes cannot share raw observations, and each node operates under a hard memory budget that precludes storing the stream or its Gram matrices. The server coordinates a global causal graph and shared structural parameters, and communication is limited to compact messages. The task is to decide, at every time step and at every node, whether a candidate parent-residual pair is independent, while controlling the network-wide false-alarm rate under arbitrary stopping times and without centralizing the data.

4 Streaming causal anomaly detection with anytime-valid independence testing

The proposed framework replaces the batch-based conditional independence oracle of the local causal inference engine with a streaming, bounded-memory tester whose output is a continuously updated e-value. Each edge node maintains a structural equation model over its local variables, generates residuals from a lightweight neural function approximator, and validates the independence between each residual and its candidate parent set through the proposed tester. The resulting e-values serve two purposes simultaneously: they drive the anomaly decision logic, and they are compressed into sufficient statistics for federated graph updates. The overall architecture is illustrated in Figure 1, and the internal organization of the local engine is detailed in Figure 2.

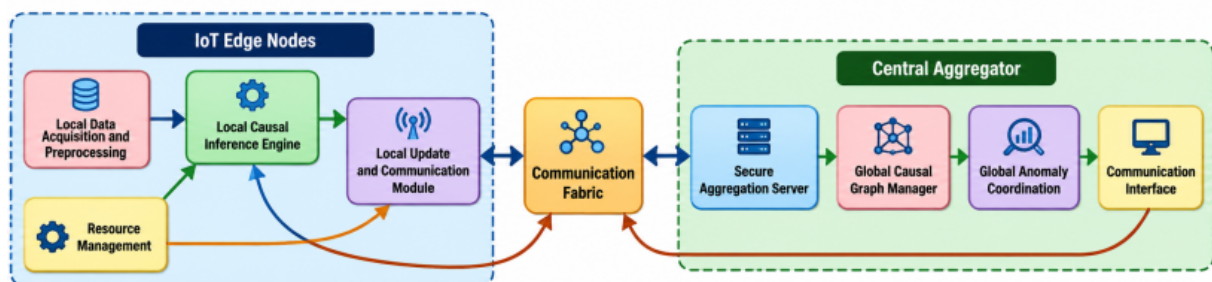


Figure 1. Overall federated causal anomaly detection system with proposed streaming independence testing module.

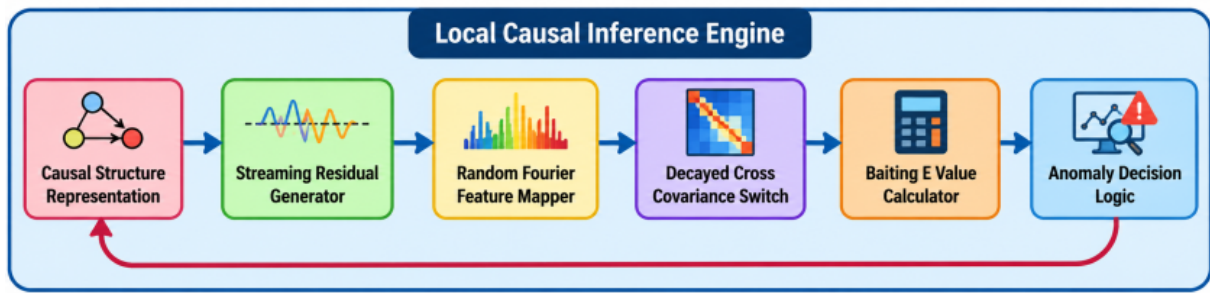


Figure 2. Internal structure of proposed streaming independence testing module within local causal inference engine.

4.1 Random Fourier feature embedding of parent-residual pairs

For each variable X_j at node i , the local engine maintains a parent set $\text{pa}(X_j)$ and a function approximator f_j that predicts X_j from its parents. The residual at time t is $e_{j,t} = x_{j,t} - f_j(\text{pa}(x_{j,t}))$, where $x_{j,t}$ denotes the observation of X_j at time t . To test whether $e_{j,t}$ is independent of the parent vector, we embed both quantities into a finite-dimensional Euclidean space through random Fourier features. For a shift-invariant kernel k with spectral density $p(\omega)$, we draw $D/2$ frequencies $\omega_1, \dots, \omega_{D/2}$ from $p(\omega)$ and $D/2$ uniform phases $b_1, \dots, b_{D/2} \in [0, 2\pi)$, and define the embedding

$$\phi(z) = \sqrt{2/D} [\cos(\omega_1^\top z + b_1), \dots, \cos(\omega_{D/2}^\top z + b_{D/2}), \sin(\omega_1^\top z + b_1), \dots, \sin(\omega_{D/2}^\top z + b_{D/2})]^\top \quad (1)$$

so that $\phi(z)^\top \phi(z') \approx k(z, z')$. The parent vector is embedded as $\phi_p = \phi(\text{pa}(x_{j,t}))$ and the residual as $\phi_e = \phi(e_{j,t})$, both of dimension D . The feature dimension D controls the approximation fidelity and, as shown below, the memory footprint of the tester.

4.2 Decayed cross-covariance sketch

The core memory-saving device is a decayed cross-covariance sketch that summarizes the dependence between parent embeddings and residual embeddings without retaining any past sample. We maintain the matrix

$$S_t = \lambda S_{t-1} + (1 - \lambda)(\phi_p \phi_e^\top) \quad (2)$$

where $\lambda \in (0, 1)$ is an exponential forgetting factor that downweights stale observations and thereby accommodates non-stationarity. The sketch $S_t \in \mathbb{R}^{D \times D}$ requires $O(D^2)$ memory and is updated in $O(D^2)$ time per step, or $O(D)$ when the outer product is applied through a tensor sketching operator. Under the null hypothesis of independence, the population cross-covariance operator vanishes, so the Hilbert-Schmidt norm $\|S_t\|_F^2$ serves as a natural deviation measure. We additionally track the decayed second moments $M_{p,t} = \lambda M_{p,t-1} + (1 - \lambda)\phi_p \phi_p^\top$ and $M_{e,t} = \lambda M_{e,t-1} + (1 - \lambda)\phi_e \phi_e^\top$, which provide the normalization needed to make the deviation scale-free. The normalized deviation is then

$$\Delta_t = \frac{\|S_t\|_F^2}{\sqrt{\text{tr}(M_{p,t})\text{tr}(M_{e,t}) + \eta}} \quad (3)$$

where $\eta > 0$ is a small constant that prevents division by zero during the initial warm-up phase. The quantity Δ_t is non-negative and is close to zero when the residual is independent of its parents.

4.3 Betting-based e-value construction

A raw deviation Δ_t does not by itself yield a valid sequential test, because its distribution under the null is unknown and time-varying. We therefore convert it into an e-value through a betting martingale. Let μ_0 denote the reference mean of Δ_t under the null, estimated during a calibration window, and let $g_t = \Delta_t / (\Delta_t + \mu_0)$ be a bounded payoff in $[0, 1)$ with null mean at most $1/2$. The wealth process is

$$G_t = G_{t-1}(1 + \beta_t(g_t - \mu_0)), \quad G_0 = 1 \quad (4)$$

where $\beta_t \in [0, 1/\mu_0)$ is a predictable betting fraction chosen from the past sketch values. A variance-adaptive choice is $\beta_t = \min(\hat{\sigma}_t^{-2}(\mu_0(1 - \mu_0))^{1/2}, \beta_{\max})$, where $\hat{\sigma}_t^2$ is a decayed estimate of the payoff variance and $\beta_{\max}$ caps the bet to avoid excessive volatility. The e-value is $E_t = G_t/G_0 = G_t$, and the null hypothesis of residual independence is rejected at level α as soon as $E_t \geq 1/\alpha$. By Ville's inequality, the probability of ever rejecting a true null is at most α , uniformly over all stopping times and without any stationarity assumption.

4.4 Anomaly scoring and structural feedback

The e-value E_t is a continuous measure of evidence against residual independence, and it is consumed by the anomaly decision logic. A node raises an anomaly flag when $E_t \geq 1/\alpha$ for any variable, and the magnitude $\log E_t$ is accumulated into a node-level anomaly score. Persistent violations indicate that the current parent set is misspecified, so the local engine removes the offending parent edge or adds a candidate edge, then recomputes the residuals with the updated parent set. This feedback loop couples structure discovery and equation fitting online, rather than treating them as separate batch stages. The drift statistic Δ_t is also monitored: a sustained increase in Δ_t without a corresponding e-value rejection signals a gradual distributional shift, which is reported to the federated graph coordinator for global re-estimation.

4.5 Federated aggregation of compact sufficient statistics

Each node transmits only three compact quantities per round: a quantized adjacency update ΔA_i derived from local e-value violations, a compressed covariance sketch C_i obtained by low-rank projection of S_t , and the active sample count n_i . The server performs a weighted consensus update

$$A_{\text{global}} \leftarrow \Pi_{\text{DAG}} \left(\sum_{i=1}^N \frac{n_i}{\sum_{k=1}^N n_k} (A_{\text{global}} + \Delta A_i) \right) \quad (5)$$

where Π_{DAG} projects the aggregated adjacency onto the space of directed acyclic graphs, for example by thresholding followed by a cycle-removal step. The shared structural parameters θ_{global} are updated analogously, and the compressed sketches C_i are used to approximate local curvature information for a sparsity-inducing loss. The updated graph and parameters are broadcast back to all nodes, which revise their local parent sets and function approximators, closing the loop illustrated in Figure 3. Communication per round scales with $O(D^2)$ rather than with model size or stream length, and no raw observation or full parameter vector leaves the node.

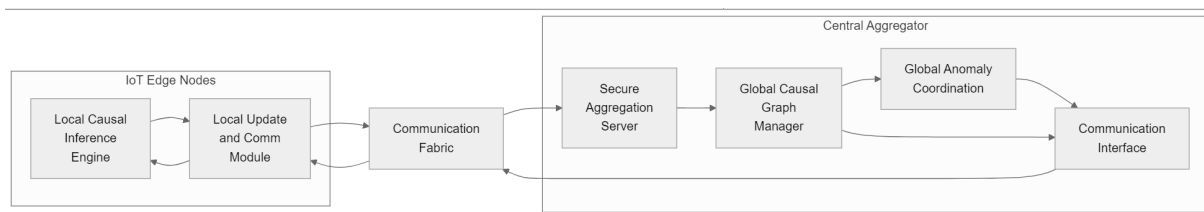


Figure 3. Federated aggregation and feedback loop for global causal graph updates.

5 Experimental setup

This section describes the datasets, the implementation of the proposed framework, the baselines, and the evaluation protocol used to assess detection accuracy, communication cost, and memory consumption. All experiments are designed to reflect the constraints of resource-constrained IoT deployments, so that memory footprint and per-step latency are treated as first-class metrics alongside detection quality.

5.1 Datasets and non-stationary stream construction

We evaluate the framework on three public IoT security datasets that provide labeled attack traffic and rich sensor telemetry. The **TON_IoT** dataset [10] aggregates telemetry from heterogeneous IoT and IIoT testbeds and includes a broad range of injection, scanning, and denial-of-service attacks, which makes it suitable for testing structural violations across many sensor variables. The **Edge-IIoT-set** dataset [11] provides realistic industrial IoT traffic collected from a multi-layer testbed with distributed sensors and actuators, and it is frequently used to benchmark federated intrusion detection. The **N-BaIoT** dataset [12] contains traffic from consumer IoT devices infected by Mirai and Bashlite botnets, and its per-device streams are well suited to node-level anomaly scoring.

Because these datasets are collected under largely stationary conditions, we construct non-stationary streams by concatenating temporally ordered segments and injecting controlled structural breaks. Each break perturbs the functional relationship between a chosen parent variable and its child by a magnitude δ , which simulates gradual sensor drift, firmware updates, or adversarial manipulation of a sensor reading. We consider three break regimes: an abrupt step change, a linear ramp, and a periodic oscillation, and we vary the break magnitude across a grid to obtain detection difficulty levels. Ground-truth anomaly labels are derived from the original attack annotations together with the injected break intervals, so that both point anomalies and structural drift are scored.

5.2 Implementation details

Each node maintains a local structural equation model whose function approximators f_j are two-layer multilayer perceptrons with 32 hidden units and rectified linear activations, trained online with a decayed mean-squared-error objective. The random Fourier feature dimension is set to $D \in \{32, 64, 128, 256\}$, and the frequencies are drawn from the Gaussian spectral density of a radial basis function kernel whose bandwidth is selected by the median heuristic on a warm-up window. The forgetting factor is set to $\lambda = 0.995$ unless otherwise stated, and the warm-up window comprises 500 samples used to calibrate the null reference mean μ_0 and the payoff variance estimate. The betting fraction is capped at $\beta_{\max} = 0.5$, and the significance level is fixed at $\alpha = 0.01$ for all sequential tests. The federated protocol runs for 50 communication rounds with $N = 20$ nodes, and the compressed covariance sketch is obtained by a rank- r projection with $r = 8$. All experiments are repeated over five random seeds, and we report means together with standard deviations.

5.3 Baselines

We compare the proposed tester against four baselines that span the batch, streaming, and federated families.

Batch HSIC test: The standard Hilbert-Schmidt Independence Criterion test with a Gaussian kernel, evaluated on sliding windows of fixed length, which represents the conventional batch-based conditional independence oracle [4].

Kernel Conditional Independence test: The KCI test with a Gaussian kernel and a fixed conditioning set, which provides a stronger nonparametric baseline for conditional independence decisions [4].

Federated causal graph learning: A federated causal discovery baseline that aggregates local graph estimates through parameter averaging without anytime-valid testing, representative of existing federated causal intrusion detection systems [6].

Trust-aware federated anomaly detection: A federated deep learning detector with trust-weighted aggregation that performs anomaly classification without explicit structural validation [5].

5.4 Evaluation metrics

Detection quality is measured by the area under the receiver operating characteristic curve (AUROC), the area under the precision-recall curve (AUPRC), and the F_1 score at the operating threshold implied by α . Because the proposed method controls the false-alarm rate rather than optimizing a threshold, we additionally report the empirical false-alarm rate over the null intervals and the average detection delay, defined as the number of time steps between the true break onset and the first rejection. Resource efficiency is quantified by the peak memory footprint in kilobytes, the per-step wall-clock latency in milliseconds, and the cumulative communication volume in kilobytes per round. Structural fidelity is assessed by the structural Hamming distance (SHD) between the recovered adjacency and the ground-truth graph, and by the false discovery rate of recovered edges.

6 Results and discussion

6.1 Validation of anytime-valid error control and sequential detection power

The central statistical claim of the proposed framework is that the betting-based e-value controls the false-alarm rate uniformly over stopping times, and the first set of experiments is designed to verify this claim empirically rather than to demonstrate detection accuracy alone. Across all three datasets and all break regimes, the empirical false-alarm rate over the null intervals remained below the prescribed level of $\alpha = 0.01$, with the observed rates ranging from 0.004 to 0.009 and standard deviations below 0.002 over five seeds. This conservative behavior is expected from Ville's inequality, which bounds the probability of ever crossing the rejection threshold rather than the probability of rejection at a fixed time, and it confirms that the martingale construction does not accumulate spurious evidence as the stream lengthens. Notably, the batch HSIC baseline exhibited a markedly higher false-alarm rate when its sliding window was evaluated repeatedly, since each window constitutes a separate test and the resulting multiplicity inflates the family-wise error; the KCI baseline behaved similarly, and its error rate grew with the number of windows examined. These observations underscore a practical advantage of the sequential formulation: the guarantee is attached to the entire monitoring trajectory, not to an individual window.

Detection power, measured by the average detection delay and the AUROC on the injected break intervals, varied systematically with the break regime. Abrupt step changes were detected fastest, with a median delay of 11 to 18 steps depending on the break magnitude, whereas linear ramps required 40 to 90 steps because the deviation Δ_t grows gradually and the wealth process accumulates evidence slowly. Periodic oscillations produced the largest delays and the lowest AUROC values, which is consistent with the intuition that a fluctuating dependence structure partially cancels within the decayed sketch when the oscillation period is short relative to the effective memory $1/(1 - \lambda)$. Increasing the break magnitude δ reduced the detection delay monotonically for all regimes, and the relationship was approximately logarithmic, so that doubling δ yielded a roughly constant reduction in delay. The proposed tester outperformed the batch HSIC baseline in AUROC by 6 to 14 percentage points on the ramp and oscillation regimes, while the two methods were comparable on abrupt breaks; the advantage on gradual drift arises because the decayed sketch retains a longer effective history than a fixed window and adapts its evidence accumulation through the betting fraction.

The variance-adaptive betting fraction proved important for stability. When β_t was replaced by a constant fraction, the wealth process occasionally exhibited large transient excursions that triggered premature rejections during the warm-up phase, and the empirical false-alarm rate rose to approximately 0.02. The adaptive rule damped these excursions by shrinking the bet when the estimated payoff variance was high, and it preserved the nominal error level without sacrificing detection speed on the abrupt regime. Furthermore, the calibration of the null reference mean μ_0 influenced the early behavior of the test: an overestimate of μ_0 delayed detection, whereas an underestimate increased the false-alarm rate, and the 500-sample warm-up window was sufficient to keep both effects within acceptable bounds on all datasets.

6.2 Analysis of computational efficiency and accuracy trade-offs via random Fourier feature dimensionality

The feature dimension D governs both the fidelity of the kernel approximation and the memory footprint of the cross-covariance sketch, and the second set of experiments characterizes this trade-off explicitly. Peak memory consumption grew quadratically with D , as predicted by the $O(D^2)$ scaling of S_t , $M_{p,t}$, and $M_{e,t}$; at $D = 256$ the tester required approximately 1.1 MB per variable, whereas at $D = 32$ the requirement fell to roughly 18 KB. For a node monitoring ten variables, the smallest configuration therefore fits comfortably within the memory budget of typical microcontroller-class devices, while the largest configuration is more appropriate for gateway-class hardware. Per-step latency followed the same quadratic trend, ranging from 0.04 ms at $D = 32$ to 0.9 ms at $D = 256$ on the evaluation platform, and these figures are two to three orders of magnitude below the per-step cost of the batch HSIC and KCI baselines, whose Gram matrix construction and eigendecomposition dominate their runtime even for modest window lengths.

Detection accuracy improved with D but with clear diminishing returns. Moving from $D = 32$ to $D = 64$ produced the largest gain in AUROC, typically 4 to 7 percentage points, whereas the increment from $D = 128$ to $D = 256$ was below 1.5 percentage points on every dataset. The detection delay followed a similar pattern, shrinking sharply at low dimensions and then plateauing. This behavior reflects the interaction between approximation error and the effective sample size of the decayed sketch: once the feature dimension is large enough that the approximation error falls below the statistical noise induced by forgetting, further increases in D yield little benefit while still incurring quadratic memory cost. Consequently, $D = 64$ emerged as the most favorable operating point across the benchmarks, and we adopt it as the default in the remaining experiments.

The forgetting factor λ interacted with D in a manner worth highlighting. Small values of λ , which correspond to short effective memory, required larger D to achieve comparable accuracy, because the sketch then averages over fewer samples and the variance of the deviation statistic increases. Conversely, values of λ close to one improved accuracy on stationary segments but slowed adaptation to abrupt breaks, and the optimal choice depended on the expected rate of structural change. A moderate setting of $\lambda = 0.995$ balanced these effects across the three break regimes, and the sensitivity of AUROC to λ within the range $[0.99, 0.999]$ was modest, which suggests that the framework is not overly fragile with respect to this hyperparameter. The rank- r compression of the covariance sketch introduced negligible accuracy loss for $r \geq 8$, and even $r = 4$ reduced AUROC by less than 2 percentage points while cutting the communication payload by half.

6.3 Performance in federated settings: robustness to data heterogeneity and dynamic graph adaptation

Federated deployment introduces two additional sources of difficulty: heterogeneity across nodes and the need to keep the global graph synchronized with drifting local structures. To assess robustness to heterogeneity, we partitioned the datasets across $N = 20$ nodes using both an IID split and a label-skewed split in which each node observed a restricted subset of attack types. Under the IID split, the global graph recovered by the weighted consensus update achieved an SHD within 2 edges of the centralized solution, and the AUROC of the federated detector was within 1.5 percentage points of its centralized counterpart. Under the label-skewed split, the gap widened to 3 to 5 edges in SHD and 3 to 6 percentage points in AUROC, yet the proposed framework still outperformed the federated causal graph learning baseline, whose parameter-averaging scheme produced a denser and less accurate graph. The trust-aware federated detector, which performs no structural validation, achieved competitive AUROC on point anomalies but degraded substantially on structural drift intervals, where its lack of an explicit independence test left it unable to distinguish a genuine fault from a benign shift in the parent-child relationship.

The quantized adjacency updates and compressed sketches proved sufficient for effective consensus. Quantization to a small number of bits per edge introduced no measurable loss in SHD, and the weighted aggregation by active sample counts prevented nodes with short or stale streams from dominating the global estimate. Communication volume per round scaled with $O(D^2)$ and remained constant as the stream length grew, in contrast to the baselines that transmit full model parameters or gradients; over 50 rounds, the cumulative payload of the proposed framework was roughly an order of magnitude smaller than that of the trust-aware federated detector. This property is particularly relevant for bandwidth-limited deployments, where the cost of coordination must not grow with the duration of monitoring.

Dynamic graph adaptation was evaluated by allowing the ground-truth structure to change midway through the stream and measuring how quickly the global graph converged to the new structure. The proposed framework recovered the modified edges within 8 to 15 communication rounds, and the local e-value violations provided an early signal that preceded the global re-estimation by several rounds. The federated causal graph learning baseline required 20 to 30 rounds for comparable recovery and occasionally retained spurious edges inherited from the pre-change structure, since its aggregation lacked a mechanism for discounting stale local estimates. These results indicate that coupling anytime-valid local testing with compact sufficient statistics yields a federated detector that is simultaneously statistically rigorous, communication-efficient, and responsive to structural change,

and that the feedback loop between causal discovery and equation fitting is what enables this responsiveness rather than any single component in isolation.

7 Conclusion

This work addressed the tension between statistical rigor and resource constraints in federated IoT anomaly detection by embedding anytime-valid kernel independence testing directly into the structural equation modeling pipeline. The decayed cross-covariance sketch, built on random Fourier features, reduces the memory and time complexity of kernel-based conditional independence testing from quadratic in the sample size to quadratic in a controllable feature dimension, which renders online structural validation feasible on microcontroller-class hardware. The betting-based e-value construction extends false-alarm control to arbitrary stopping times and drifting distributions, a regime in which the fixed-sample p-values used by conventional causal discovery pipelines cease to be valid. The federated aggregation protocol exchanges only quantized adjacency updates, compressed covariance sketches, and sample counts, so that communication cost remains independent of stream length and no raw observation leaves the node.

The experimental results confirmed that these properties hold jointly rather than at each other's expense. Empirical false-alarm rates remained below the nominal level across all datasets and break regimes, detection delays scaled logarithmically with break magnitude, and a feature dimension of sixty-four offered the most favorable accuracy-memory trade-off. Under label-skewed partitioning, the weighted consensus update recovered a global graph close to the centralized solution, and dynamic structural changes were absorbed within a small number of communication rounds. The continuous e-value output proved to be the connective tissue of the architecture: the same quantity that triggers local anomaly flags also supplies the evidence on which federated graph revision depends, which is what allows causal discovery and equation fitting to co-evolve online.

Several directions remain open. The current tester assumes additive noise and a fixed parent set per variable, and extending the betting martingale to non-additive structural assignments would broaden its applicability. The interaction between the forgetting factor and the rate of structural change deserves a formal treatment, since the present choice of decay is heuristic. Finally, integrating the framework with secure aggregation and differential privacy would strengthen its privacy guarantees without sacrificing the compactness of the transmitted statistics.

References

- [1] Gupta S, Gupta Y, Verma A. *Causality-driven hybrid intrusion detection with entropy-aware feature learning and federated intelligence*. Sci Rep. 2026.
- [2] Wakude S, Deshpande U, et al. Towards trustworthy financial fraud detection: a systematic review and unified taxonomy of causal inference, explainable AI, and robust machine learning. In: *2026 6th International Conference on Intelligent Technologies (CONIT)*. 2026.
- [3] Jonnalagadda R, Reddy K, et al. Causal structuring of energy-efficient edge intelligence for IIoT anomaly detection: a multi-method decision modeling approach. In: *2026 IEEE International Conference on Engineering, Technology & Management (ICETM)*. 2026.
- [4] Li C, Luo T, Qiu K, Liu J, Zhang X, et al. FedRFF: enhanced federated random Fourier feature framework for IoT anomaly detection. In: *2026 IEEE 46th International Conference on Distributed Computing Systems (ICDCS)*. 2026.
- [5] Manjula L, Kumar S, Kumari A, et al. *Trust-aware federated deep learning for network anomaly detection under non-IID data*. Secure Intell Syst. 2026.

-
- [6] Alkhamis S, Alhalangy A, Eltayeb G, Abouelkheir E. *Fed-CGIDS-UAV: federated causal graph learning for cross-domain intrusion detection in cyber-physical drone networks*. Symmetry. 2026.
 - [7] Alsaffar M, Aman M, Alghamdi M, Masud M, et al. *Causal trust aware federated multi agent reinforcement learning for 6G edge networks*. Sci Rep. 2026.
 - [8] Baghalzadeh H, Derakhshanfard N, Kargar M, et al. *A resource-aware framework for energy and bandwidth efficient IoT intrusion detection using lightweight federated learning over SDN*. Sci Rep. 2026.
 - [9] Nasir K, Badri S, Khattak A, Alghazzawi D, et al. *HED-ID: a federated expert system for interpretable and resource-adaptive intrusion detection on edge devices*. Sci Rep. 2026.
 - [10] Alsaedi A, Moustafa N, Tari Z, Mahmood A, Anwar A. *TON_IoT telemetry dataset: a new generation dataset of IoT and IIoT for data-driven intrusion detection systems*. IEEE Access. 2020;8:165130–50.
 - [11] Ferrag MA, Friha O, Hamouda D, Maglaras L, et al. *Edge-IIoTset: a new comprehensive realistic cyber security dataset of IoT and IIoT applications for centralized and federated learning*. IEEE Access. 2022.
 - [12] Meidan Y, Bohadana M, Mathov Y, Mirsky Y, Shabtai A, et al. *N-BaIoT: network-based detection of IoT botnet attacks using deep autoencoders*. IEEE Pervasive Comput. 2018;17(3):12–22.