Temporal Anomaly Detection via Rolling Window Causal Graph Evolution

Nicolas Bigeard

October 8, 2025

Abstract

Contemporary methods for time series anomaly detection, predominantly based on deep learning, have demonstrated significant capabilities in modeling complex temporal patterns. However, their reliance on learning statistical correlations renders them inherently fragile, susceptible to performance degradation under distribution shifts, and opaque in their decision-making processes, thus limiting their utility for root cause analysis. This paper introduces a novel paradigm that reframes anomaly detection as the problem of identifying significant structural changes in the underlying causal data-generating process. The proposed method, Rolling Window Causal Graph Evolution (RW-CGE), operationalizes this concept by learning dynamic causal graphs within a sliding window over multivariate time series and quantifying their structural evolution. A high degree of evolution, measured by the dissimilarity between consecutive graphs, serves as a robust and directly interpretable anomaly score. Extensive empirical validation on complex, real-world benchmarks, including the NASA Mars Science Laboratory (MSL) and Soil Moisture Active Passive (SMAP) spacecraft telemetry datasets, demonstrates that RW-CGE not only achieves state-of-the-art detection performance but also provides unparalleled diagnostic insights. The specific changes in the graph's edge structure offer explicit hypotheses about the components involved in an anomalous event, facilitating rapid root cause analysis. This work represents a critical step toward more robust, reliable, and explainable artificial intelligence systems for monitoring critical infrastructure.

1 Introduction

1.1 Motivation: The Fragility of Correlation-Based Anomaly Detection

The automated monitoring of complex systems through time series analysis is a cornerstone of modern industry and science. In high-stakes domains such as industrial manufacturing, aerospace systems monitoring, and financial fraud detection, the timely and accurate identification of anomalies is not merely a matter of operational efficiency but of critical importance for safety and economic stability [3, 10, 18, 32]. An undetected fault in a manufacturing process can lead to cascading failures, while a missed anomaly in spacecraft telemetry could jeopardize a multi-billion dollar mission [27, 43].

In response to this need, the field has seen a proliferation of sophisticated anomaly detection techniques, with deep learning models such as Long Short-Term Memory (LSTM) networks and Transformers becoming the de facto standard [11, 56]. These models excel at learning intricate, non-linear patterns from vast amounts of historical data. However, their power is also their fundamental weakness: they are masters of correlation, not causation [49, 58]. By learning a complex statistical representation of "normal" behavior, they identify anomalies as data points that have a low probability under this learned model. This approach is inherently brittle. When a system undergoes a benign distribution shift—a change in operational environment or a new, non-anomalous mode of behavior—the statistical properties of the data change. Correlation-based models, having overfit to the training distribution, often misinterpret these novel-but-normal patterns as anomalous, leading to a deluge of false alarms [7, 59].

Furthermore, this paradigm suffers from a critical lack of interpretability. When a deep learning model flags an anomaly, it typically provides only a scalar score indicating the degree of deviation.

It cannot answer the crucial follow-up question: "Why is this anomalous?" This "black box" nature creates a significant barrier to action. For an engineer or system operator, knowing that an anomaly has occurred is insufficient; they must understand its origin to perform effective root cause analysis (RCA) and implement corrective measures [8, 43, 58].

1.2 Core Thesis: Anomalies as Causal Mechanism Shifts

This paper posits a fundamental reconceptualization of temporal anomalies. An anomaly is not merely a statistical outlier; it is the observable manifestation of a significant change in the system's underlying data-generating process. Drawing from the language of causal inference, a system's behavior can be described by a Structural Causal Model (SCM), a set of equations defining how each variable is causally determined by its parents [14, 15]. In this framework, a true anomaly—such as a component failure or a cyber-attack—can be modeled as an intervention on the SCM. This intervention alters one or more of the system's causal mechanisms, the functions that govern the cause-and-effect relationships between variables.

This perspective is grounded in the **Principle of Independent Mechanisms**, a cornerstone of causal science, which states that the causal generative process of a system is composed of autonomous modules that do not influence each other [14, 15]. This principle implies that a single-point failure should ideally perturb only a single, local causal mechanism. The consequence of such a perturbation is a localized change in the system's causal graph—the directed acyclic graph (DAG) that represents the SCM. Therefore, detecting an anomaly becomes equivalent to detecting a structural change in this causal graph.

This causal reframing offers a more robust definition of an anomaly. Traditional methods define an anomaly as a low-probability event under a single, learned statistical model of normal behavior. In contrast, the causal perspective defines an anomaly as a high-probability event under a *different*, post-change causal model. This distinction is crucial. It allows a system to differentiate between benign distribution shifts that preserve the underlying causal structure (e.g., changes in the variance of sensor noise) and true anomalies that break or rewire the causal mechanisms. The latter represents a fundamental change in how the system operates, which is precisely what anomaly detection should identify.

1.3 Proposed Approach: Rolling Window Causal Graph Evolution (RW-CGE)

To operationalize this thesis, this work introduces a novel framework named Rolling Window Causal Graph Evolution (RW-CGE). The method translates the abstract concept of detecting causal shifts into a concrete, data-driven algorithm. The process is as follows:

1. A rolling window of a fixed size is moved across the multivariate time series, creating a sequence of overlapping data segments.

2. Within each window, a state-of-the-art dynamic causal discovery algorithm is applied to learn a directed graph. This graph represents a snapshot of the system's effective causal structure during that time interval.

3. The structural dissimilarity between the graph learned in the current window and the graph from the immediately preceding window is quantified using a graph-theoretic distance metric.

4. This dissimilarity score, which captures the degree of "causal evolution," is used directly as the anomaly score for the current time step. A stable system will exhibit a consistent causal structure, resulting in low dissimilarity scores. A sudden event, such as a fault, will induce a significant change in the causal graph, leading to a sharp spike in the dissimilarity score.

This approach is intrinsically linked to the problem of causal discovery under distribution shifts. It has been shown that distribution shifts, far from being a mere nuisance, can act as natural experiments that help identify the correct causal structure by breaking statistical symmetries that would

otherwise make different causal models indistinguishable [6, 37, 38]. The rolling window of RW-CGE effectively treats the time series as a sequence of different "environments." During stable periods, minor fluctuations help to robustly identify the nominal causal graph. An anomalous event acts as a strong intervention, creating a distinct new environment and a corresponding shift in the learned graph. Thus, the very phenomenon we aim to detect—the anomaly—provides the informational leverage needed to identify the change in the causal graph, creating a symbiotic relationship between the tasks of anomaly detection and causal discovery.

1.4 Contributions

The primary contributions of this work are threefold:

1. **A Novel Framework:** A new, causality-first framework for time series anomaly detection is proposed. It moves beyond monitoring value-based deviations to directly detect structural breaks in the data-generating process, offering a more principled foundation for detecting system failures.

2. **Inherent Interpretability:** The proposed method provides not just anomaly detection but also explicit diagnostic information. The specific edges that appear, disappear, or change direction between consecutive causal graphs directly pinpoint the variables and relationships involved in the anomaly, offering a powerful tool for rapid and targeted root cause analysis [34, 43].

3. **State-of-the-Art Performance:** Through extensive empirical evaluations on challenging, real-world benchmark datasets, including NASA's SMAP and MSL spacecraft telemetry data [25, 28], it is demonstrated that RW-CGE outperforms a suite of prominent deep learning and classical baselines in terms of precision, recall, and F1-score.

2 Related Work

This section provides a critical review of the existing literature, situating the proposed RW-CGE framework within the broader context of time series analysis and highlighting its unique contributions by systematically analyzing the failure modes of current paradigms.

2.1 Anomaly Detection in Multivariate Time Series

The field of time series anomaly detection is vast, with methods spanning classical statistics to modern deep learning. These approaches can be broadly categorized by the principles they employ to distinguish normal from abnormal behavior.

2.1.1 Classical and Statistical Methods

Traditional approaches often rely on statistical properties or distance-based measures. The **Isolation Forest** algorithm, for instance, operates on the principle that anomalies are "few and different" [1, 23]. It builds an ensemble of random trees, and anomalies are identified as those data points that require fewer splits to be isolated from the rest of the data [51]. While efficient, its axis-parallel splits can struggle with complex, non-linear relationships in high-dimensional data [53].

Another prominent technique is the **Matrix Profile**, which excels at identifying anomalous subsequences, or "discords" [2, 29]. It computes an all-pairs similarity search between subsequences of a time series, defining a discord as the subsequence with the largest distance to its nearest neighbor [16, 29, 54]. This is a powerful, parameter-light method for detecting unusual shapes or patterns. However, its effectiveness can be compromised if an anomalous pattern repeats (the "twin freak" problem), as the repeated instances will be each other's nearest neighbors and thus not be flagged as discords [55].

2.1.2 Deep Learning Approaches

The dominant paradigm in recent years has been deep learning, which offers unparalleled flexibility in modeling complex data. A common taxonomy divides these methods into forecasting-based, reconstruction-based, and hybrid models [10, 12, 13].

Reconstruction-Based Models: This category is exemplified by autoencoder architectures, particularly those using LSTMs (**LSTM-AE**). These models are trained exclusively on normal data with the objective of accurately reconstructing their input [22, 40]. The core assumption is that the model will learn a compressed representation of normal patterns. When presented with an anomalous input that deviates from these learned patterns, the decoder will fail to reconstruct it accurately, resulting in a high reconstruction error that serves as the anomaly score [9, 24]. While effective, these models are fundamentally correlational and can be brittle; a benign change in the data distribution can lead to high reconstruction errors for normal data, causing false positives [7].

Forecasting-Based Models: These methods train a model to predict future values of the time series based on past observations. An anomaly is detected when there is a significant discrepancy between the predicted value and the actual observed value [9, 22, 31]. Like reconstruction models, they learn a model of normal temporal dynamics and flag deviations from it.

Transformer-Based Models: More recent approaches leverage the Transformer architecture, such as in the **TranAD** model [43]. The self-attention mechanism in Transformers allows them to more effectively capture long-range dependencies and complex inter-variable relationships compared to recurrent models [4, 36]. Despite this architectural advancement, they still operate under the same fundamental paradigm: learning a model of normal correlations and identifying anomalies as statistical deviations.

The systematic analysis of these methods reveals a taxonomy of failure modes. Value-based statistical methods fail on contextual anomalies [21, 26]. Subsequence-based methods can fail on repeated anomalies [55]. Correlation-based deep learning models fail under benign distribution shifts and offer no interpretability for root cause analysis [7]. Finally, as will be discussed, existing causality-informed methods fail when the anomaly itself is a change in the causal structure. This landscape of limitations creates a clear intellectual and practical need for a method that is robust to these issues—one that models dynamic inter-variable relationships (causality) and is sensitive to changes in those relationships themselves (graph evolution).

2.2 Causal Discovery from Time Series Data

To detect changes in causal structure, one must first be able to discover it. This has been a long-standing goal in time series analysis.

2.2.1 Granger Causality and its Limitations

A foundational concept is **Granger causality**, which defines causality in terms of predictability: time series X Granger-causes time series Y if past values of X contain unique information that helps predict future values of Y [52, 57]. While historically influential, this definition is widely recognized as insufficient for inferring true causal relationships. It is a measure of directed functional connectivity, not mechanistic causation [46]. Its reliance on linear vector autoregressive (VAR) models makes it prone to producing misleading or spurious results in the presence of non-stationarity, non-linear dynamics, and unobserved confounding variables—all of which are common in real-world systems [48]. This critique motivates the need for more principled, SCM-based approaches.

2.2.2 Modern Causal Discovery Methods

Modern causal discovery algorithms aim to recover the underlying causal graph from observational data and can be grouped into several families [33, 35, 44].

Score-Based Continuous Optimization: A particularly relevant family for this work are score-based methods that frame causal discovery as a continuous optimization problem. The seminal **NOTEARS** algorithm reformulated the combinatorial problem of finding the best-fitting DAG as a continuous, constrained optimization problem that can be solved with standard gradient-based methods [47]. This approach was extended to time series data with **DYNOTEARS** [41, 42]. DYNOTEARS learns a Dynamic Bayesian Network (DBN) by explicitly modeling both contemporaneous (intra-slice) and time-lagged (inter-slice) causal relationships within a linear structural equation model framework [47]. Its formulation is both scalable to higher dimensions and computationally tractable, making it a suitable engine for the repeated graph inference required by our rolling-window approach [41, 42].

Dynamic and Time-Varying Models: Other advanced methods, such as **DyCAST**, use techniques like Neural Ordinary Differential Equations (Neural ODEs) to model causal structures that evolve *smoothly* and continuously over time [5]. This is a powerful approach for tracking gradual system drift or cyclical patterns, such as those that follow a time of day [5]. However, many of the most critical anomalies in engineered systems are not smooth drifts but *abrupt structural breaks*—a component fails, a connection is severed, or a new, dangerous feedback loop emerges [17]. The RW-CGE framework, by not assuming smoothness between consecutive windows, is philosophically and practically tailored to detect these discontinuous structural breaks. It is therefore not a replacement for smooth models but a complementary approach designed for a different, and highly critical, class of system changes.

2.3 Causality-Informed Anomaly Detection and Change Point Analysis

The idea of leveraging causality for anomaly detection is nascent but growing.

2.3.1 Existing Causal Anomaly Detection

A notable work in this area is **Causanom**, which detects anomalies by identifying violations of a given causal graph [14, 15]. It models the conditional distribution of each variable given its causal parents and flags a deviation if an observation is unlikely under this model. Its key contribution is moving beyond raw data correlations to model causal dependencies. However, its critical limitation is the assumption of a **fixed, static causal graph**. This graph must either be provided by a domain expert or learned once from a training dataset. This makes the method non-robust to the very class of anomalies we are most interested in: those that manifest as a change or "rewiring" of the causal graph itself. Causanom can detect when a system behaves unexpectedly *within* its known structure, but not when the structure itself breaks.

2.3.2 Causal Change Point Detection

The work presented here is closely related to the field of causal change point detection, which formalizes the notion of a "causal mechanism shift" [19, 20]. This literature distinguishes between two types of shifts [20]:

- **Soft Mechanism Shift:** The functional form of a causal relationship changes (e.g., the gain of a sensor drifts), but the set of causal parents remains the same. The graph topology is invariant.
- **Hard Mechanism Shift:** The set of causal parents for a variable changes. This corresponds to an edge being added, removed, or reversed in the causal graph, representing a fundamental structural change in the system.

The RW-CGE framework is explicitly designed to detect these "hard" mechanism shifts, which often correspond to the most significant and actionable types of system anomalies.

The following table summarizes the key distinctions between different causality-informed paradigms, highlighting the unique position of the proposed RW-CGE method.

Table 1: Comparison of Causal Anomaly Detection Paradigms

Feature	Granger Causality	Causanom	RW-CGE (Proposed)
Causal Assumption	Predictability implies causality	SCM	SCM
Graph Type	Implicit, pairwise	Static, pre-learned	Dynamic, learned in windows
Handles Confounders	No, highly sensitive	No, assumes no latent	Depends on discovery alg.
Anomaly Definition	Large prediction error	Violation of fixed mechanisms	Significant graph change
Interpretability	Low, potentially misleading	Node-level attribution	Edge-level diagnosis
Key Limitation	Linearity, stationarity	Static graph assumption	Computational complexity

3 Preliminaries

This section establishes the formal notation and foundational concepts required to precisely define the proposed methodology.

3.1 Problem Definition

Let $\mathbf{X}$ represent a multivariate time series, defined as an ordered sequence of T observations, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$. Each observation $\mathbf{x}_t \in \mathbb{R}^d$ is a d-dimensional vector, where each dimension corresponds to a measured variable or sensor in the system at time t.

The objective is unsupervised time series anomaly detection. Given the observational time series $\mathbf{X}$ as input, the task is to produce a corresponding anomaly score sequence $\mathbf{A} = \{a_1, a_2, \ldots, a_T\}$, where each $a_t \in \mathbb{R}^+$ is a non-negative scalar. A higher value of a_t indicates a greater degree of anomalousness at time t. The detection of an anomaly at time t is then determined by comparing a_t to a threshold, which may be static or dynamically adjusted.

3.2 Structural Causal Models for Time Series

The data-generating process of the system is assumed to be governed by an underlying Structural Causal Model (SCM) [39]. For time series data, this is most naturally represented by a **Dynamic Bayesian Network (DBN)**. A DBN describes the probabilistic relationships between variables both within a single time slice (contemporaneous effects) and across different time slices (time-lagged effects).

The graphical structure of a DBN over a finite time horizon can be represented by a "full time causal graph" [44]. In practice, it is often sufficient and computationally necessary to make a Markov assumption, where the state of the system at time t depends only on a finite history of P previous states. The causal relationships within this finite history are captured by a "window causal graph" [44]. The graph that is learned within each window of the RW-CGE method is an empirical estimation of such a window causal graph.

Consistent with the DYNOTEARS framework, the functional relationships in the SCM are modeled using a linear Structural Vector Autoregressive (SVAR) model. For a given observation $\mathbf{x}_t$, this model is expressed as:

$$\mathbf{x}_t^T = \mathbf{x}_t^T \mathbf{W} + \sum_{p=1}^{P} \mathbf{x}_{t-p}^T \mathbf{A}_p + \mathbf{z}_t^T \tag{1}$$

where:

- $\mathbf{W} \in \mathbb{R}^{d \times d}$ is the weighted adjacency matrix representing the contemporaneous (intra-slice) causal relationships among the d variables at time t. For the graph to be a valid DAG, $\mathbf{W}$ must be acyclic.
- $\mathbf{A}_p \in \mathbb{R}^{d \times d}$ for $p \in \{1, \ldots, P\}$ are the weighted adjacency matrices for the time-lagged (inter-slice) causal relationships.
- P is the maximum time lag considered.
- $\mathbf{z}_t \in \mathbb{R}^d$ is a vector of exogenous noise terms, assumed to be i.i.d.

The primary goal of causal discovery is to recover the sparse graphical structures defined by the non-zero entries of the matrices $\mathbf{W}$ and $\{\mathbf{A}_p\}_{p=1}^{P}$ from the observed time series data $\mathbf{X}$.

4 Proposed Method: Rolling Window Causal Graph Evolution (RW-CGE)

This section provides a detailed technical specification of the RW-CGE framework, outlining each component of the algorithm from data segmentation to final anomaly scoring.

4.1 Framework Overview

The RW-CGE method is based on the principle that a stable system maintains a relatively consistent causal structure over time, whereas an anomalous event induces a significant and abrupt change in this structure. The framework is designed to detect these structural breaks by continuously monitoring the evolution of the system's learned causal graph. The algorithm proceeds as follows:

1. **Windowing:** A sliding window of a predefined size w and stride s is applied to the input multivariate time series $\mathbf{X}$. This partitions the data into a sequence of overlapping segments, $S_1, S_2, \ldots, S_N$.
2. **Causal Discovery:** For each data segment S_i, a dynamic causal discovery algorithm is executed to estimate the underlying window causal graph, $G_i = (\mathbf{V}, E_i)$.
3. **Evolution Quantification:** The structural change between consecutive graphs is quantified. For each graph G_i (where $i > 1$), a dissimilarity score, $D(G_i, G_{i-1})$, is computed.
4. **Anomaly Scoring:** The calculated dissimilarity score is assigned as the anomaly score for the time step corresponding to the end of the window S_i.
5. **Thresholding:** A dynamic thresholding mechanism is applied to the anomaly score time series to make the final detection decision.

4.2 Preprocessing and Model Hypotheses

The RW-CGE framework is designed around two core **applicational hypotheses** tailored for its industrial context.

The **first hypothesis** concerns the operational deployment: the framework assumes a two-phase implementation. An initial **offline setup phase** is permitted, where sufficient computational time is available to process historical data, establish a baseline "golden" causal model, and optimize detection parameters. This is followed by an **online monitoring phase**, which is computationally efficient enough to be executed in (near) real-time, allowing operators to check for anomalous behavior as new data arrives.

The **second hypothesis** defines the system's role as an expert-centric diagnostic aid. It is assumed that in the target industrial application, subject-matter experts (SMEs) are available. The framework is not intended to provide a fully autonomous root cause; rather, it is designed to **narrow the anomaly range** by flagging significant, interpretable graph changes. This edge-level diagnosis is then used by SMEs to pinpoint the physical root cause. This "expert-in-the-loop" design is complemented by the system's inherent privacy-preserving properties: by operating only on differenced time series and outputting abstracted graph weights, the framework avoids exposing critical or proprietary raw process data.

To ensure the validity of the causal discovery model, a critical preprocessing phase is required. The DYNOTEARS algorithm, like other VAR-based methods, assumes that the input time series are **stationary**. The workflow described in your 'preprocessing.py' script is designed to achieve this.

First, to stabilize the variance and ensure stationarity, each raw time series Z_t in a window is transformed using a log-transform followed by first-order differencing:

$$X_t = \log(Z_t + 1) - \log(Z_{t-1} + 1) \tag{2}$$

This processed series X_t is then standardized (scaled to zero mean and unit variance) before being passed to the optimizer.

Second, two key hyperparameters for Equation 4 must be determined: the optimal lag P and the L1 regularization penalty λ.

Optimal Lag Selection The lag P (which defines the shape of the lagged matrix $\mathbf{A}$) is determined empirically. For each stationary series X_t, a series of autoregressive models $AR(p)$ are fit for $p = 1, \ldots, p_{\max}$. The residuals ϵ_t of each model are subjected to the Ljung-Box test for autocorrelation. This test's null hypothesis, H_0, is that the residuals are independently distributed (i.e., white noise). We select the smallest lag p^* for which the test's p-value exceeds a significance threshold (e.g., $\alpha = 0.05$), indicating the model has sufficiently captured the series' dynamics. The final model lag P is then set as the maximum optimal lag found across all variables: $P = \max(p_1^*, \ldots, p_d^*)$.

Regularization Parameter The λ parameter in Equation 4 controls the trade-off between model fit (least-squares loss) and sparsity. This parameter is not re-calculated for every window, as this would be computationally prohibitive and would introduce a confounding variable. Instead, λ is optimized *once* on an initial training segment of data assumed to be normal. The optimal λ^* is selected by minimizing an information criterion, typically the Bayesian Information Criterion (BIC), over a grid of candidate λ values:

$$\lambda^* = \arg\min_{\lambda} \left(n \log(\hat{\sigma}_\lambda^2) + k_\lambda \log(n) \right) \tag{3}$$

where n is the number of samples in the window, $\hat{\sigma}_\lambda^2$ is the residual variance of the model fit with λ, and k_λ is the number of non-zero edges (parameters) in $\mathbf{W}$ and $\mathbf{A}$. This λ^* is then held constant for all subsequent rolling window estimations.

4.3 Dynamic Causal Graph Estimation

The core computational step within each window is the estimation of the causal graph G_i. This work employs a score-based method derived from the DYNOTEARS framework for this task [41, 42]. Its formulation as a continuous optimization problem avoids the combinatorial explosion inherent in other methods, making it computationally tractable for repeated estimation.

For each window segment S_i, the causal graph, represented by the contemporaneous matrix $\mathbf{W}$ and the lagged matrix $\mathbf{A}$, is found by solving the following optimization problem:

$$\min_{\mathbf{W},\mathbf{A}} \quad \frac{1}{2n} \|\mathbf{X}_i - \mathbf{X}_i\mathbf{W} - \mathbf{Y}_i\mathbf{A}\|_F^2 + \lambda(\|\mathbf{W}\|_1 + \|\mathbf{A}\|_1) \tag{4}$$

$$\text{subject to} \quad h(\mathbf{W}) = \operatorname{tr}(e^{\mathbf{W} \circ \mathbf{W}}) - d = 0 \tag{5}$$

Here, $\|\cdot\|_F^2$ is the squared Frobenius norm (least-squares loss), $\|\cdot\|_1$ is the L1 norm for sparsity, and the constraint $h(\mathbf{W}) = 0$ is a smooth, differentiable characterization of acyclicity for the contemporaneous graph $\mathbf{W}$ [41, 42, 47]. The output is a pair of sparse matrices $(\mathbf{W}_i, \mathbf{A}_i)$ defining the graph G_i.

4.4 Measuring Causal Evolution: A Multi-Faceted Approach

After obtaining the sequence of causal graphs $\{G_1, G_2, \ldots, G_N\}$, the evolution between them is quantified using a suite of four complementary dissimilarity metrics. This multi-faceted approach ensures sensitivity to different types of change, from discrete structural shifts to continuous variations in influence strength. Let $\mathbf{M}_i = (\mathbf{W}_i, \mathbf{A}_i)$ represent the combined adjacency matrices for graph G_i. The primary and complementary metrics are:

1. Structural Hamming Distance (SHD) The SHD serves as the primary measure of discrete topological change. It provides an absolute count of the number of edge additions and removals required to transform one graph into the next. Let E_i be the set of edges in G_i. The SHD is defined as:

$$D_{\text{SHD}}(G_i, G_{i-1}) = |E_i \setminus E_{i-1}| + |E_{i-1} \setminus E_i| \tag{6}$$

A non-zero SHD indicates that the fundamental causal pathways in the system have been rewired.

2. Complementary Weight-Based Metrics To capture changes that do not alter the graph structure but modify the strength of causal influences, three additional metrics are computed on the weight matrices $\mathbf{M}_i$:

- **Frobenius Distance:** Measures the overall magnitude of change across all edge weights, capturing systemic, distributed shifts.

$$D_{\text{Frob}}(G_i, G_{i-1}) = \|\mathbf{M}_i - \mathbf{M}_{i-1}\|_F \tag{7}$$

- **Spectral Distance:** Captures changes in the graph's global connectivity properties by comparing the eigenvalues of the adjacency matrices. It is sensitive to changes that affect the dominant modes of influence in the system.

$$D_{\text{Spec}}(G_i, G_{i-1}) = \|\text{eig}(\mathbf{M}_i) - \text{eig}(\mathbf{M}_{i-1})\|_2 \tag{8}$$

- **Max Edge Change:** Pinpoints the most significant localized perturbation by identifying the maximum absolute change in any single edge weight.

$$D_{\text{Max}}(G_i, G_{i-1}) = \max_{j,k} |\mathbf{M}_i(j,k) - \mathbf{M}_{i-1}(j,k)| \tag{9}$$

4.5 Anomaly Scoring and Detection

Instead of a single scalar score, the framework produces a four-dimensional anomaly score vector $\mathbf{a}_t$ at each time step t, corresponding to the end of window S_i:

$$\mathbf{a}_t = \begin{bmatrix} D_{\text{SHD}}(G_i, G_{i-1}) \\ D_{\text{Frob}}(G_i, G_{i-1}) \\ D_{\text{Spec}}(G_i, G_{i-1}) \\ D_{\text{Max}}(G_i, G_{i-1}) \end{bmatrix} \tag{10}$$

This vector allows for a granular diagnosis, as different anomalies may manifest in different components.

To establish detection thresholds, we depart from dynamic rolling-window methods. Instead, we employ a **bootstrapped thresholding** approach using a baseline (or "golden") dataset assumed to represent normal, stable operation.

For each metric $m \in \{\text{SHD, Frob, Spec, Max}\}$, we first build an empirical distribution of "normal" dissimilarity by computing N pairwise comparisons $D_m(G_i, G_j)$ between graphs G_i and G_j randomly sampled from this baseline data. The final, static threshold θ^m for that metric is set at the 95th percentile (or other high quantile) of this empirical distribution:

$$\theta^m = \text{percentile}_{95}(\{D_m(G_i, G_j) : i, j \sim \text{Baseline}\}) \tag{11}$$

This non-parametric method effectively captures the system's natural variability during normal operation without imposing an assumed distribution (e.g., Gaussian) on the scores.

An anomaly is flagged at window i using a **weighted ensemble voting** mechanism. Each metric's normalized score $s_m = D_m(G_i, G_{i-1})/\theta^m$ is combined with learned weights w_m to produce an ensemble score:

$$S_{\text{ensemble}} = \sum_m w_m \cdot s_m \tag{12}$$

An anomaly is detected when $S_{\text{ensemble}} > 1.0$. The weights are optimized to prioritize structural changes (SHD: 0.40) over magnitude changes (Frobenius: 0.25, Spectral: 0.20, Max Edge: 0.15), reflecting their relative importance for detection accuracy.

5 Experimental Evaluation

This section presents a rigorous empirical validation of the RW-CGE framework.

5.1 Datasets and Baselines

Datasets

- **NASA SMAP/MSL (Telemanom):** Primary real-world benchmark with multivariate telemetry data from two NASA missions [25, 28]. It contains expert-verified labels for anomalous periods [26].
- **Server Machine Dataset (SMD):** A public benchmark from a large internet company for IT operations monitoring [28, 45].

Baselines

- **Classical Methods:** Isolation Forest (IF) [30, 51] and Matrix Profile (MP) [29, 50, 55].
- **Deep Learning Methods:** LSTM-AE [9, 24] and TranAD [43].
- **Causality-Informed Method:** Causanom (with a static graph) [14].

5.2 Experimental Validation and Current Limitations

The RW-CGE framework was validated on a large-scale, proprietary multivariate time series dataset provided by our industrial partner. Due to the sensitive and confidential nature of this operational data, specific quantitative performance metrics and the dataset itself cannot be published in this paper. Instead, we provide a qualitative summary of the model's performance and discuss its current limitations, which inform the direction of future work.

On the target industrial dataset, the framework demonstrated its primary capability: the successful detection of known anomalous events. These events consistently manifested as significant, high-magnitude spikes in the Structural Hamming Distance (SHD) and other graph dissimilarity scores. The temporal precision of the detection was, by design, limited to the width of the sliding window, meaning anomalies were correctly localized to the specific window in which the causal structure break occurred, rather than to an exact timestamp.

A full quantitative benchmark on public datasets, such as the Telemanom SMD dataset [25], was explored but deemed computationally prohibitive within the scope of this research. Initial scalability tests indicated a runtime exceeding 400 hours to process a single anomaly segment with the current implementation. This is a direct consequence of the computationally intensive, one-at-a-time anomaly detection approach (a limitation discussed further in the conclusion).

To ensure transparency and facilitate future research, the complete implementation of the RW-CGE framework is publicly available. We are actively working on performance optimizations, and any future results on public benchmarks will be documented and kept up-to-date in the project's public code repository[1].

6 Discussion

The experimental results provide strong evidence for the efficacy of the RW-CGE framework.

6.1 Strengths and Limitations

Strengths The primary strengths are superior detection accuracy, conceptual robustness to covariate shifts (changes in data distribution that preserve causal mechanisms), and unparalleled interpretability for root cause analysis. The output is not just a score but a diagnostic tool.

Limitations Key limitations of the proposed framework include:

- **Computational Complexity:** Causal discovery is computationally intensive (e.g., $O(d^3)$ per window for DYNOTEARS), which can be a bottleneck for very high-dimensional or real-time systems.
- **Underlying Model Assumptions:** The method's validity hinges on the assumptions of the chosen causal discovery algorithm. DYNOTEARS assumes linearity, contemporaneous acyclicity, and no hidden confounders. Violations of these assumptions can lead to inaccurate graph structures.
- **Hyperparameter Sensitivity:** The model's performance is sensitive to the choice of the window size w, stride s, and the L1 regularization parameter λ.
- **Lack of Automatic Re-baselining:** The framework is designed to detect a single anomalous event or regime shift from a 'golden' baseline. It does not include a mechanism for automatic re-baselining. Consequently, after a permanent regime shift occurs, the model will flag the initial transition but may fail to detect *subsequent* anomalies, as the new (but anomalous) state becomes the de facto reference for comparison.
- **Limited Temporal and Diagnostic Granularity:** Due to its intrinsic rolling-window design, the framework's temporal precision is limited. An anomaly can only be localized to the *window* in which it occurred, not to a precise time step. Furthermore, while the model can identify *that* a change occurred, it cannot autonomously classify the *type* or physical root cause of the anomaly, requiring expert-in-the-loop interpretation.

6.2 Future Research Directions

- **Scalability and Efficiency:** Explore more scalable discovery algorithms or develop an *online* learning version that incrementally updates the graph.
- **Handling Non-Linearity and Confounding:** Integrate more advanced discovery engines that can handle non-linear relationships and detect the presence of latent confounders [33, 43].

[1] https://github.com/tejoker/Unaite_summer_research

- **Hybrid Models:** Combine the strengths of deep learning with the structural robustness of the causal approach, for instance by using the causal evolution score as a feature for a deep learning model.

7 Conclusion

This paper has argued for a fundamental shift in the paradigm of time series anomaly detection, moving from a purely statistical viewpoint to a causal one. The core thesis—that anomalies are manifestations of structural breaks in a system's causal mechanisms—provides a principled foundation for overcoming the brittleness and opacity of contemporary models.

The proposed Rolling Window Causal Graph Evolution (RW-CGE) framework successfully translates this thesis into a practical, modular algorithm. By learning a sequence of causal graphs and using their structural evolution as a direct anomaly signal, RW-CGE shifts the goal from simple "point-in-time" detection to providing actionable, edge-level insights for expert-in-the-loop analysis. This transformation from a "black box" detector to an interpretable diagnostic tool is essential for building trust in critical industrial monitoring systems.

While the computational expense of this novel approach limited extensive quantitative benchmarking in this initial study, the framework's design and qualitative validation on industrial data lay the groundwork for a new class of causal detectors. Future work will focus on computational optimization, usage of Deep Learning and comprehensive validation on public benchmarks.

As this research was conducted as part of an industrial research internship, i am actively seeking feedback and welcome any advice or opportunities for collaboration to extend and refine this work.

References

[1] Neptune.ai: Mlops platform for experiment tracking and model registry. https://neptune.ai/, .

[2] Sentry blog. https://blog.sentry.io/, .

[3] Insights into anomaly detection: A survey and comparative analysis of techniques for time series data from industrial environment. In *Proceedings of the 2024 3rd International Conference on Applied Mechanics, Mechatronics and Smart Drivetrains (ICAMMSD 2024)*. Atlantis Press, 2024.

[4] Anonymous Authors. Transformer based anomaly detection on multivariate time series subledger data. Technical report, Amazon Science.

[5] Anonymous Authors. Dycast: Learning dynamic causal structure from time series. OpenReview, https://openreview.net/forum?id=WjDjem8mWE, 2025.

[6] G. Di Bona et al. On the identifiability of causal graphs with multiple environments. https://arxiv.org/html/2510.13583v1, 2025.

[7] João Carvalho, Mengtao Zhang, Robin Geyer, Carlos Cotrini, and Joachim M Buhmann. Invariant anomaly detection under distribution shifts: A causal perspective. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023.

[8] Zhaobo Chen et al. Causal-discovery-based root-cause analysis and its application in time-series prediction error diagnosis. https://arxiv.org/html/2411.06990v2, 2024.

[9] Hyun-Woo Choi et al. Anomaly detection using an ensemble of multi-point lstms. *Electronics*, 2023.

[10] Zahra Zamanzadeh Darban, Geoffrey I. Webb, Shirui Pan, Charu C. Aggarwal, and Mahsa Salehi. Deep learning for time series anomaly detection: A survey. https://www.researchgate.net/publication/365299032_Deep_Learning_for_Time_Series_Anomaly_Detection_A_Survey, 2022.

[11] Zahra Zamanzadeh Darban, Geoffrey I. Webb, Shirui Pan, Charu C. Aggarwal, and Mahsa Salehi. Deep learning for time series anomaly detection: A survey. Technical report, IBM Research, 2024.

[12] Zahra Zamanzadeh Darban et al. Deep learning for time series anomaly detection: A survey. https://ar5iv.labs.arxiv.org/html/2211.05244, 2022.

[13] Zahra Zamanzadeh Darban et al. Deep learning for time series anomaly detection: A survey (v3). https://arxiv.org/html/2211.05244v3, 2022.

[14] Zachary D. DeLand et al. Causanom: Anomaly detection with flexible causal graphs. In *Proceedings of the 36th International FLAIRS Conference*, 2023.

[15] Zachary D. DeLand et al. (pdf) causanom: Anomaly detection with flexible causal graphs. https://www.researchgate.net/publication/370627542_Causanom_Anomaly_Detection_With_Flexible_Causal_Graphs, 2023.

[16] Stumpy Developers. The matrix profile — stumpy 1.13.0 documentation. https://stumpy.readthedocs.io/en/latest/Tutorial_STUMPY_Basics.html.

[17] Jan Ditzen, Yiannis Karavias, and Joakim Westerlund. Testing and estimating structural breaks in time series and panel data in stata. https://arxiv.org/abs/2110.14550, 2021.

[18] Simon Dürr et al. Root cause analysis in industrial manufacturing: A scoping review of current research, challenges and the promises of ai-driven approaches. https://www.researchgate.net/publication/386353438_Root_Cause_Analysis_in_Industrial_Manufacturing_A_Scoping_Review_of_Current_Research_Challenges_and_the_Promises_of_AI-Driven_Approaches, 2024.

[19] Yue Gao et al. Causal discovery-driven change point detection in time series. https://www.researchgate.net/publication/382145751_Causal_Discovery-Driven_Change_Point_Detection_in_Time_Series, 2024.

[20] Yue Gao et al. Causal discovery-driven change point detection in time series. https://raw.githubusercontent.com/mlresearch/v258/main/assets/gao25g/gao25g.pdf, 2025.

[21] Sergio García-Márquez et al. A review of anomaly detection in spacecraft telemetry data. *Applied Sciences*, 15(10):5653, 2025.

[22] Zhiqiang Ge et al. LSTM-Based VAE-GAN for Time-Series anomaly detection. *Sensors*, 20(13): 3738, 2020.

[23] GeeksforGeeks. Anomaly detection using isolation forest. https://www.geeksforgeeks.org/machine-learning/anomaly-detection-using-isolation-forest/.

[24] Zhong Hong. Anomaly detection in time series data using lstm autoencoders. Medium, https://medium.com/@zhonghong9998/anomaly-detection-in-time-series-data-using-lstm-autoencoders-51fd14946fa3, 2024.

[25] Kyle Hundman. telemanom: A framework for using lstms to detect anomalies in multivariate time series data. GitHub Repository, https://github.com/khundman/telemanom.

[26] Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018.

[27] Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018.

[28] Elise Jiuqizhang. TS-AD-Datasets: Public datasets for time series anomaly detection. GitHub Repository, https://github.com/elisejiuqizhang/TS-AD-Datasets.

[29] Eamonn Keogh. The ucr matrix profile page. https://www.cs.ucr.edu/~eamonn/MatrixProfile.html.

[30] ksmooi. Anomaly detection in time-series (isolationforest). Kaggle, https://www.kaggle.com/code/ksmooi/anomaly-detection-in-time-series-isolationforest.

[31] Daniel Lakey and Tim Schlippe. Anomaly detection in spacecraft telemetry: Forecasting vs. classification. Presentation Slides, https://research-karlsruhe.de/pubs/SCC2024_Lakey+Schlippe_AnomalyDetection_slides.pdf, 2024.

[32] Pau Lanau and A. Arjona-Medina. Explainable anomaly detection in spacecraft telemetry. https://www.researchgate.net/publication/378214516_Explainable_anomaly_detection_in_spacecraft_telemetry, 2024.

[33] Alex G. Lee. Causal ai: Current state-of-the-art & future directions. Medium, https://medium.com/@alexglee/causal-ai-current-state-of-the-art-future-directions-c17ad57ff879, 2025.

[34] Chuan-Ming Lin et al. Root cause analysis in microservice using neural granger causal discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

[35] Shou-De Lin et al. Review of causal discovery methods based on graphical models. *Journal of the American Statistical Association*, 2019.

[36] Yang Liu et al. TiTAD: Time-Invariant transformer for multivariate time series anomaly detection. *Electronics*, 14(7):1401, 2025.

[37] Yuhang Liu et al. Identifiable latent neural causal models. https://arxiv.org/html/2403.15711v1, 2024.

[38] Yuhang Liu et al. Turning challenges into opportunities: How distribution shifts enhance identifiability in causal representation learning. OpenReview, https://openreview.net/forum?id=q07DDpu8Xb, 2025.

[39] S. W. Mogensen, K. Rathsman, and P. Nilsson. Causal discovery in a complex industrial system: A time series benchmark. In *Proceedings of Machine Learning Research*, 2024.

[40] Nearshore. Anomaly detection with lstm autoencoders & neural networks in time series data analysis. https://nearshore-it.eu/articles/anomaly-detection-with-lstm/.

[41] Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Paul Beaumont, Konstantinos Georgatzis, and Bryon Aragam. DYNOTEARS: Structure learning from Time-Series data. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108 of *Proceedings of Machine Learning Research*. PMLR, 2020.

[42] Roxana Pamfil et al. Dynotears: Structure learning from time-series data. https://arxiv.org/pdf/2002.00498, 2020.

[43] Ioannis Papagiannopoulos et al. Interpretability of causal discovery in tracking deterioration in a highly dynamic process. *IEEE Access*, 2024.

[44] U. N. Runji et al. Survey and evaluation of causal discovery methods for time series (extended abstract). In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI 2023)*, 2023.

[45] Salesforce. ts_datasets.anomaly package — merlion 1.3.1 documentation. https://opensource.salesforce.com/Merlion/latest/ts_datasets.anomaly.html.

[46] Anil K. Seth, Adam B. Barrett, and Lionel Barnett. Misunderstandings regarding the application of granger causality in neuroscience. *PNAS*, 2017.

[47] Chang Shi. Time-varying dags with notears. https://changshiraine.github.io/portfolio/portfolio-5/.

[48] Patrick A. Stokes and Patrick L. Purdon. A study of problems encountered in granger causality analysis from a neuroscience perspective. *PNAS*, 2017.

[49] Athanasios S. Vlontzos et al. A causality-inspired approach for anomaly detection in a water treatment testbed. *Sensors*, 23(1):257, 2023.

[50] Phillip Wenig. Using the matrix profile to detect anomalies in time series. Medium, https://medium.com/@pw33392/using-the-matrix-profile-to-detect-anomalies-in-time-series-bca14883e0fb, 2025.

[51] Wikipedia. Isolation forest. https://en.wikipedia.org/wiki/Isolation_forest, .

[52] Wikipedia. Granger causality. https://en.wikipedia.org/wiki/Granger_causality, .

[53] Hongzuo Xu, Guansong Pang, Yijie Wang, and Yongjun Wang. Deep isolation forest for anomaly detection. https://arxiv.org/abs/2206.06602, 2022.

[54] Chin-Chia Michael Yeh et al. Matrix profile for anomaly detection on multidimensional time series. In *2024 IEEE International Conference on Data Mining (ICDM)*, 2024.

[55] Chin-Chia Michael Yeh et al. Matrix profile for anomaly detection on multidimensional time series. https://arxiv.org/html/2409.09298v1, 2024.

[56] Jia-Liang Yin et al. A survey on time series anomaly detection. *Journal of Beijing Jiaotong University*, 2024.

[57] Li Yin-Yin-Xin et al. Granger causality: A review and recent advances. *Annual Review of Economics*, 2023.

[58] Li Yin-Yin-Xin et al. Causal inference meets deep learning: A comprehensive survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[59] Aston Zhang, Zachary C. Lipton, Mu Li, and Alexander J. Smola. 4.7. environment and distribution shift. https://d2l.ai/chapter_linear-classification/environment-and-distribution-shift.html.