Information Processing and Management 64 (2027) 105120

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

TFCAD: Time–Frequency Causal Anomaly Detection for Multivariate Time Series☆

Wenxuan He, Mingjing Du*, Wenbin Zhang

Jiangsu Key Laboratory of Educational Intelligent Technology, School of Artificial Intelligence and Computer Science, Jiangsu Normal University, Xuzhou, 221116, China

ARTICLE INFO

Keywords:
Multivariate time series
Anomaly detection
Granger causal discovery
Time–frequency domain

ABSTRACT

Unsupervised anomaly detection in multivariate time series (MTS) plays a crucial role in ensuring the reliability and safety of complex industrial systems. However, existing MTS anomaly detection methods often fail to extract comprehensive features or explicitly model directional predictive dependencies beyond static pairwise correlations. To address these challenges, this paper proposes TFCAD (Time–Frequency Causal Anomaly Detection for multivariate time series), a novel framework designed to bridge time–frequency representation learning with Granger causal discovery for robust anomaly detection. Specifically, TFCAD transforms raw sequences into rich time–frequency representations, from which it dynamically uncovers directed Granger causal relationships among variables. This learned Granger-based structure is subsequently injected into the detection model via a graph convolutional network, serving as a predictive prior that provides explicit guidance and substantially enhances the reconstruction fidelity of normal patterns. Crucially, by decoupling the Granger causal discovery from the reconstruction task, TFCAD mitigates gradient conflicts between structural learning and anomaly reconstruction and provides interpretable predictive priors. Extensive experiments on 6 benchmarks comprising over 3.4 million observations show that TFCAD outperforms 13 state-of-the-art baselines by average margins of 23.87% in F1 score and 26.01% in AUC-ROC. The source code is publicly available at https://github.com/Du-Team/TFCAD.

1. Introduction

The extensive integration of monitoring devices and advancements in smart digitization have resulted in a surge of multivariate time series (MTS) data within both industrial landscapes and modern autonomous applications (Chen et al., 2025). Simultaneously, unsupervised anomaly detection, as a powerful unsupervised learning technique, has attracted increasing attention from researchers by revealing hidden patterns and detecting anomalies in unlabeled time series data (Egilmez et al., 2017). Notably, real-world MTS inherently exhibit complex non-equilibrium dynamics, characterized by fundamental time irreversibility (Yao et al., 2021). This temporal asymmetry suggests that inter-variable dependencies can exhibit directional and time-varying behavior. Consequently, traditional static correlations or strictly time-domain representations are highly insufficient for capturing these evolving structural dependencies, necessitating a dynamic time–frequency approach.

☆ **Acknowledgments**
This work is supported by the Qinglan Project of Jiangsu Province of China, the National Natural Science Foundation of China (No. 62006104), Jiangsu Normal University Faculty Research Excellence Development Program (No. JSNUGZL2026032).

* Corresponding author.
E-mail addresses: hewx@jsnu.edu.cn (W. He), dumj@jsnu.edu.cn (M. Du), zwbwen@jsnu.edu.cn (W. Zhang).

https://doi.org/10.1016/j.ipm.2026.105120
Received 2 April 2026; Received in revised form 15 August 2026; Accepted 22 August 2026
Available online 1 September 2026

Due to the exceptional efficacy of deep learning, these techniques have been increasingly adopted for anomaly detection in MTS (Wu et al., 2023). Among these, Graph Neural Networks (GNNs) have emerged as an effective technique to model multivariate dependencies by leveraging the inherent relationships between variables. Concurrently, frequency-domain analysis provides an effective approach for extracting multi-scale temporal characteristics through spectral analysis. These approaches exhibit remarkable efficacy in managing intricate, perturbed, and high-dimensional temporal sequences when contrasted with conventional paradigms.

Despite such progress, current MTS anomaly detection frameworks still face formidable hurdles, especially in capturing complex patterns and inter-variable correlations during the feature learning stage. First, regarding feature extraction, recent SOTA methods — such as AMFormer (Zhong et al., 2024) and DADA (Shentu et al., 2025) — predominantly rely on single-domain strategies, focusing exclusively on temporal information or spectral patterns. This single-domain perspective overlooks complementary information across both domains, resulting in an insufficient feature representation. To accurately identify anomalies within non-equilibrium systems, it is crucial to separate transient anomalous signals from the complex, non-stationary structural background. Drawing a cross-disciplinary analogy to background purification in hyperspectral anomaly detection (Huang et al., 2021) — where multi-scale morphological profiles effectively isolate small targets from cluttered backgrounds — we recognize that robust MTS anomaly detection requires a similar multi-scale time–frequency paradigm. To bridge this gap and realize such a mechanism, we explicitly introduce our proposed Multi-scale Wavelet Encoder (MWE). By leveraging a lifting-scheme wavelet transform, the MWE hierarchically decomposes the irreversible signals, effectively isolating high-frequency anomalous transients from low-frequency background trends before causal discovery. Second, in terms of inter-variable modeling, even advanced graph-based approaches, exemplified by MST-GAT (Ding et al., 2023) and MTGFLOW (Zhou et al., 2024), typically treat all variable pairs equally, without explicitly modeling directional predictive dependencies among variables.

To overcome these challenges, this paper introduces TFCAD (time–frequency Causal Anomaly Detection for multivariate time series), a novel framework designed to unify time–frequency representation learning with Granger causal structure learning for robust anomaly detection. TFCAD adopts a two-phase architecture where causal discovery serves as an auxiliary module to provide interpretable, structured guidance for the primary reconstruction-based anomaly detection process. In the first phase, the MWE is employed to decompose raw signals into hierarchical time–frequency representations to capture both global trends and localized fluctuations. These rich representations then feed into a dedicated Time–Frequency Causal AutoEncoder (TFCAE) module, which discovers latent Granger causal relationships among variables by imposing structured constraints on the learned Granger-based directed graph. In the second and primary phase, anomaly detection is formulated as a causality-guided reconstruction task. The causal structure identified in the first phase is embedded into the reconstruction network via a Causality-Guided Graph Convolutional Network (CGGCN). The entire framework is trained via a composite loss function with structurally decoupled gradient paths across the two phases, allowing causal insights to guide detection without confounding the learning objectives.

In summary, the main contributions of this work are as follows:

- We propose TFCAD, an unsupervised MTS anomaly detection framework that effectively integrates time–frequency feature extraction and Granger-based structure learning to enhance detection performance.
- We introduce the MWE module, which utilizes lifting-scheme-based wavelet transformations to capture time–frequency characteristics of non-stationary time series across different temporal scales, providing rich information for downstream modeling.
- We design the TFCAE module to discover interpretable Granger-based directed structures from time–frequency representations and leverage them to dynamically guide GCN-based feature refinement, enabling the model to learn more discriminative representations.

1.1. Research objectives

The primary goal of this research is to develop a robust and interpretable unsupervised anomaly detection framework for MTS that overcomes the limitations of single-domain feature extraction and superficial correlation modeling. To achieve this, the following specific objectives are defined:

- To design a lifting-scheme-based mechanism that decomposes non-stationary MTS into hierarchical representations, capturing both global trends and localized transient fluctuations that are typically obscured in purely temporal or spectral domains.
- To integrate the discovered Granger-based structural dependencies into a graph-based reconstruction model, constraining feature aggregation to follow directed predictive dependencies to improve detection sensitivity and reduce false positives in complex industrial systems.
- To establish a decoupled training strategy that allows causal discovery to provide structural guidance to the anomaly detection process without confounding the respective learning objectives, thereby mitigating gradient conflicts during graph structure learning.

The remaining parts of the paper are organized as follows: Section 2 discusses the related work, Section 3 introduces the theoretical background, Section 4 details the methodology, Section 5 presents the experimental evaluations, Section 6 provides discussions and implications, and Section 7 concludes the paper.

2. Related work

2.1. Deep multivariate time series anomaly detection

Deep learning methods have become central to MTS anomaly detection due to their strong data representation capabilities (He et al., 2024; Wei et al., 2025). Among these methods, reconstruction-based approaches are widely adopted. They typically use an encoder–decoder architecture to model the underlying distribution of normal samples, flagging observations whose reconstruction error exceeds a statistically defined threshold as anomalous (Tuli et al., 2022).

Early endeavors in this domain, exemplified by OmniAnomaly (Su et al., 2019), concentrate predominantly on modeling temporal dependencies, yet afford limited consideration to the intricate cross-dimensional interactions among features. To bridge this lacuna, subsequent methods — MSCRED (Zhang et al., 2019) and MST-GAT (Ding et al., 2023) — simultaneously model inter-variable and temporal dynamics by integrating convolutional encoders with graph-attention mechanisms, thereby unifying spatial and temporal representation learning within a single framework (He et al., 2026). However, a fundamental limitation pervades these methods: their extreme dependence on a scalar, aggregated reconstruction error; this deficiency intrinsically precludes the discrimination of distinct anomaly categories whose statistical error signatures are commensurate.

More recently, the field has evolved toward the integration of diverse learning paradigms to augment discriminative capacity. For instance, MultiRC (Hu et al., 2024) integrates reconstruction learning with contrastive learning, while IGCL (Zhao et al., 2025) introduces an importance-based generative contrastive framework. In parallel, Wang et al. (2025) develop a multi-granularity contrastive zero-shot learning model based on attribute decomposition, highlighting the efficacy of hierarchical feature representation in zero-shot and unsupervised learning contexts. Other methods leverage generative models, such as GLAD (Yao et al., 2024), which uses a diffusion model, and DADA (Shentu et al., 2025), which employs dual adversarial decoders. Beyond these specific architectures, the emergence of Large Language Models (LLMs), such as Time-LLM (Jin et al., 2024), has introduced a foundation-model paradigm for sequence analysis.

Despite their novelty, these advanced paradigms face distinct hurdles. On one hand, hybrid architectures like MultiRC (Hu et al., 2024) and IGCL (Zhao et al., 2025) often require exhaustive hyperparameter tuning to prevent gradient domination, while complex generative priors in GLAD (Yao et al., 2024) can introduce high-dimensional instabilities that hinder convergence. On the other hand, LLM-based approaches are frequently constrained by high computational latency and a lack of explicit, domain-specific causal interpretability in industrial monitoring settings.

In contrast, our proposed TFCAD framework circumvents these optimization pitfalls through a synergistic yet gradient-decoupled design. By utilizing a non-differentiable acyclic projection, TFCAD effectively isolates the causal discovery phase from downstream reconstruction errors, preventing downstream reconstruction gradients from interfering with graph structure learning.

2.2. Time–frequency domain time series analysis

Frequency domain analysis has become a prominent approach in time series anomaly detection due to its ability to capture multi-scale representations (Fang et al., 2024). In contrast to Fourier-based representations, wavelet transforms afford optimal time–frequency localization, thereby enabling the exact isolation of transient characteristics. The classic Discrete Wavelet Transform (DWT), employed by methods such as MEGA (J. Wang et al., 2023), effectively decomposes signals into distinct frequency sub-bands. However, DWT is limited by its reliance on fixed wavelet basis functions, which may not optimally adapt to diverse data characteristics. The Lifting Wavelet Transform (LWT) improves upon this with a more flexible spatial-domain construction, yet traditional implementations still depend on hand-crafted prediction and update operators, limiting their full adaptive potential.

Recent advancements have focused on making these transformations more adaptive. AdaWaveNet (Yu et al., 2024) represents a significant leap, introducing a multi-scale wavelet transformation using a lifting scheme with neural networks. The model automatically discovers the optimal wavelet bases for specific datasets, overcoming the limitations of fixed functions. To remedy the spectral granularity limitation, CATCH (Wu et al., 2025) introduces a frequency-domain patching paradigm that tokenizes narrow-band components, enabling precise anomaly localization within specific spectral slots.

To harness the theoretical strengths of the lifting scheme in time–frequency localization, we introduce the Multi-scale Wavelet Encoder. Instead of relying on complex parameter-heavy transformations, MWE employs a streamlined lifting mechanism to efficiently decompose non-stationary data into multi-scale representations that are highly amenable to downstream causal discovery, creating a robust foundation for the overall framework.

2.3. Causal graph neural networks

Causal graph neural networks have emerged as a prominent paradigm for MTS analysis, which infer underlying causal relationships to enhance learning. This paradigm encompasses a spectrum of methods. Based on how the causal structure is derived, existing methods can be broadly categorized into two types: predefined graph methods that rely on a predefined, static graph constructed prior to model training, and learned graph methods that jointly learn graph discovery and representation learning. The predefined graph approach is limited because the static graph may not reflect dynamic Granger causal relationships, hindering model adaptability.

Consequently, learned graph methods have become the focus of modern research. A significant breakthrough is NOTEARS (Zheng et al., 2018), which recasts the combinatorial challenge of inferring a Directed Acyclic Graph (DAG) into a continuous optimization

Table 1
Summary of the reviewed state-of-the-art methods for multivariate time series anomaly detection.

Category	Representative methods	Characteristics	Pros (Advantages)	Cons (Limitations)
Spatial–Temporal reconstruction	OmniAnomaly, MSCRED, MST-GAT	Utilizes encoder–decoder architectures with convolutional or graph-attention mechanisms to model inter-variable and temporal dynamics.	Unifies spatial and temporal representation learning within a single framework.	Extreme dependence on a scalar aggregated reconstruction error; fails to discriminate distinct anomaly categories.
Hybrid contrastive & Generative models	MultiRC, IGCL, GLAD, DADA	Integrates diverse learning paradigms such as contrastive learning, diffusion models, or adversarial dual decoders.	Significantly augments the discriminative capacity and distribution modeling for zero-shot and unsupervised scenarios.	Requires exhaustive hyperparameter tuning; complex generative priors often introduce high-dimensional instabilities hindering convergence.
LLM-based sequence analysis	Time-LLM	Adapts Large Language Models (LLMs) to serve as foundation models for time series sequence analysis.	Offers powerful generalizability and leverages pre-trained foundation knowledge.	Suffers from high computational latency and lacks explicit, domain-specific causal interpretability for industrial monitoring.
Time–frequency & Wavelet analysis	MEGA, AdaWaveNet, CATCH	Decomposes non-stationary signals into distinct frequency sub-bands or adaptive multi-scale representations using wavelet transforms.	Provides optimal time–frequency localization; excels at capturing transient characteristics and specific spectral anomalies.	Traditional methods rely on fixed basis functions, which limit adaptability; heavy transformations can be computationally complex.
Causal graph neural networks	NOTEARS, DAG-GNN, AERCA	Casts DAG structure learning into a continuous optimization framework; utilizes Granger causality or variational autoencoders to infer dynamic dependencies.	Enables fully data-driven causal discovery; highly valuable for isolating structural dependencies and root cause analysis.	Typically operates on raw time series rather than enriched time–frequency representations; complex loss balancing can affect stability.

framework by proposing a smooth characterization of acyclicity. Building on this, DAG-GNN (Yu et al., 2019) proposes a deep generative model that learns DAG structures from complex nonlinear data by applying a variant of the acyclicity constraint within a variational autoencoder parameterized by a GNN. This approach provides a principled basis for fully data-driven causal discovery. Recently, the Agentic Stream of Thought (ASoT) framework (Meier et al., 2025) has been proposed for structured knowledge-based causal discovery, which leverages agentic reasoning to integrate structural priors into complex system analysis.

A major application of causal discovery in time series is root cause analysis. AERCA (Han et al., 2025), for example, integrates Granger causal discovery to model anomalies as interventions and explicitly learns causal relationships to pinpoint their origins. This work, along with related research on quantifying causal influences (Janzing et al., 2013), demonstrates the value of using a formal causal framework not only to detect anomalies but also to understand them.

While building upon the principles established by methods like DAG-GNN and AERCA, our proposed TFCAE module introduces a distinct approach. The key innovation lies in performing causal discovery not on the raw time-series data, but on the rich, multi-scale time–frequency representations generated by the MWE module. This strategic choice allows TFCAE to identify complex dependencies that may be more prominent in the time–frequency domain. Furthermore, unlike AERCA, which primarily focuses on root cause analysis, our framework utilizes the learned causal graph to directly guide the graph convolution operations in the downstream anomaly detection model. This yields a synergistic learning framework in which rich time–frequency features support the learning of directed predictive dependencies in the Granger sense, while the resulting graph serves as a structural prior for downstream anomaly detection.

To provide a clear and structured comparison of the existing literature, the methodological characteristics, primary advantages, and inherent limitations of these state-of-the-art approaches are comprehensively summarized in Table 1.

3. Theoretical background

Before introducing the proposed model, we briefly outline the theoretical foundations on which it is built. Specifically, we review the wavelet transform via the lifting scheme and the formulation of Granger causality, as these concepts underpin our method to time–frequency representation and causal structure detection.

3.1. Wavelet transform via lifting scheme

The wavelet transform is a powerful tool for the multi-scale analysis of signals, offering simultaneous localization in both time and frequency domains. This capability is vital for processing non-stationary time series. For a given signal $y(x)$, its wavelet decomposition can be expressed as:

$$y(x) = \sum_{a,b} c_{a,b}\psi_{a,b}(x), \tag{1}$$

where $\psi_{a,b}(x)$ are the wavelet basis functions at scale a and translation b, and $c_{a,b}$ are the corresponding wavelet coefficients. Traditional wavelet transforms depend on fixed, hand-crafted basis functions (e.g., Haar, Daubechies), which inherently restricts their adaptability to signals with diverse characteristics.

To overcome this limitation, our work leverages the lifting scheme—a second-generation wavelet construction that is performed entirely in the spatial domain. This approach provides the flexibility to design multi-scale, nonlinear wavelet transforms. The scheme consists of three primary stages. First, in the split stage, the input signal is partitioned into two disjoint subsets, typically the even-indexed samples ($\boldsymbol{x}^{\text{eve}}$) and odd-indexed samples ($\boldsymbol{x}^{\text{odd}}$). Subsequently, the predict phase deploys a prediction operator $P(\cdot)$ to exploit the inter-set correlation, enabling the estimation of one subset from its counterpart. The resultant prediction residual yields the detail coefficients $\boldsymbol{d}$, which encode the high-frequency content.

$$\boldsymbol{d} = \boldsymbol{x}^{\text{odd}} - P(\boldsymbol{x}^{\text{eve}}). \tag{2}$$

Finally, the update stage applies an update operator $U(\cdot)$ to the derived detail coefficients $\boldsymbol{d}$, refining the complementary subset so that essential signal properties — e.g., moments — are preserved, yielding smoothed approximation coefficients $\boldsymbol{c}$.

$$\boldsymbol{c} = \boldsymbol{x}^{\text{eve}} + U(\boldsymbol{d}). \tag{3}$$

Critically, by implementing the prediction and update operators as neural network modules, the lifting scheme evolves from a fixed linear transformation into a flexible, nonlinear feature extraction mechanism. This design allows the model to effectively decompose the signal into hierarchical components, capturing complex time–frequency characteristics through the deep stacking of these layers. Consequently, it generates a rich, multi-scale representation that preserves the signal's intrinsic geometry, constituting the nucleus of our feature-extraction pipeline.

3.2. Granger causality

Granger causality serves as a fundamental statistical criterion for evaluating whether the historical observations of one time series significantly enhance the predictive accuracy of another. In its classical formulation rooted in linear vector autoregressive models, a time series $\mathcal{X}$ is defined to Granger-cause $\mathcal{Y}$ if the inclusion of lagged values of $\mathcal{X}$ results in a statistically significant reduction in the forecasting error of $\mathcal{Y}$, conditional on the past history of $\mathcal{Y}$ alone. While originally predicated on linear dependencies, this framework has recently been extended to address nonlinear dynamics. Recent approaches leverage deep neural network architectures to approximate complex, data-driven functional mappings, thereby generalizing the detection of causal structures beyond traditional linear constraints.

Let a multivariate time series be denoted by $\mathcal{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_T)$, where $\boldsymbol{x}_t \in \mathbb{R}^M$ is an M-dimensional vector at time t. The data-generating process for the jth variable is formalized by the structural equation:

$$x_t^{(j)} := f^{(j)}(\boldsymbol{x}_{\le t-1}^{(1)}, \dots, \boldsymbol{x}_{\le t-1}^{(M)}) + u_t^{(j)}, \tag{4}$$

where $\boldsymbol{x}_{\le t-1}^{(i)}$ represents the history of the ith variable up to time $t-1$, $f^{(j)}(\cdot)$ is a nonlinear function capturing the temporal dependencies, and $u_t^{(j)}$ is an exogenous variable representing external, unmodeled influences at time t. Within this framework, variable i is defined to Granger-cause variable j if the predictive function $f^{(j)}(\cdot)$ depends explicitly on the history $\boldsymbol{x}_{\le t-1}^{(i)}$. Accurately identifying these directed dependencies is fundamental to elucidating the underlying dynamics of the multivariate system and forms the foundational basis of our proposed causal discovery module. Accordingly, throughout this work, the Granger-based graph is interpreted as a representation of directed predictive dependencies, where an edge denotes incremental predictive information from one variable's history to another rather than an interventionally identified physical relation.

4. Methodology

The proposed TFCAD framework, illustrated in Fig. 1, consists of two synergistic phases: causal discovery and anomaly detection. The first phase performs causal discovery in the time–frequency domain to capture the underlying data dependencies. In the second phase, the learned Granger causal structure is incorporated into a reconstruction-based anomaly detection model through a Causality-Guided Graph Convolutional Network (CGGCN), providing structural guidance for the reconstruction and detection processes.

4.1. Multi-scale wavelet encoder

The initial stage of our model employs a multi-scale wavelet encoder, which is designed to transform the raw input time series into a rich, multi-resolution time–frequency representation. This transformation is crucial for disentangling multi-scale representations and augmenting the architecture's capacity to resolve complex temporal dynamics. For an input window $\boldsymbol{X} \in \mathbb{R}^{W \times M}$, where W is the window size and M is the number of variables, this process unfolds through the following conceptual steps.

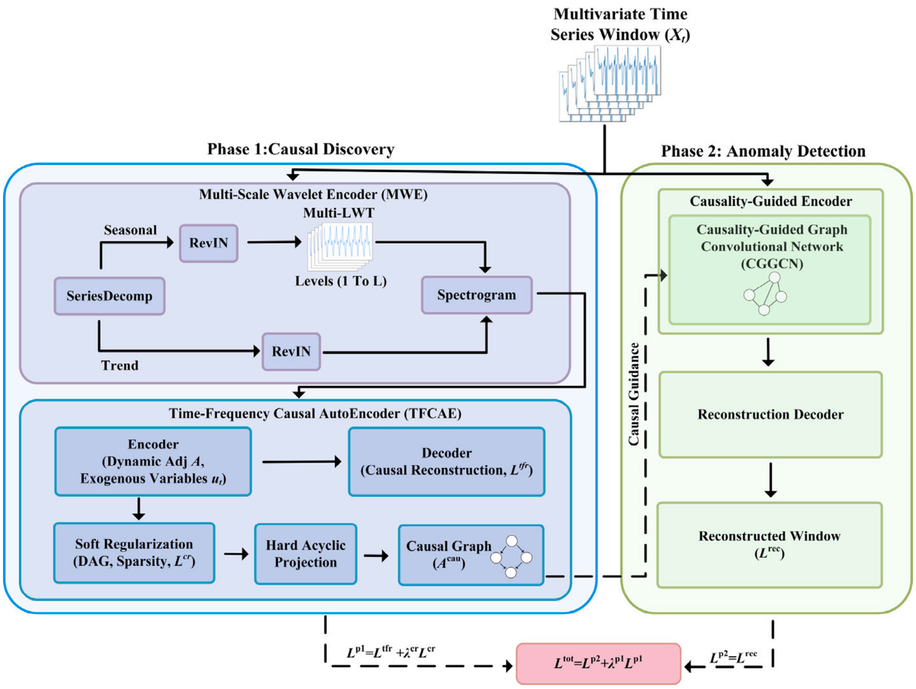


Fig. 1. The overall framework of TFCAD. The architecture consists of two main phases: (1) Causal Discovery (left, blue block), which extracts multi-scale representations via MWE and infers directed dependencies using TFCAE; and (2) Anomaly Detection (right, green block), where the discovered causal structure explicitly guides CGGCN for robust sequence reconstruction. The input data (top) flows into Phase 1 to generate a causal graph, which provides predictive structural priors for Phase 2 via the dashed "Causal Guidance" line. Non-differentiable hard acyclic projection acts as a gradient blocker, structurally decoupling the two phases and preventing downstream reconstruction gradients from propagating back into graph structure learning. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.1.1. *Time series decomposition*

The process commences by decomposing the input signal $\boldsymbol{X}$ into a low-frequency trend component, $\boldsymbol{X}^{\text{tre}}$, and a high-frequency seasonal component, $\boldsymbol{X}^{\text{sea}}$. This initial step isolates the long-term progression of the series from its periodic and more volatile fluctuations, allowing each to be processed more effectively:

$$\boldsymbol{X}^{\text{tre}}, \boldsymbol{X}^{\text{sea}} = \text{SeriesDecomp}(\boldsymbol{X}). \tag{5}$$

4.1.2. *Component-wise normalization*

Distribution shifts across different time series instances are a common challenge that can impede model generalization. To address this, we apply Reversible Instance Normalization (RevIN). This normalization is performed independently on both the trend and seasonal components. This dual-normalization strategy enhances robustness by standardizing the distinct statistical properties of both long-term and short-term dynamics, while preserving the ability to reverse the transformation for accurate final reconstruction. This step yields the normalized components $\hat{\boldsymbol{X}}^{\text{tre}}$ and $\hat{\boldsymbol{X}}^{\text{sea}}$.

4.1.3. *Multi-level wavelet analysis of seasonal component*

The subsequent multi-level wavelet decomposition is strategically applied exclusively to the normalized seasonal component, $\hat{\boldsymbol{X}}^{\text{sea}}$. This design choice focuses the analytical power of the wavelet transform on the seasonal component, which typically encapsulates the more complex, non-stationary dynamics of the signal. The analysis is realized through a cascade of L lifting-scheme-based wavelet blocks. At each level $l \in \{1, \dots, L\}$, the approximation coefficients from the previous level, $\boldsymbol{C}_{l-1}$ (with $\boldsymbol{C}_0 = \hat{\boldsymbol{X}}^{\text{sea}}$), are decomposed further. The detail coefficients $\boldsymbol{D}_l$ and the next level's approximation coefficients $\boldsymbol{C}_l$ are computed by multi-scale prediction $P(\cdot)$ and update $U(\cdot)$ operators, which are architecturally implemented as one-dimensional convolutional neural networks. To preserve temporal coherence across all frequency scales, the detail coefficients $\boldsymbol{D}_l$ produced at each level are up-sampled via linear interpolation to match the original window size W.

4.1.4. *Feature recombination and integration*

Finally, the comprehensive feature representation is constructed. The temporally-aligned detail coefficients from all levels, $\{\boldsymbol{D}_1, \boldsymbol{D}_2, \dots, \boldsymbol{D}_L\}$, are first stacked to form a multi-resolution tensor, $\boldsymbol{Z}^{\text{det}} \in \mathbb{R}^{M \times L \times W}$. To ensure the final representation retains crucial low-frequency information, the normalized trend component $\hat{\boldsymbol{X}}^{\text{tre}}$ is explicitly incorporated into the feature construction process. In practice, this component is broadcast to match the spatial dimensions of the detail-coefficient tensor and then added with a small, fixed weighting factor, which enriches the representation without overwhelming the fine-grained wavelet features. An overall scaling factor α is applied to the combined output to calibrate its magnitude across the model.

$$\boldsymbol{Z} = \alpha(\boldsymbol{Z}^{\text{det}} + \beta \cdot \hat{\boldsymbol{X}}^{\text{tre}}). \tag{6}$$

The MWE module culminates in a multi-scale tensor $\boldsymbol{Z} \in \mathbb{R}^{M\times L\times W}$ that holistically captures scale-resolved dynamics of the input window, furnishing a rich representation for subsequent tasks.

4.2. Time–frequency causal AutoEncoder

The TFCAE module is designed to learn a Granger-based directed dependency structure directly from the time–frequency representations generated by the MWE module. Operating as an encoder–decoder framework, it infers directed predictive dependencies, translating time–frequency domain characteristics into a Granger-based structural representation. Subsequently, a hybrid constraint strategy is employed to generate a sparse and acyclic directed graph, providing structural guidance for the downstream GCN.

4.2.1. Encoder–decoder framework for granger-based structure learning

The TFCAE operates as an encoder–decoder framework designed to simulate abductive and deductive reasoning, enabling the model to infer directed predictive dependencies from the time–frequency domain.

Encoder (Abductive Reasoning). The encoder's primary role is to infer the unobserved exogenous variables, $\boldsymbol{u}_t$, from the observed data, $\boldsymbol{x}_t$. Given the multi-scale tensor $\boldsymbol{Z}$ produced by the MWE, a 2D-CNN first processes the time–frequency representation of each variable to generate node embeddings. A multi-head self-attention mechanism then contextualizes these embeddings, followed by an edge prediction network that computes a dynamic, weighted adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{M\times M}$. Simultaneously, the encoder predicts the value at the current time step t based on historical context, $\hat{\boldsymbol{x}}_t^{\text{pre}}$. The exogenous variables are then inferred as the prediction residuals transformed by the learned Granger causal structure:

$$\boldsymbol{u}_t = \boldsymbol{A}(\boldsymbol{x}_t - \hat{\boldsymbol{x}}_t^{\text{pre}}). \tag{7}$$

This process follows an abductive-style mapping from the prediction error to the latent disturbance $\boldsymbol{u}_t$, which represents residual information not explained by the historical predictor.

Decoder (Deductive Reasoning). The decoder's function is to evaluate the learned Granger-based directed structure through reconstruction of the observation $\boldsymbol{x}_t$ from the inferred latent disturbance $\boldsymbol{u}_t$. A parallel 2D-CNN module generates a baseline prediction $\hat{\boldsymbol{x}}_t^{\text{bas}}$ from historical data, representing a baseline prediction from historical context before incorporating the latent adjustment $\boldsymbol{u}_t$. The inferred exogenous variables $\boldsymbol{u}_t$ are then processed by an adjustment predictor, $f_\theta(\cdot)$, which computes an adjustment term. The final reconstruction $\hat{\boldsymbol{x}}_t^{\text{rec}}$ is obtained by adding this adjustment to the baseline:

$$\hat{\boldsymbol{x}}_t^{\text{rec}} = \hat{\boldsymbol{x}}_t^{\text{bas}} + f_\theta(\boldsymbol{u}_t). \tag{8}$$

This step models deductive reasoning, using the feed-forward network $f_\theta(\cdot)$ to predict the final observation. The module is optimized by minimizing the reconstruction loss between the observation and its reconstruction:

$$\mathcal{L}^{\text{tfr}} = \|\boldsymbol{x}_t - \hat{\boldsymbol{x}}_t^{\text{rec}}\|_2^2. \tag{9}$$

4.2.2. Multi-constraint causal regularization

To encourage the learned adjacency matrix to form a sparse and acyclic Granger-based structure, we employ a hybrid strategy that integrates differentiable soft regularizations during optimization with a hard acyclicity projection applied post-inference.

Soft Constraints via Regularization. During training, we utilize two soft constraints, incorporated into a composite regularization loss term, $\mathcal{L}^{\text{cr}}$, to guide the model toward learning a desirable sparse and acyclic graph structure.

Smooth Acyclicity Constraint. Since a causal graph is inherently defined as a DAG, guaranteeing acyclicity is a fundamental requirement in learning causal relations. To guide the learned structure toward a DAG, we adopt the continuous, smooth, and exact characterization of acyclicity introduced by Zheng et al. (2018). This constraint is formulated as:

$$\mathcal{L}^{\text{dag}} = \text{tr}(e^{\boldsymbol{A}\odot\boldsymbol{A}}) - M = 0, \tag{10}$$

where $\odot$ denotes the Hadamard product, $\text{tr}(\cdot)$ is the matrix trace, M represents the number of variables (i.e., nodes in the causal graph).

Let $\tilde{\boldsymbol{A}} = \boldsymbol{A} \odot \boldsymbol{A}$ be the non-negative weighted adjacency matrix. In graph theory, the trace of the kth power of a matrix, $\text{tr}(\tilde{\boldsymbol{A}}^k)$, exactly quantifies the sum of weighted closed cycles of length k in the graph. By expanding the matrix exponential via its Taylor series, we have:

$$\text{tr}(e^{\tilde{\boldsymbol{A}}}) = \text{tr}\left(\sum_{k=0}^{\infty} \frac{\tilde{\boldsymbol{A}}^k}{k!}\right) = \text{tr}(\boldsymbol{I}) + \sum_{k=1}^{\infty} \frac{\text{tr}(\tilde{\boldsymbol{A}}^k)}{k!} = M + \sum_{k=1}^{\infty} \frac{\text{tr}(\tilde{\boldsymbol{A}}^k)}{k!}. \tag{11}$$

Since the graph contains no cycles if and only if $\mathrm{tr}(\tilde{\boldsymbol{A}}^k) = 0$ for all $k \geq 1$, the trace of the matrix exponential evaluates to exactly M if and only if the graph is strictly acyclic.

This formulation offers a distinct advantage: it provides a differentiable function. The exact gradient can be efficiently computed as $\nabla_{\boldsymbol{A}} \mathcal{L}^{\mathrm{dag}} = (e^{\boldsymbol{A} \odot \boldsymbol{A}})^T \odot 2\boldsymbol{A}$, allowing for seamless integration into the gradient-based optimization process.

Sparsity Constraint. To avoid overly dense or noisy causal structures, we apply a penalty on the L_1 norm of the adjacency matrix:

$$\mathcal{L}^{\mathrm{spa}} = \|\boldsymbol{A}\|_1. \tag{12}$$

Consequently, the total regularization loss $\mathcal{L}^{\mathrm{cr}}$ is defined as a weighted sum of these components:

$$\mathcal{L}^{\mathrm{cr}} = \lambda^{\mathrm{dag}} \mathcal{L}^{\mathrm{dag}} + \lambda^{\mathrm{spa}} \mathcal{L}^{\mathrm{spa}}. \tag{13}$$

Hard Acyclicity Projection. Although soft constraints guide the optimization toward a structurally sound graph, they cannot guarantee a strictly acyclic structure because the model must simultaneously satisfy other learning objectives, such as reconstruction error, which may push the adjacency matrix away from perfect acyclicity. To obtain a strictly acyclic DAG for guiding the graph convolutional network, a hard acyclicity projection step is applied.

This algorithm first ranks all potential edges by their absolute weights in descending order. Starting with an empty graph, it iteratively adds the edge (v_i, v_j) if and only if no path exists from v_j to v_i, thereby preventing cycle formation. This procedure serves as a heuristic approximation for the NP-hard maximum weight acyclic subgraph problem. The resulting projected adjacency matrix is denoted by $\boldsymbol{A}^{\mathrm{cau}} = \mathrm{Proj}(\boldsymbol{A})$.

The soft and hard constraints serve complementary purposes: the soft regularization shapes the adjacency matrix toward a near-DAG form, enabling the projection to preserve high-weight directed edges. Without such preconditioning, the projection would discard substantial information to maintain acyclicity. As the projection is non-differentiable, it breaks gradient flow, making the TFCAE a causal-prior generator rather than a fully end-to-end component.

4.3. Causal-guided anomaly detection

This subsection presents the main anomaly detection architecture, which adopts a graph-based encoder–decoder paradigm and integrates the learned Granger-based structural prior to guide representation learning for robust anomaly detection. The architecture is composed of a causality-guided encoder, and a reconstruction decoder.

4.3.1. Causality-guided encoder

The input $\boldsymbol{X}$ is then mapped to a latent representation $\boldsymbol{H}$ by the encoder. This encoder is instantiated by the CGGCN. Crucially, this module does not learn the graph structure from scratch; instead, it integrates the strictly acyclic Granger-based directed graph, $\boldsymbol{A}^{\mathrm{cau}}$, derived from the multi-scale causal discovery module (Section 4.2). By using $\boldsymbol{A}^{\mathrm{cau}}$ as a structural prior, the encoder is constrained to aggregate information primarily along directed predictive dependencies in the Granger sense. This guidance reduces reliance on purely correlation-based connections and allows the learned latent variables to encode the inter-variable dependency patterns represented by the structural prior.

4.3.2. Reconstruction decoder

The decoder mirrors the encoder structure, mapping the causality-guided latent representation $\boldsymbol{H}$ back to the original input dimension to produce the reconstructed output $\hat{\boldsymbol{x}}_t$. The network is trained by minimizing the reconstruction loss $\mathcal{L}^{\mathrm{rec}}$. By enforcing reconstruction based on structure-guided features, the model learns a high-fidelity representation of normal behavior. Anomalies are then detected by identifying time points with large reconstruction errors, indicating deviations that violate the learned structural and temporal patterns.

4.4. Training objective and anomaly score

This subsection details the formal objective functions used to train the proposed model and describes the mechanism for computing the final anomaly score during inference.

4.4.1. Training objective

The model comprises two functionally distinct components: causal discovery and anomaly detection. The overall training objective aggregates two primary loss terms corresponding to the core learning phases of the model.

Phase 1: Causal Discovery and Causal Reconstruction Loss. This component is dedicated to guiding the causal discovery within the TFCAE module. It is formulated as the sum of the reconstruction loss of the TFCAE decoder ($\mathcal{L}^{\mathrm{tfr}}$) and a multi-constraint causal regularization term ($\mathcal{L}^{\mathrm{cr}}$), which is critical for ensuring the learned graph is a DAG:

$$\mathcal{L}^{\mathrm{p1}} = \mathcal{L}^{\mathrm{tfr}} + \lambda^{\mathrm{cr}} \mathcal{L}^{\mathrm{cr}}, \tag{14}$$

where λ^{cr} is a hyperparameter balancing the importance of the causal structure.

Phase 2: Anomaly Detection Loss. This objective drives the optimization of the main anomaly detection autoencoder to ensure robust data reconstruction.

The loss function is based on the standard reconstruction loss ($\mathcal{L}^{\mathrm{rec}}$), which measures the pixel-wise discrepancy between the input and the reconstructed output.

The composite loss function for Phase 2 is thus defined as:

$$\mathcal{L}^{\mathrm{p2}} = \mathcal{L}^{\mathrm{rec}}. \tag{15}$$

The final loss function is a weighted combination of these two objectives:

$$\mathcal{L}^{\mathrm{tot}} = \mathcal{L}^{\mathrm{p2}} + \lambda^{\mathrm{p1}} \mathcal{L}^{\mathrm{p1}}, \tag{16}$$

where λ^{p1} is a hyperparameter balancing the influence of the causal guidance and the core detection objective. It is important to note that although both objectives are included in the overall training objective, their gradient paths are structurally decoupled. Specifically, the introduction of the non-differentiable greedy acyclic projection in the TFCAE module acts as a gradient blocker and therefore blocks gradient propagation from the anomaly detection loss to the TFCAE module. This design ensures that causal discovery is driven by intrinsic time–frequency characteristics of the data rather than being influenced by reconstruction objectives, while still allowing the learned Granger causal structure to provide dynamic guidance to the anomaly detection encoder.

After training on normal data, the model is used for anomaly inference. Given a test observation $\boldsymbol{x}_t$, the anomaly score is computed as the squared L_2 norm of the reconstruction error produced by the anomaly detection autoencoder:

$$\mathcal{S}_t = \|\boldsymbol{x}_t - \hat{\boldsymbol{x}}_t\|_2^2, \tag{17}$$

where $\hat{\boldsymbol{x}}_t$ denotes the reconstructed output. A data point is flagged as anomalous if its score exceeds a threshold τ.

4.5. Complexity analysis

Let T denote the raw sequence length, M the number of variables, W the sliding-window length, s the stride, B the batch size, E the number of training epochs, and L the number of lifting-wavelet levels. The number of windows is $N_w = \left\lfloor \frac{T-W}{s} \right\rfloor + 1$. For one training batch, the overall cost is composed of four parts: multi-scale spectral encoding, graph message passing, causal DAG regularization, and reconstruction projection. Accordingly, the per-batch time complexity is $O\big(B\big(MLW + M^2\ W + M^3 + (MW)^2\big)\big)$. In practice, the dominant term is typically $O(B(MW)^2)$, while $O(BM^3)$ is the second-largest term due to the matrix-exponential-based acyclicity constraint. Therefore, the per-epoch training complexity is $O\big(N_w\big(MLW + M^2\ W + M^3 + (MW)^2\big)\big)$, and the full training complexity over E epochs is $O\big(E\,N_w\big(MLW + M^2\ W + M^3 + (MW)^2\big)\big)$. During inference, the causal-regularization branch is not evaluated, so the complexity is reduced to $O\big(N_w\big(M^2\ W + (MW)^2\big)\big)$.

For memory usage, the dominant activation terms are from multi-scale features and graph relation tensors, yielding $O\big(B(MLW + M^2 d_c + MW)\big)$, where d_c is the causal embedding dimension. The parameter complexity is mainly determined by large projection layers and is approximately $O\big((MW)^2\big)$.

4.6. Theoretical analysis of multi-task gradient conflicts

Rather than jointly optimizing all modules under both the causal discovery and anomaly reconstruction objectives, TFCAD adopts a structurally decoupled optimization strategy. In a fully end-to-end (E2E) setting, both objectives simultaneously affect the continuous adjacency matrix $\boldsymbol{A}$ learned by the causal module. Since the two objectives optimize different functional targets, their gradients with respect to $\boldsymbol{A}$ may exhibit inconsistent update directions, introducing interference into the learning of the graph structure. The gradient-decoupling mechanism in TFCAD prevents the downstream reconstruction objective from directly perturbing the causal structure learning process, while the hard acyclic projection further ensures that the structural prior delivered to the anomaly detection module satisfies the required acyclicity constraint.

Let $\boldsymbol{A} \in \mathbb{R}^{M\times M}$ denote the continuous adjacency matrix learned by the causal module. Under E2E coupling, the training objective can be written explicitly in terms of its dependence on the continuous adjacency matrix as a combination of the causal discovery loss (Phase 1, $\mathcal{L}^{\mathrm{p1}}$) and the downstream anomaly reconstruction loss ($\mathcal{L}^{\mathrm{rec}}$) driven by the CGGCN. The total objective is:

$$\mathcal{L}^{\mathrm{tot}} = \lambda^{\mathrm{p1}} \mathcal{L}^{\mathrm{p1}}(\boldsymbol{A}) + \mathcal{L}^{\mathrm{rec}}(\boldsymbol{A}, \boldsymbol{X}) \tag{18}$$

where λ^{p1} is the balancing hyperparameter. Accordingly, the gradient acting on the shared adjacency matrix is:

$$\nabla_{\boldsymbol{A}} \mathcal{L}^{\mathrm{tot}} = \lambda^{\mathrm{p1}} \nabla_{\boldsymbol{A}} \mathcal{L}^{\mathrm{p1}} + \nabla_{\boldsymbol{A}} \mathcal{L}^{\mathrm{rec}} \tag{19}$$

The potential interference between the two objectives can be characterized by their gradient directions in the adjacency space. Specifically, we denote the gradients of the Phase-1 objective and the reconstruction objective with respect to the continuous adjacency matrix as

$$\begin{aligned} \boldsymbol{g}^{\mathrm{p1}} &= \nabla_{\boldsymbol{A}} \mathcal{L}^{\mathrm{p1}}, \\ \boldsymbol{g}^{\mathrm{rec}} &= \nabla_{\boldsymbol{A}} \mathcal{L}^{\mathrm{rec}}. \end{aligned} \tag{20}$$

Their directional relationship can be measured by the cosine similarity

$$\rho = \frac{\langle \boldsymbol{g}^{\mathrm{p1}}, \boldsymbol{g}^{\mathrm{rec}} \rangle}{\|\boldsymbol{g}^{\mathrm{p1}}\|_2\, \|\boldsymbol{g}^{\mathrm{rec}}\|_2}. \tag{21}$$

Table 2
Summary of the six benchmark datasets used for evaluation. The table details the training and testing set sizes, the number of feature dimensions, and the overall anomaly ratio for each dataset. These datasets encompass diverse real-world domains (e.g., aerospace, server monitoring, and industrial control), providing a comprehensive testbed for evaluating the robustness of the proposed framework.

Dataset name	Train set size	Test set size	Dimension number	Anomaly ratio (%)
ASD	102 331	51 840	19	4.61
MSL	58 317	73 729	55	10.72
PSM	132 481	87 841	25	27.80
SMAP	135 183	427 617	25	13.13
SMD	708 405	748 420	38	4.16
SWaT	475 200	449 919	51	11.98

When $\rho < 0$, the angle between the two gradients exceeds 90°, indicating that the two objectives impose conflicting update directions on the shared adjacency matrix. In this case, the reconstruction gradient contains a component opposing the Phase-1 gradient, such that the descent updates induced by the two objectives partially counteract each other, thereby introducing interference into graph structure optimization. Such interference arises because the E2E formulation requires the same adjacency matrix to simultaneously support structural learning and downstream reconstruction, while the optimization directions associated with these two objectives are not necessarily aligned.

To avoid such cross-objective interference, TFCAD decouples the downstream reconstruction optimization from graph structure learning. Specifically, the projected graph $\boldsymbol{A}^{\text{cau}}$ is used as a gradient-isolated structural prior for the CGGCN, and the stop-gradient operator, denoted by $\mathrm{sg}(\cdot)$, is applied before graph-based feature aggregation:

$$\boldsymbol{H} = \mathrm{sg}\left(\boldsymbol{A}^{\text{cau}}\right) \boldsymbol{X}^{\mathrm{T}} \boldsymbol{\Theta}. \tag{22}$$

Consequently, the downstream reconstruction loss does not propagate gradients back to the continuous adjacency matrix through the structural-prior path, i.e.,

$$\nabla_{\boldsymbol{A}} \mathcal{L}^{\text{rec}} = 0. \tag{23}$$

The continuous adjacency matrix is therefore learned solely under the Phase-1 objective, while the parameters of the CGGCN are optimized for reconstruction conditional on the projected structural prior. The two operations serve complementary purposes: gradient blocking isolates graph structure learning from downstream reconstruction updates, whereas the hard projection guarantees the strict acyclicity of the structural prior supplied to the anomaly detection module.

5. Experimental evaluation

5.1. Experimental settings

5.1.1. Datasets

We employ six datasets in the experimental phase: Application Server Dataset (ASD) (Li et al., 2021), Mars Science Laboratory Rover (MSL) (Hundman et al., 2018), Pooled Server Metrics (PSM) (Abdulaal et al., 2021), Server Machine Dataset (SMD) (Su et al., 2019), Soil Moisture Active Passive satellite (SMAP) (Hundman et al., 2018), and Secure Water Treatment (SWaT) (Mathur & Tippenhauer, 2016). A summary of their detailed features is provided in Table 2.

- ASD is collected from a prominent internet provider; this set is used to track operational telemetry from 12 server units over 45 days. It is comprised of 19 performance dimensions, with ground truth provided by system specialists based on incident reports.
- MSL is sourced from NASA and contains science laboratory telemetry from the Curiosity rover. It involves 55 signals from 27 distinct components, with anomalies expert-verified by aerospace engineers.
- PSM is sourced from eBay's cloud infrastructure; this dataset is used to capture 25 performance variables from a single unified entity. It is widely utilized to evaluate detection models in high-traffic server environments.
- SMAP is focused on environmental monitoring and is sourced from NASA. Moisture levels are tracked via 25 distinct features across 55 entities, and it serves as a benchmark for the detection of transient satellite anomalies.
- SMD is maintained as a 5-week record to monitor 38 metrics across 28 independent compute nodes. The sequences are split into identical-sized training and testing sets, which are sampled at 1-minute intervals.
- SWaT is obtained from a water purification testbed. Industrial control system behaviors are modeled under 51 sensor/actuator points, and various simulated cyber-attack scenarios are incorporated for security analysis.

5.1.2. Baselines

To comprehensively evaluate the performance of TFCAD, we conduct a comparative analysis against thirteen baseline models. These baselines span diverse research lines, ranging from classical statistical time series models (PCA (Darban et al., 2025), ARIMA (Schmidl et al., 2022)) and classic deep reconstruction and forecasting models (LSTM (Malhotra et al., 2015), VAE (Xu et al., 2018), LSTM-VAE (Park et al., 2018), AMFormer (Zhong et al., 2024)), to adversarial and contrastive learning frameworks (USAD (Audibert et al., 2020), DADA (Shentu et al., 2025), COCA (Wang et al., 2023), MGCLAD (Qin et al., 2023)), as well as structure-aware approaches leveraging graphs, causality, or frequency analysis (MTGFLOW (Zhou et al., 2024), AERCA (Han et al., 2025), CATCH (Wu et al., 2025)). Specifically, since ARIMA is inherently a univariate model, we independently build a model for each dimension of the multivariate time series and aggregate the anomaly scores (e.g., by taking the maximum or average prediction error) to determine whether the overall system is anomalous.

5.1.3. Metrics

The performance of all methods is evaluated using standard evaluation metrics: Precision (P), Recall (R), F1-score (F1), and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). To ensure statistically rigorous comparisons, all experiments are repeated using three distinct random seeds, and the results are reported as "mean ± standard deviation".

5.1.4. Implementation details

To ensure a fair and rigorous comparison, we adopt a standardized data preprocessing and evaluation protocol across all models. Specifically, the Spectral Residual (SR) transform is applied independently to each channel to enhance temporal saliency, with missing values replaced by zeros. All features are then Min–Max normalized to $[0, 1]$, chronologically split into training (70%) and validation (30%) sets, and segmented into input sequences using a sliding window with a step size of 1. For training configurations, unless otherwise specified in the original publications or official implementations, models are trained for up to 250 epochs with a batch size of 128 using the Adam optimizer (Kingma & Ba, 2015) with an initial learning rate of 10^{-3}. Regarding model-specific configurations, we strictly follow the architectural designs and hyperparameter settings recommended in the original publications or official implementations. By combining this unified experimental pipeline with faithful reproduction of the recommended configurations, we ensure that each baseline is evaluated under uniform conditions while maintaining consistency with the original experimental settings.

5.1.5. Hyperparameter settings

Regarding the model hyperparameters, our Multi-scale Wavelet Encoder builds upon the decomposition principles of AdaWaveNet (Yu et al., 2024). However, while AdaWaveNet simply sums the separated components at the output layer, our MWE module integrates them at the latent feature level to construct a comprehensive representation for downstream causal discovery. To prevent the large-magnitude low-frequency trend component from overwhelming the fine-grained high-frequency wavelet details during this fusion, we empirically set the scaling factors to $\alpha = 1$ and $\beta = 0.1$. For the causal regularization parameters, we follow the standard practices established in continuous DAG learning literature (Yu et al., 2019; Zheng et al., 2018). Specifically, the weights for the smooth acyclicity constraint (λ^{dag}), the sparsity constraint (λ^{spa}), and the overall causal regularization (λ^{cr}) are all uniformly set to 1. Empirical observations (as validated in Section 5.7) confirm that this configuration consistently yields strictly acyclic DAG structures across all benchmarks without the need for dataset-specific tuning. Finally, the trade-off parameter λ^{p1}, which balances the causal discovery and anomaly detection objectives, is set to 1 by default; its specific impact is further analyzed in Section 5.3.

5.2. Comparison experiments

Table 3 tabulates the comparative performance of the proposed method against the baselines, where we utilize bold font to indicate the best results and underlining for the second-best values.

As indicated by the results in Table 3, TFCAD exhibits robust performance superiority, outperforming other methods across all six benchmark datasets. Specifically, on the challenging SMD dataset, our model achieves a remarkable F1-score of 0.8852, verifying its effectiveness in complex multivariate server monitoring scenarios. TFCAD secures the highest AUC-ROC values across all datasets (e.g., 0.8759 on ASD and 0.8509 on SMD), demonstrating its exceptional capability in distinguishing anomalous events from normal patterns. Beyond the aggregate metrics, it is worth noting that while some baselines achieve exceptionally high Precision or Recall on specific tasks (e.g., USAD achieving a Recall of 1.0000 on MSL but yielding a very low F1-score of 0.2235), they often fail to maintain a reasonable trade-off, resulting in suboptimal F1-scores and lower AUC-ROC. In contrast, TFCAD avoids such bias by effectively harmonizing Precision and Recall, ensuring comprehensive and reliable detection capabilities.

Notably, on the PSM dataset, the F1-score gain over the second-best baseline (AMFormer) is relatively marginal (a 1.0% improvement). The fundamental reason for this variance is twofold. First, this dataset is dominated by relatively simple, isolated point anomalies, causing advanced baselines to approach an F1 performance ceiling (e.g., F1 > 0.95), leaving limited room for substantial gains. Second, the Granger-based structural prior in TFCAD is designed to encode directed predictive dependencies during GCN aggregation. Its contribution is therefore less pronounced in the observed isolated-anomaly setting of PSM, whereas larger gains are observed on ASD and SMD, where TFCAD outperforms the runner-up by 4.4% and 3.8% in F1, respectively. These results are consistent with the role of the learned structural prior in providing informative inter-variable context for downstream representation learning.

Table 3
Overall anomaly detection performance comparison on six benchmark datasets (mean ± std). Bold and underlined values indicate the best and second-best results, respectively. Proposed TFCAD outperforms all baseline methods, highlighting the superior discriminative capability achieved by integrating time–frequency analysis with dynamic causal guidance.

Methods	ASD				MSL			
	P	R	F1	AUC-ROC	P	R	F1	AUC-ROC
ARIMA	0.7086 ± 0.0000	0.5400 ± 0.0000	0.5645 ± 0.0000	0.5708 ± 0.0000	0.8215 ± 0.0000	0.9622 ± 0.0000	0.8556 ± 0.0000	0.5627 ± 0.0000
PCA	<u>0.8820</u> ± 0.0000	0.8578 ± 0.0000	0.8596 ± 0.0000	0.7569 ± 0.0000	0.8260 ± 0.0000	0.9630 ± 0.0000	0.8670 ± 0.0000	0.5828 ± 0.0000
VAE	0.6329 ± 0.0083	0.8679 ± 0.0010	0.6923 ± 0.0034	0.4974 ± 0.0001	<u>0.8532</u> ± 0.0086	0.9579 ± 0.0000	0.8714 ± 0.0069	0.5248 ± 0.0003
LSTM	0.7789 ± 0.0260	0.8829 ± 0.0057	0.8115 ± 0.0192	0.7958 ± 0.0068	0.8515 ± 0.0042	0.9579 ± 0.0000	0.8697 ± 0.0030	0.5248 ± 0.0002
LSTM-VAE	0.2142 ± 0.0218	0.5759 ± 0.0466	0.2795 ± 0.0166	0.7138 ± 0.0062	0.6377 ± 0.0018	0.9157 ± 0.0123	0.6774 ± 0.0025	0.6029 ± 0.0100
AMFormer	0.6458 ± 0.0203	0.6381 ± 0.0264	0.6046 ± 0.0325	0.5745 ± 0.0276	0.8490 ± 0.0161	<u>0.9903</u> ± 0.0137	<u>0.8821</u> ± 0.0234	0.5454 ± 0.0180
DADA	0.6164 ± 0.0126	0.6975 ± 0.0850	0.6135 ± 0.0291	0.6793 ± 0.0006	0.8151 ± 0.0020	0.9198 ± 0.0051	0.8246 ± 0.0028	0.6125 ± 0.0147
COCA	0.3677 ± 0.0049	0.5315 ± 0.0324	0.3968 ± 0.0018	0.7240 ± 0.0034	0.4883 ± 0.0250	0.9610 ± 0.0292	0.5824 ± 0.0156	0.5230 ± 0.0120
USAD	0.5785 ± 0.0249	0.8010 ± 0.0477	0.6289 ± 0.0052	0.4760 ± 0.0006	0.1387 ± 0.0000	**1.0000** ± 0.0000	0.2235 ± 0.0000	0.4415 ± 0.0039
MTGFLOW	0.8660 ± 0.0485	<u>0.9354</u> ± 0.0037	<u>0.8928</u> ± 0.0228	<u>0.8401</u> ± 0.0328	0.7886 ± 0.0291	0.9713 ± 0.0000	0.8347 ± 0.0206	<u>0.6154</u> ± 0.0134
MGCLAD	0.6462 ± 0.0177	0.8978 ± 0.0019	0.7265 ± 0.0112	0.5439 ± 0.0103	0.8516 ± 0.0074	0.9498 ± 0.0115	0.8642 ± 0.0018	0.6026 ± 0.0028
AERCA	0.8228 ± 0.0036	0.9295 ± 0.0000	0.8617 ± 0.0035	0.5070 ± 0.0021	0.8086 ± 0.0150	0.9816 ± 0.0064	0.8650 ± 0.0168	0.5243 ± 0.0102
CATCH	0.8531 ± 0.0029	0.8811 ± 0.0065	0.8547 ± 0.0076	0.6384 ± 0.0072	0.8373 ± 0.0153	0.9382 ± 0.0115	0.8379 ± 0.0078	0.4556 ± 0.0013
TFCAD	**0.9395** ± 0.0027	**0.9365** ± 0.0060	**0.9325** ± 0.0030	**0.8759** ± 0.0031	**0.8628** ± 0.0108	0.9835 ± 0.0099	**0.8973** ± 0.0041	**0.6327** ± 0.0001

Methods	PSM				SMAP			
	P	R	F1	AUC-ROC	P	R	F1	AUC-ROC
ARIMA	0.2790 ± 0.0000	**0.9999** ± 0.0000	0.4363 ± 0.0000	0.6403 ± 0.0000	0.6759 ± 0.0000	**1.0000** ± 0.0000	0.7513 ± 0.0000	0.4938 ± 0.0000
PCA	0.2790 ± 0.0000	<u>0.9998</u> ± 0.0000	0.4363 ± 0.0000	0.7151 ± 0.0000	0.5938 ± 0.0000	0.7649 ± 0.0000	0.6233 ± 0.0000	0.4589 ± 0.0000
VAE	0.9485 ± 0.0012	0.8969 ± 0.0000	0.9220 ± 0.0006	0.6772 ± 0.0027	0.6757 ± 0.0027	0.9827 ± 0.0012	0.7362 ± 0.0045	0.5336 ± 0.0008
LSTM	0.9804 ± 0.0044	0.8381 ± 0.0037	0.9037 ± 0.0040	0.6537 ± 0.0593	0.6885 ± 0.0040	0.9632 ± 0.0051	0.7259 ± 0.0022	0.5259 ± 0.0003
LSTM-VAE	0.9130 ± 0.0331	0.8512 ± 0.0313	0.8810 ± 0.0322	<u>0.7243</u> ± 0.0168	0.3934 ± 0.0250	0.8909 ± 0.0167	0.4413 ± 0.0203	0.5195 ± 0.0074
AMFormer	0.9782 ± 0.0156	0.9409 ± 0.0004	<u>0.9592</u> ± 0.0073	0.6807 ± 0.0112	0.6452 ± 0.0016	0.9715 ± 0.0221	0.6940 ± 0.0091	<u>0.5392</u> ± 0.0077
DADA	**0.9993** ± 0.0004	0.8213 ± 0.0861	0.9003 ± 0.0520	0.6950 ± 0.0215	**0.7249** ± 0.0195	0.9354 ± 0.0134	0.7455 ± 0.0111	0.5389 ± 0.0244
COCA	0.8508 ± 0.2111	0.8505 ± 0.1401	0.8325 ± 0.0361	0.6082 ± 0.0287	0.5403 ± 0.0013	0.9719 ± 0.0000	0.6036 ± 0.0006	0.5006 ± 0.0047
USAD	0.9359 ± 0.0045	0.9053 ± 0.0118	0.9204 ± 0.0083	0.6364 ± 0.0153	0.3081 ± 0.0024	<u>0.9927</u> ± 0.0000	0.3757 ± 0.0028	0.5312 ± 0.0008
MTGFLOW	0.9357 ± 0.0466	0.9733 ± 0.0315	0.9533 ± 0.0091	0.6216 ± 0.0153	0.6357 ± 0.0137	0.9563 ± 0.0093	0.6913 ± 0.0206	0.5156 ± 0.0309
MGCLAD	0.9553 ± 0.0631	0.7907 ± 0.1058	0.8610 ± 0.0375	0.5252 ± 0.0150	0.6668 ± 0.0155	0.9698 ± 0.0000	0.7098 ± 0.0134	0.5232 ± 0.0043
AERCA	0.9693 ± 0.0017	0.8925 ± 0.0000	0.9294 ± 0.0008	0.6009 ± 0.0038	0.7010 ± 0.0030	0.9858 ± 0.0134	0.7570 ± 0.0005	0.5264 ± 0.0003
CATCH	<u>0.9926</u> ± 0.0003	0.8943 ± 0.0000	0.9409 ± 0.0001	0.7006 ± 0.0017	0.6879 ± 0.0028	0.9884 ± 0.0043	<u>0.7608</u> ± 0.0092	0.4922 ± 0.0001
TFCAD	0.9559 ± 0.0018	0.9827 ± 0.0001	**0.9691** ± 0.0009	**0.7346** ± 0.0036	<u>0.7173</u> ± 0.0023	0.9798 ± 0.0044	**0.7801** ± 0.0006	**0.5805** ± 0.0044

Methods	SMD				SWaT			
	P	R	F1	AUC-ROC	P	R	F1	AUC-ROC
ARIMA	0.7957 ± 0.0000	0.7445 ± 0.0000	0.7596 ± 0.0000	0.6307 ± 0.0000	0.9613 ± 0.0000	0.7567 ± 0.0000	0.8468 ± 0.0000	<u>0.6471</u> ± 0.0000
PCA	0.8427 ± 0.0000	**0.9098** ± 0.0000	<u>0.8528</u> ± 0.0000	0.6919 ± 0.0000	0.5122 ± 0.0000	0.8962 ± 0.0000	0.6518 ± 0.0000	0.2129 ± 0.0000
VAE	0.7111 ± 0.0134	0.6406 ± 0.0024	0.6349 ± 0.0049	0.7071 ± 0.0007	0.9639 ± 0.0016	0.7519 ± 0.0000	0.8448 ± 0.0006	0.5989 ± 0.0038
LSTM	0.7825 ± 0.0260	0.6577 ± 0.0323	0.6786 ± 0.0115	0.7524 ± 0.0071	0.3404 ± 0.0588	0.8874 ± 0.0404	0.4897 ± 0.0553	0.2046 ± 0.0016
LSTM-VAE	0.7549 ± 0.0070	0.6578 ± 0.0158	0.6521 ± 0.0124	0.7913 ± 0.0145	0.8385 ± 0.0723	0.8120 ± 0.0272	0.8236 ± 0.0211	0.2393 ± 0.0048
AMFormer	0.8543 ± 0.0279	0.8010 ± 0.0083	0.8052 ± 0.0256	0.6934 ± 0.0046	0.3812 ± 0.0012	0.7586 ± 0.0000	0.5074 ± 0.0010	0.3364 ± 0.0030
DADA	<u>0.9367</u> ± 0.0083	0.6354 ± 0.0050	0.6842 ± 0.0011	0.8090 ± 0.0050	**0.9928** ± 0.0003	0.7118 ± 0.0000	0.8292 ± 0.0001	0.6161 ± 0.0030
COCA	0.5360 ± 0.0683	0.6319 ± 0.0137	0.4489 ± 0.0312	0.5188 ± 0.0059	0.9592 ± 0.0117	0.7427 ± 0.0391	0.8367 ± 0.0204	0.6067 ± 0.0040
USAD	0.2245 ± 0.0029	0.6064 ± 0.0023	0.2360 ± 0.0003	0.3933 ± 0.0001	0.9630 ± 0.0505	0.6618 ± 0.0238	0.7838 ± 0.0001	0.6104 ± 0.0056
MTGFLOW	0.8375 ± 0.0179	0.8114 ± 0.0175	0.7980 ± 0.0272	0.7468 ± 0.0971	0.9690 ± 0.0292	0.6944 ± 0.0324	0.8090 ± 0.0321	0.6269 ± 0.0107
MGCLAD	0.7799 ± 0.0009	0.6434 ± 0.0390	0.6636 ± 0.0270	0.7463 ± 0.0082	0.6820 ± 0.3359	0.8280 ± 0.0693	0.7167 ± 0.1793	0.2259 ± 0.0017
AERCA	0.8118 ± 0.0662	0.7104 ± 0.0427	0.7146 ± 0.0174	0.6328 ± 0.0018	0.2315 ± 0.0066	<u>0.8977</u> ± 0.0096	0.3680 ± 0.0076	0.2141 ± 0.0054
CATCH	**0.9439** ± 0.0006	0.7449 ± 0.0071	0.7852 ± 0.0038	<u>0.8218</u> ± 0.0006	<u>0.9848</u> ± 0.0098	0.8385 ± 0.0029	<u>0.9058</u> ± 0.0025	0.2270 ± 0.0003
TFCAD	0.8982 ± 0.0036	<u>0.8861</u> ± 0.0086	**0.8852** ± 0.0050	**0.8509** ± 0.0009	0.9315 ± 0.0199	**0.9006** ± 0.0131	**0.9156** ± 0.0028	**0.6792** ± 0.0060

To evaluate the statistical significance of our results, we perform the Nemenyi test based on F1-scores at a 0.05 significance level (Critical Difference, CD = 8.1). As illustrated in Fig. 2, TFCAD achieves the highest average rank (1.0000), statistically outperforming several competitive baselines, including LSTM-VAE (11.1671), USAD (11.1671), and COCA (11.3329). These results clearly demonstrate the superior performance and stability of our model.

This superior performance is fundamentally rooted in TFCAD's dual focus on multi-scale feature extraction and Granger-based structure learning. The lifting-based wavelet transform plays a crucial role in capturing complex, multi-scale temporal patterns and transient features in the time–frequency domain, offering a more discriminative representation of the time series compared to methods that operate solely in the time domain or rely on fixed transformations. More importantly, the causality-guided convolution module enables TFCAD to identify and leverage Granger-based directed predictive dependencies between variables. By incorporating these directional dependencies rather than relying solely on symmetric pairwise correlations, TFCAD obtains a more informative structural representation of normal multivariate behavior. This, in turn, enhances the reconstruction process, resulting in improved sensitivity to anomalous events and more effective anomaly discrimination.

5.3. Parameter sensitivity

Our experimental analysis scrutinizes the role of essential hyperparameters, specifically isolating the effects of window size, causal guidance frequency, and the Phase 1 loss weight on the final results.

Window Size. To rigorously evaluate the model's capacity to capture temporal contexts, we analyze the impact of standard large window sizes ranging from 32 to 128, with results illustrated in Fig. 3(a). The optimal window size exhibits dataset-specific characteristics, closely tied to the nature of the underlying anomalies. For datasets characterized by transient, short-lived anomalies (such as MSL, PSM, SMD, and SWaT), the model achieves its peak F1-score at a smaller window size of 32. As the window expands, the performance gradually declines due to the "dilution effect", where excessive normal patterns smooth out short anomalous spikes. Conversely, datasets characterized by complex contextual dependencies or strong periodic trends, such as ASD and SMAP, benefit from broader temporal receptive fields, achieving their peak performance at window sizes of 96 and 128, respectively. Crucially, despite the natural dilution effect in extended windows, the overall performance degradation remains notably slow and stable across

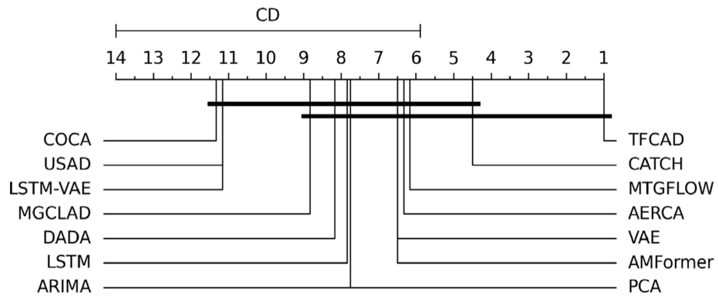
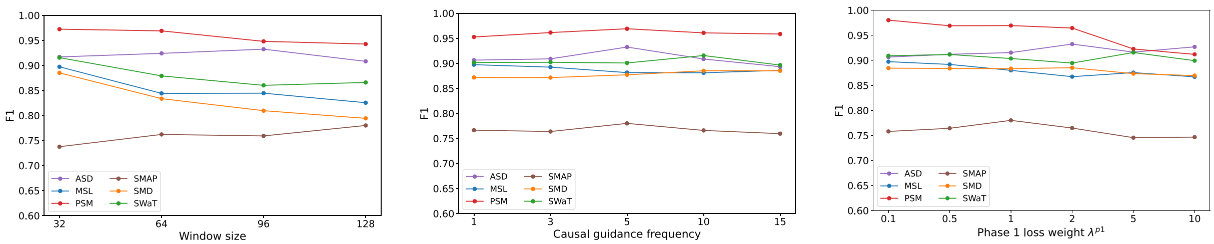


Fig. 2. Statistical significance evaluation using the Nemenyi test based on F1-scores. The horizontal axis represents the average rank, while models connected by the same thick line show no statistically significant difference. TFCAD achieves the best average rank and statistically outperforms multiple competitive baselines, confirming its robust performance.



(a) Window size vs. performance. (b) Causal guidance frequency vs. performance. (c) Phase 1 loss weight vs. performance.

Fig. 3. Parameter sensitivity analysis evaluating the impact of three key hyperparameters on the model's F1-score across diverse datasets: (a) Window size; (b) Causal guidance frequency; and (c) Phase 1 loss weight. The relatively flat curves indicate that TFCAD maintains robust and stable detection across a wide range of configurations, minimizing the need for exhaustive hyperparameter tuning.

all datasets. We attribute this robustness directly to the Multi-scale Wavelet Encoder. By leveraging the multi-scale lifting scheme, MWE effectively disentangles high-frequency anomalous details from large-magnitude, low-frequency trends, allowing TFCAD to maintain sensitivity to local anomalies even within prolonged observation windows.

Causal Guidance Frequency. This parameter determines the batch interval at which the causal graph discovered by TFCAE is used to update the GCN during training. As illustrated in Fig. 3(b), the model remains stable across a wide range of guidance frequencies, reflecting strong robustness overall. A closer examination of the trends shows that datasets such as MSL, SMD, and SWaT display relatively flat curves with only minor fluctuations, whereas ASD, PSM, and SMAP exhibit more pronounced improvements when the guidance frequency is set to 5. This pattern reflects a fundamental trade-off: if the guidance is too infrequent, the GCN may rely on outdated causal structures; if it is applied too frequently, the rapid updates may disrupt the GCN's internal learning dynamics. Setting the guidance frequency to 5 provides an effective balance, enabling the model to incorporate evolving causal information while maintaining stable and efficient learning.

Phase 1 Loss Weight (λ^{p1}). The hyperparameter λ^{p1} acts as a trade-off parameter that governs the balance between the causal discovery objective (Phase 1) and the primary anomaly detection reconstruction (Phase 2). As depicted in Fig. 3(c), we evaluate its impact across the range $\{0.1, 0.5, 1, 2, 5, 10\}$. When λ^{p1} is exceptionally small (e.g., 0.1), datasets like SMAP suffer from insufficient structural guidance, preventing GCN from effectively modeling inter-variable dependencies. Conversely, an excessively large λ^{p1} (e.g., 5 or 10) overly penalizes the model, overshadowing the primary reconstruction task and causing severe performance drops in datasets like PSM and SMD. Overall, the model demonstrates strong robustness and consistency when $\lambda^{\mathrm{p1}} \in [0.5, 2]$. Within this stable optimal range, $\lambda^{\mathrm{p1}} = 1$ serves as a robust, dataset-agnostic default configuration.

5.4. Ablation experiments

To assess the individual contributions of the major components in TFCAD, we construct five model variants by selectively removing or replacing specific modules.

***w/o* MWE & TFCAE**. This variant removes both the Multi-scale Wavelet Encoder and Time–Frequency Causal AutoEncoder modules, allowing the GCN to operate directly on raw time-series inputs; this setting isolates the overall benefit brought by the time–frequency feature extraction and the causal guidance.

***w/o* MWE & *w/* MLP**. This variant substitutes the Multi-scale Wavelet Encoder with a pure MLP-based feature extractor of comparable parameter size. This confirms the improvements are driven by the time–frequency decomposition itself, not just an increase in model capacity.

Table 4
Ablation study evaluating the effectiveness of key model components (mean ± std). Bold and underlined values indicate the best and second-best results. The consistent superiority of the full TFCAD framework supports the contributions of both multi-scale time–frequency extraction and learned structural guidance to anomaly detection performance.

Methods	ASD			MSL		
	P	R	F1	P	R	F1
w/o MWE & TFCAE	0.9017 ± 0.0099	0.6344 ± 0.0000	0.7237 ± 0.0041	0.7809 ± 0.0033	0.9656 ± 0.0000	0.8268 ± 0.0020
w/o MWE & *w/* MLP	0.9258 ± 0.0072	0.8317 ± 0.0075	0.8667 ± 0.0002	0.8270 ± 0.0010	0.9524 ± 0.0000	0.8515 ± 0.0010
w/o CG & *w/* Rand	0.6968 ± 0.0021	0.5964 ± 0.0000	0.5489 ± 0.0008	0.7151 ± 0.0281	0.9275 ± 0.0048	0.7391 ± 0.0178
w/o CG & *w/* Corr	0.8926 ± 0.0051	0.8264 ± 0.0000	0.8483 ± 0.0025	0.8245 ± 0.0011	0.9524 ± 0.0000	0.8490 ± 0.0006
w/ E2E	0.8703 ± 0.0125	0.9189 ± 0.0103	0.8866 ± 0.0004	0.8148 ± 0.0041	0.9792 ± 0.0000	0.8550 ± 0.0018
TFCAD	**0.9395** ± 0.0027	**0.9365** ± 0.0060	**0.9325** ± 0.0030	**0.8628** ± 0.0108	**0.9835** ± 0.0099	**0.8973** ± 0.0041
Methods	PSM			SMAP		
	P	R	F1	P	R	F1
w/o MWE & TFCAE	**0.9670** ± 0.0360	0.8639 ± 0.0593	0.9113 ± 0.0170	0.6752 ± 0.0057	0.9633 ± 0.0000	0.7193 ± 0.0045
w/o MWE & *w/* MLP	0.9415 ± 0.0013	0.9058 ± 0.0000	0.9233 ± 0.0006	0.6560 ± 0.0006	0.9633 ± 0.0000	0.7073 ± 0.0012
w/o CG & *w/* Rand	0.9268 ± 0.0027	0.9165 ± 0.0148	0.9216 ± 0.0062	0.6666 ± 0.0169	0.9499 ± 0.0035	0.7033 ± 0.0149
w/o CG & *w/* Corr	0.9405 ± 0.0030	0.9058 ± 0.0000	0.9229 ± 0.0015	0.6546 ± 0.0004	0.9633 ± 0.0000	0.7080 ± 0.0001
w/ E2E	0.9148 ± 0.0102	0.9629 ± 0.0060	0.9382 ± 0.0082	0.6914 ± 0.0028	**0.9802** ± 0.0128	0.7419 ± 0.0061
TFCAD	0.9559 ± 0.0018	**0.9827** ± 0.0001	**0.9691** ± 0.0009	**0.7173** ± 0.0023	0.9798 ± 0.0044	**0.7801** ± 0.0006
Methods	SMD			SWaT		
	P	R	F1	P	R	F1
w/o MWE & TFCAE	0.8944 ± 0.0025	0.7761 ± 0.0017	0.8059 ± 0.0000	0.9108 ± 0.0012	0.8691 ± 0.0000	0.8894 ± 0.0006
w/o MWE & *w/* MLP	0.8882 ± 0.0100	0.8519 ± 0.0049	0.8573 ± 0.0002	0.9218 ± 0.0226	0.8676 ± 0.0146	0.8937 ± 0.0030
w/o CG & *w/* Rand	0.7964 ± 0.0030	0.7708 ± 0.0005	0.7365 ± 0.0013	0.8875 ± 0.0272	0.8700 ± 0.0303	0.8782 ± 0.0021
w/o CG & *w/* Corr	0.8889 ± 0.0101	0.8495 ± 0.0044	0.8564 ± 0.0001	0.9069 ± 0.0003	0.8779 ± 0.0000	0.8922 ± 0.0001
w/ E2E	0.8659 ± 0.0021	0.8323 ± 0.0130	0.8341 ± 0.0018	0.9261 ± 0.0258	0.8627 ± 0.0216	0.8929 ± 0.0004
TFCAD	**0.8982** ± 0.0036	**0.8861** ± 0.0086	**0.8852** ± 0.0050	**0.9315** ± 0.0199	**0.9006** ± 0.0131	**0.9156** ± 0.0028

***w/o* CG & *w/* Rand**. This variant replaces the learned Causal Graph (CG) with a randomly generated adjacency matrix. This baseline evaluates whether the performance gains are associated with the learned structural dependencies rather than merely introducing arbitrary graph structures into the GCN.

***w/o* CG & *w/* Corr**. This variant retains the main architecture but replaces the dynamically learned Causal Graph with a static correlation graph built from Pearson correlation coefficients, enabling a controlled comparison between the learned directed predictive structure and a conventional symmetric correlation-based prior.

***w/* E2E**. This variant removes the gradient-decoupling mechanism and the non-differentiable hard acyclic projection, allowing the causal discovery and anomaly reconstruction objectives to be optimized in a fully end-to-end manner. This setting is used to examine whether allowing both objectives to jointly influence the continuous adjacency matrix induces gradient conflicts during graph structure learning, and to evaluate the resulting effect on anomaly detection performance.

The results presented in Table 4 clearly demonstrate the effectiveness of each core component in TFCAD through a progressive comparison of model variants.

The *w/o* MWE & TFCAE variant shows the most substantial performance degradation across datasets. For example, its F1 score drops sharply from 0.9325 to 0.7237 on ASD, and from 0.8973 to 0.8268 on MSL. Operating the GCN directly on raw time-series inputs — without multi-scale time–frequency features or causal guidance — leads to a substantial loss in detection capability. It underscores the indispensable value of the proposed two modules for effectively capturing the complex patterns required for accurate anomaly detection.

A finer-grained validation is provided by the *w/o* MWE & *w/* MLP variant. Although replacing the MWE with a pure MLP-based feature extractor maintains a comparable parameter capacity, it still performs worse than the full model (e.g., 0.7073 vs. 0.7801 on SMAP, and 0.9233 vs. 0.9691 on PSM). This significant difference rigorously proves that the performance improvements are inherently driven by the time–frequency decomposition architecture itself, rather than merely an increase in neural network capacity.

To explicitly investigate the sensitivity of the causality-guided GCN to arbitrary graph structures, we introduce the *w/o* CG & *w/* Rand variant. This variant suffers a significant performance drop, with its F1 score plummeting to 0.5489 on ASD and 0.7365 on SMD. These results indicate that the GCN is sensitive to arbitrary graph structures and that simply introducing graph connectivity is insufficient to reproduce the performance gains provided by the learned structural prior.

Furthermore, the *w/o* CG & *w/* Corr variant also exhibits noticeable performance loss (e.g., an F1 score of 0.8490 on MSL and 0.8564 on SMD, falling short of TFCAD's 0.8973 and 0.8852, respectively). Unlike the learned directed structure, the static Pearson correlation graph is symmetric and does not encode directional predictive dependencies. The observed performance gap indicates that the learned directed predictive structure provides a more effective structural prior for downstream anomaly detection than the conventional correlation-based graph.

Finally, the results of the *w/* E2E variant reveal a distinct performance degradation compared to the full TFCAD model, with the F1 score dropping to 0.8866 on ASD and 0.8341 on SMD (versus TFCAD's 0.9325 and 0.8852). Notably, on the SMAP dataset,

Table 5
Gradient interaction diagnostics for the E2E variant across six benchmark datasets. ρ measures the directional alignment between the Phase-1 structural gradient and the reconstruction gradient with respect to the continuous adjacency matrix, while κ denotes the percentage of sampled mini-batches with $\rho < 0$.

Dataset	ρ	κ (%)
ASD	−0.3887	56.45
MSL	−0.1765	47.40
PSM	−0.0524	53.60
SMAP	−0.3282	48.04
SMD	−0.3032	62.71
SWaT	−0.0340	45.39

although the end-to-end optimization yields a marginally higher Recall (0.9802), it causes a noticeable decrease in Precision (0.6914 vs. TFCAD's 0.7173), leading to a lower overall F1-score (0.7419 vs. 0.7801). These results show that directly coupling the structural and reconstruction objectives can adversely affect anomaly detection performance. However, performance degradation alone does not reveal whether this effect arises from gradient conflict; therefore, we further examine the interaction between the two objectives directly in the adjacency space.

5.5. gradient conflict diagnostics

To directly examine the optimization interaction observed in the E2E setting, we further analyze the gradients of the Phase-1 structural objective and the reconstruction objective with respect to the continuous adjacency matrix $\boldsymbol{A}$ during E2E training. Following the gradient definitions in Section 4.6, every 10 mini-batches we separately obtain $\boldsymbol{g}^{\mathrm{p1}}$ and $\boldsymbol{g}^{\mathrm{rec}}$ before the standard backward pass and compute their cosine similarity, ρ. This diagnostic procedure does not alter the forward computation, loss functions, optimizer, or parameter updates of the original E2E training process. In addition to the mean ρ, we quantify how frequently conflicting gradient directions occur using

$$\kappa = \frac{1}{N}\sum_{q=1}^{N} \mathbb{I}\left(\rho_q < 0\right), \tag{24}$$

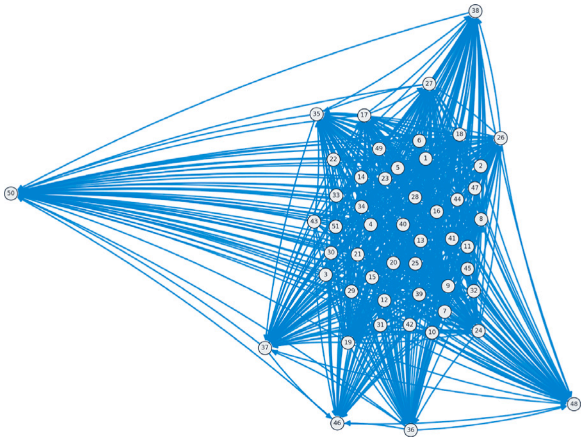
where N denotes the number of sampled mini-batches, q indexes the sampled mini-batches, and $\mathbb{I}(\cdot)$ is the indicator function. A negative ρ_q indicates that the angle between the two objective gradients exceeds 90°, such that their descent directions partially oppose each other in the adjacency space. The diagnostic experiment is repeated over three independent runs, and the reported values are averaged across runs.

As shown in Table 5, the mean ρ is negative on all six datasets, while κ ranges from 45.39% to 62.71%. The degree of directional opposition varies across datasets: ASD and SMD exhibit more pronounced negative alignment, whereas the mean ρ values on PSM and SWaT are closer to zero. Nevertheless, conflicting gradient directions occur repeatedly across all benchmarks. These observations provide direct optimization-level evidence that the Phase-1 structural objective and the downstream reconstruction objective are not consistently aligned in the adjacency space. Together with the performance degradation of the E2E variant, the results support the gradient-decoupling design as a means of preventing downstream reconstruction updates from interfering with graph structure learning.

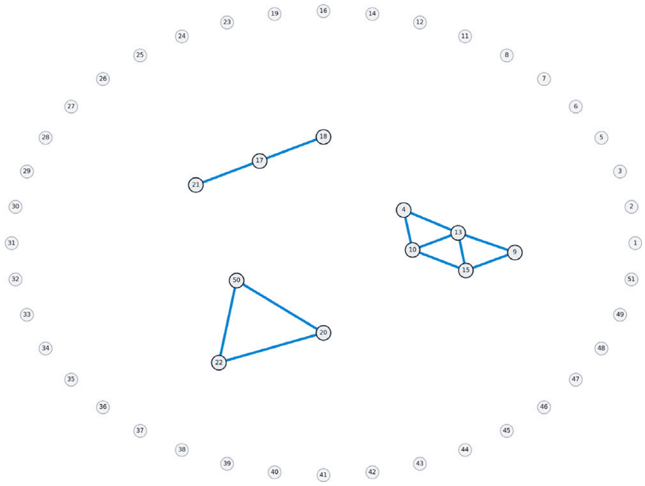
5.6. Visual analysis of structure learning

To further investigate the effectiveness of TFCAD in capturing the underlying inter-dependencies among multivariate time series, we provide a qualitative visualization by contrasting the Granger-based DAG learned by TFCAD (Fig. 4(a)) with the static correlation graph derived from the *w/o* CG & *w/* Corr variant in Section 5.4 (Fig. 4(b)). This controlled visual analysis is conducted on the SWaT dataset under the same experimental settings, utilizing the exact same data split and random seed for both methods. The two visualized graphs correspond to the final graph priors extracted from their respective trained checkpoints. Specifically, nodes denote the SWaT sensor/actuator variables, while edges encode the dependency structure used by each method (directed edges in TFCAD and undirected edges in the Pearson correlation-based graph of *w/o* CG & *w/* Corr). Additionally, the spatial layout in our graph visualization is generated using a force-directed layout to improve readability, rather than for physical or causal interpretation. Geometric proximity and edge lengths are optimization artifacts of the drawing algorithm and do not represent the strength of causality, physical coupling, or propagation delay. Therefore, our visual analysis focuses on node identity, edge direction, and overall connectivity statistics, rather than their spatial distribution.

From the resulting structures, TFCAD generates a graph with 663 directed edges and 0 isolated nodes, whereas the Pearson-based correlation graph of the *w/o* CG & *w/* Corr variant produces only 12 undirected edges alongside 40 isolated nodes. We observe that TFCAD produces a significantly richer and more connected topology. This indicates that TFCAD provides a more informative structural prior with directed predictive dependencies in the Granger sense, whereas the Pearson-based correlation captures only symmetric pairwise co-variation and cannot represent directionality. Consequently, the GCN can leverage these directed predictive dependencies to better distinguish normal operational patterns from anomalous deviations.



(a) Granger-based DAG learned by TFCAD (Edges: 663, Isolated: 0).



(b) Static correlation graph (Edges: 12, Isolated: 40).

Fig. 4. Topological comparison of extracted dependencies on the SWaT dataset: (a) the Granger-based DAG learned by TFCAD, and (b) the static correlation graph. Nodes represent sensor/actuator variables, while edges denote the extracted dependencies. The denser and directed topology in (a) provides a richer structural prior than the sparse, symmetric correlation graph in (b), offering more detailed dependency information for downstream representation learning.

Table 6
Dataset-wise acyclicity statistics on learned directed graphs. The 100% DAG rate across all benchmarks demonstrates that the proposed hybrid constraint strategy consistently enforces strict acyclicity in the structural priors supplied to the downstream GCN.

Dataset	Avg. Number of edges	DAG rate (%)	Avg. In/Out degree
ASD	91.3	100.0	4.81/4.81
MSL	763.0	100.0	13.87/13.87
PSM	160.0	100.0	6.40/6.40
SMAP	157.6	100.0	6.30/6.30
SMD	367.9	100.0	9.68/9.68
SWaT	663.0	100.0	13.00/13.00

5.7. Empirical validation of causal graph acyclicity

Moving beyond the topological connectivity discussed in the previous section, we now focus on the acyclicity requirement imposed on the learned directed predictive structure. This section examines the hybrid constraint strategy from Section 4.2.2 and evaluates whether the resulting structural priors satisfy the required DAG constraint.

Table 7
Training time comparison on six benchmark datasets (measured in minutes). The results demonstrate that despite incorporating complex multi-scale time–frequency extraction and dynamic causal discovery, TFCAD maintains practical time efficiency comparable to standard deep learning baselines, achieving a highly favorable balance between superior detection performance and computational cost.

Model	ASD	MSL	PSM	SMAP	SMD	SWaT
VAE	24.0000 ± 0.0000	28.5000 ± 0.7071	178.5000 ± 27.1823	53.0000 ± 0.0000	430.0000 ± 4.9533	310.0000 ± 7.0711
LSTM	55.8650 ± 12.1537	55.4600 ± 7.4499	105.5000 ± 8.3883	137.0750 ± 4.4458	371.5300 ± 29.1300	29.3500 ± 26.4458
LSTM-VAE	54.5000 ± 2.6066	43.0000 ± 5.4558	43.5000 ± 0.7071	96.0000 ± 8.0955	271.5000 ± 22.1356	2.5000 ± 0.7071
AMFormer	51.5000 ± 13.1335	53.0000 ± 8.3848	47.5000 ± 2.5772	41.0000 ± 10.7990	363.5000 ± 13.8148	302.5000 ± 26.4788
DADA	0.0108 ± 0.0000	0.0100 ± 0.0000	0.0600 ± 0.0000	0.0101 ± 0.0001	0.0145 ± 0.0002	0.1750 ± 0.0212
USAD	122.0000 ± 7.0711	688.5000 ± 2.1213	107.5000 ± 7.5772	447.0000 ± 16.7990	1321.0000 ± 90.5097	45.5000 ± 0.7071
COCA	248.5000 ± 20.4056	111.5000 ± 2.1213	109.5000 ± 10.8148	210.0000 ± 4.2426	1129.0000 ± 98.9949	21.5000 ± 6.3640
MGCLAD	26.5000 ± 3.3345	8.5000 ± 4.9497	88.0000 ± 4.0416	37.0000 ± 3.9411	111.0000 ± 9.9239	36.0000 ± 5.4558
MTGFLOW	69.0000 ± 13.2548	47.5000 ± 16.2635	31.5000 ± 0.7071	91.5000 ± 35.8614	447.0000 ± 67.2864	214.0000 ± 24.4508
AERCA	26.5000 ± 3.5355	47.5000 ± 2.1213	10.0000 ± 0.0000	24.0000 ± 2.8284	173.0000 ± 8.4853	510.0000 ± 18.3848
CATCH	11.5000 ± 2.1213	98.5000 ± 7.7782	72.5000 ± 9.7021	6.5000 ± 3.5355	509.0000 ± 16.8772	228.5000 ± 7.4767
ARIMA	0.4331 ± 0.2431	0.4534 ± 0.2454	18.4491 ± 4.8010	0.2790 ± 0.1910	14.3211 ± 0.2705	60.0891 ± 3.9149
PCA	0.0001 ± 0.0000	0.0006 ± 0.0001	0.0017 ± 0.0007	0.0002 ± 0.0001	0.0094 ± 0.0004	0.0509 ± 0.0066
TFCAD	15.0000 ± 1.4142	56.0000 ± 7.0711	9.0000 ± 0.0000	43.5000 ± 0.7071	107.0000 ± 4.2426	79.5000 ± 23.3345

Table 8
Performance of root cause analysis on the SMD dataset. The results compare the effectiveness of different structural priors for node-level anomaly localization under the same evaluation protocol.

Method	AC@1	AC@3	AC@5
w/o CG & *w/* Rand	0.6942 ± 0.0000	0.7967 ± 0.0022	0.8395 ± 0.0022
w/o CG & *w/* Corr	0.7238 ± 0.0018	0.8369 ± 0.0017	0.8695 ± 0.0018
TFCAD	**0.7500** ± 0.0039	**0.8945** ± 0.0058	**0.9167** ± 0.0029

Consistent with the visual analysis in Section 5.6, our validation approach involves extracting the final learned adjacency matrix from the fully trained TFCAD checkpoint. To reconstruct the precise topological structure utilized by the downstream GCN, we apply a standard noise-filtering threshold (0.01) to eliminate negligible connection weights and exclude self-loops. An automated topological cycle check using NetworkX is then performed on each resulting directed graph.

Table 6 summarizes the outcomes of this rigorous analysis across all six benchmark datasets. The results show that our method generates an acyclic directed graph in 100% of the evaluated cases. This perfect acyclicity score provides strong empirical evidence for the robustness of our method in ensuring structural integrity.

This robustness directly follows from the two-stage constraint mechanism. The soft constraints guide the optimization process toward structures that closely approximate a DAG, reducing the structural discrepancy introduced during post-processing. The subsequent hard acyclic projection then enforces a strict DAG structure. As a unified mechanism, the soft regularization and hard projection ensure that the structural prior supplied to the downstream GCN satisfies the required sparsity and acyclicity constraints.

5.8. Time-efficiency analysis

To empirically validate this theoretical complexity, we summarize the total training time of all models across the six benchmark datasets in Table 7.

Overall, the measured runtime is strictly bounded by the number of variables M and window size W, fully aligning with our theoretical estimations. The empirical results demonstrate that despite integrating multi-scale wavelet lifting and graph-based causal discovery, the time efficiency of TFCAD remains acceptable and comparable to that of standard baselines. This confirms that TFCAD maintains an elegant balance between superior anomaly detection performance and practical time efficiency.

5.9. Root cause analysis

The learned directed structural prior can support Root Cause Analysis (RCA) by providing dependency-aware guidance for identifying anomalous variables associated with detected system-level anomalies. To quantitatively evaluate this capability, we conduct an experiment on the SMD dataset using its official node-level interpretation labels. When a system-level anomaly is detected, variables are ranked according to their node-level reconstruction errors under the guidance of the learned projected graph $\boldsymbol{A}^{\mathrm{cau}}$. Performance is evaluated using the Recall at Top-K (AC@K) metric, which measures whether the labeled anomalous variables are retrieved among the top-ranked candidates.

As shown in Table 8, TFCAD achieves the best RCA performance among the three evaluated variants, reaching an AC@5 of 0.9167. Replacing the learned graph with random connectivity reduces AC@1 from 0.7500 to 0.6942, while the correlation-based variant obtains an AC@3 of 0.8369 compared with 0.8945 for TFCAD. These results indicate that the learned directed predictive structure provides more effective guidance for ranking anomalous variables than either arbitrary connectivity or a symmetric correlation-based prior. The improved AC@K performance further demonstrates the practical value of the learned structural prior for node-level anomaly localization and downstream RCA.

6. Discussion and implications

This section discusses the broader significance of TFCAD from both theoretical and practical perspectives, highlighting how it diverges from existing MTS anomaly detection paradigms.

6.1. Theoretical implications

The theoretical contribution of TFCAD lies in its integration of non-linear time–frequency analysis with Granger-based structural learning. Unlike traditional single-domain methods that rely solely on temporal or spectral information, TFCAD demonstrates that learning directed predictive dependencies in the time–frequency domain can effectively capture latent inter-variable relationships that are often obscured in raw signals. By leveraging lifting-scheme-based wavelet transformations, the model transforms non-stationary data into a representation that is more amenable to structured dependency learning.

Crucially, the architecture of TFCAD extends beyond a mere engineering combination of existing modules. Conceptually, our framework combines time–frequency representation learning and Granger-based structural learning through a gradient-decoupled architecture. In a fully end-to-end configuration, the causal discovery and reconstruction objectives can impose conflicting update directions on the shared continuous adjacency matrix, introducing interference into graph structure optimization. TFCAD addresses this interaction through gradient decoupling and hard acyclicity projection, which isolate graph structure learning from downstream reconstruction updates while providing a strictly acyclic structural prior to the anomaly detection module. Therefore, the core conceptual novelty lies in how structural priors are extracted from non-stationary signals and incorporated into downstream reconstruction without cross-objective interference.

6.2. Practical implications

From a practical standpoint, TFCAD offers a promising perspective for addressing the industry need for more robust and interpretable anomaly detection in complex systems. In real-world environments like server monitoring (SMD) or water treatment (SWaT), operators are often overwhelmed by false positives triggered by spurious correlations. TFCAD helps mitigate this by utilizing the learned graphs to filter out non-essential interactions, ensuring that the detection focus remains on violations of valid predictive dependencies. Furthermore, the discovered DAGs can serve as a valuable structural reference for downstream root-cause analysis, assisting engineers in narrowing down potential source metrics rather than treating the system as a complete black box.

7. Conclusion

In this paper, we present TFCAD, an unsupervised anomaly detection framework for multivariate time series. By decomposing non-stationary signals into multi-scale time–frequency representations and leveraging the learned Granger-based directed predictive structure, the proposed framework captures structurally meaningful inter-variable interactions. The integration of structure-guided graph convolution with reconstruction-based learning enables the model to achieve improved sensitivity to anomalous patterns, resulting in more discriminative and stable anomaly scores in complex multivariate settings. Extensive evaluations on six benchmark datasets demonstrate that TFCAD outperforms thirteen baselines across multiple evaluation metrics.

Despite its promising performance, several limitations warrant further investigation. First, the matrix-exponential-based acyclicity constraint imposes a non-trivial computational overhead during joint optimization, particularly as the variable dimensionality scales. Second, the current architecture assumes strictly synchronous time series, which may limit its direct applicability to asynchronously sampled industrial systems. Third, the benchmark datasets used in this study do not provide ground-truth causal DAGs, interventional observations, or complete physical system topologies. Accordingly, the learned graphs are interpreted as directed predictive dependencies in the Granger sense rather than as directly verified physical causal structures. Future work will investigate datasets with explicit structural annotations or interventional information to enable more direct evaluation of causal graph recovery. Addressing these constraints to improve scalability, flexibility, and structural validation remains a vital direction for our future research.

CRediT authorship contribution statement

Wenxuan He: Writing – original draft, Software, Methodology, Conceptualization. **Mingjing Du:** Writing – original draft, Supervision, Project administration, Funding acquisition. **Wenbin Zhang:** Writing – review & editing, Validation.

Data availability

Data will be made available on request.

References

Abdulaal, A., Liu, Z., & Lancewicki, T. (2021). Practical approach to asynchronous multivariate time series anomaly detection and localization. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery and data mining* (pp. 2485–2494). http://dx.doi.org/10.1145/3447548.3467174.

Audibert, J., Michiardi, P., Guyard, F., Marti, S., & Zuluaga, M. A. (2020). USAD: Unsupervised anomaly detection on multivariate time series. In *Proceedings of the 26th ACM SIGKDD conference on knowledge discovery and data mining* (pp. 3395–3404). http://dx.doi.org/10.1145/3394486.3403392.

Chen, L., Gao, X., Liu, J., Zhang, Y., Diao, X., Artis, T. W., Lu, J., & Meng, Z. (2025). A multivariate time series anomaly detection method with multi-grain dynamic receptive field. *Knowledge-Based Systems*, *309*, Article 112768. http://dx.doi.org/10.1016/j.knosys.2024.112768.

Darban, Z. Z., Webb, G. I., Pan, S., Aggarwal, C., & Salehi, M. (2025). Deep learning for time series anomaly detection: A survey. *ACM Computing Surveys*, *57*(1), 1–42. http://dx.doi.org/10.1145/3691338.

Ding, C., Sun, S., & Zhao, J. (2023). MST-GAT: A multimodal spatial-temporal graph attention network for time series anomaly detection. *Information Fusion*, *89*, 527–536. http://dx.doi.org/10.1016/j.inffus.2022.08.011.

Egilmez, H. E., Pavez, E., & Ortega, A. (2017). Graph learning from data under Laplacian and structural constraints. *IEEE Journal of Selected Topics in Signal Processing*, *11*(6), 825–841. http://dx.doi.org/10.1109/JSTSP.2017.2726975.

Fang, Y., Xie, J., Zhao, Y., Chen, L., Gao, Y., & Zheng, K. (2024). Temporal-frequency masked autoencoders for time series anomaly detection. In *Proceedings of the 40th IEEE international conference on data engineering* (pp. 1228–1241). http://dx.doi.org/10.1109/ICDE60146.2024.00099.

Han, X., Absar, S., Zhang, L., & Yuan, S. (2025). Root cause analysis of anomalies in multivariate time series through Granger causal discovery. In *International conference on learning representations*. https://openreview.net/forum?id=k38Th3x4d9.

He, S., Du, M., Jiang, X., Zhang, W., & Wang, C. (2024). VAEAT: Variational autoencoder with adversarial training for multivariate time series anomaly detection. *Information Sciences*, *676*, Article 120852. http://dx.doi.org/10.1016/j.ins.2024.120852.

He, S., He, W., Du, M., Jiang, X., & Dong, Y. (2026). GDCMAD: Graph-based dual-contrastive representation learning for multivariate time series anomaly detection. *Information Sciences*, *728*, Article 122790. http://dx.doi.org/10.1016/j.ins.2025.122790.

Hu, S., Zhao, K., Qiu, X., Shu, Y., Hu, J., Yang, B., & Guo, C. (2024). MultiRC: Joint learning for time series anomaly prediction and detection with multi-scale reconstructive contrast. http://dx.doi.org/10.48550/arXiv.2410.15997, arXiv preprint.

Huang, J., Liu, K., Xu, M., Perc, M., & Li, X. (2021). Background purification framework with extended morphological attribute profile for hyperspectral anomaly detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *14*, 8113–8124. http://dx.doi.org/10.1109/JSTARS.2021.3103858.

Hundman, K., Constantinou, V., Laporte, C., Colwell, I., & Söderström, T. (2018). Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 387–395). http://dx.doi.org/10.1145/3219819.3219845.

Janzing, D., Balduzzi, D., Grosse-Wentrup, M., & Schölkopf, B. (2013). Quantifying causal influences. *The Annals of Statistics*, *41*(5), 2324–2358. http://dx.doi.org/10.1214/13-aos1145.

Jin, M., Wang, S., Ma, L., Chu, Z., Zhang, J. Y., Shi, X., Chen, P.-Y., Liang, Y., Li, Y.-F., Pan, S., & Wen, Q. (2024). Time-LLM: Time series forecasting by reprogramming large language models. In *International conference on learning representations*. https://openreview.net/forum?id=Unb5CVPtae.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *International conference on learning representations*. http://arxiv.org/abs/1412.6980.

Li, Z., Zhao, Y., Han, J., Su, Y., Jiao, R., Wen, X., & Pei, D. (2021). Multivariate time series anomaly detection and interpretation using hierarchical inter-metric and temporal embedding. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery and data mining* (pp. 3220–3230). http://dx.doi.org/10.1145/3447548.3467075.

Malhotra, P., Vig, L., Shroff, G., & Agarwal, P. (2015). Long short term memory networks for anomaly detection in time series. In *Proceedings of the 23rd European symposium on artificial neural networks* (pp. 89–94). https://www.esann.org/sites/default/files/proceedings/legacy/es2015-56.pdf.

Mathur, A. P., & Tippenhauer, N. O. (2016). SWaT: A water treatment testbed for research and training on ICS security. In *Proceedings of the 2016 international workshop on cyber-physical systems for smart water networks* (pp. 31–36). http://dx.doi.org/10.1109/CYSWATER.2016.7469060.

Meier, S., Raut, P. N., Mahr, F., Thielen, N., Franke, J., & Risch, F. (2025). Structured knowledge-based causal discovery: Agentic streams of thought. *Information Processing & Management*, *62*(5), Article 104202. http://dx.doi.org/10.1016/j.ipm.2025.104202.

Park, D., Hoshi, Y., & Kemp, C. C. (2018). A multimodal anomaly detector for robot-assisted feeding using an LSTM-based variational autoencoder. *IEEE Robotics and Automation Letters*, *3*(2), 1544–1551. http://dx.doi.org/10.1109/LRA.2018.2801475.

Qin, S., Chen, L., Luo, Y., & Tao, G. (2023). Multiview graph contrastive learning for multivariate time-series anomaly detection in IoT. *IEEE Internet of Things Journal*, *10*(24), 22401–22414. http://dx.doi.org/10.1109/JIOT.2023.3303946.

Schmidl, S., Wenig, P., & Papenbrock, T. (2022). Anomaly detection in time series: A comprehensive evaluation. *Proceedings of the VLDB Endowment*, *15*(9), 1779–1797. http://dx.doi.org/10.14778/3538598.3538602.

Shentu, Q., Li, B., Zhao, K., Shu, Y., Rao, Z., Pan, L., Yang, B., & Guo, C. (2025). Towards a general time series anomaly detector with adaptive bottlenecks and dual adversarial decoders. In *International conference on learning representations*. https://openreview.net/forum?id=aKcd7ImG5e.

Su, Y., Zhao, Y., Niu, C., Liu, R., Sun, W., & Pei, D. (2019). Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 2828–2837). http://dx.doi.org/10.1145/3292500.3330672.

Tuli, S., Casale, G., & Jennings, N. R. (2022). TranAD: Deep transformer networks for anomaly detection in multivariate time series data. *Proceedings of the VLDB Endowment*, *15*(6), 1201–1214. http://dx.doi.org/10.14778/3514061.3514067.

Wang, R., Liu, C., Mou, X., Gao, K., Guo, X., Liu, P., Wo, T., & Liu, X. (2023). Deep contrastive one-class time series anomaly detection. In *Proceedings of the 2023 SIAM international conference on data mining* (pp. 694–702). http://dx.doi.org/10.1137/1.9781611977653.ch78.

Wang, J., Shao, S., Bai, Y., Deng, J., & Lin, Y. (2023). Multiscale wavelet graph autoencoder for multivariate time-series anomaly detection. *IEEE Transactions on Instrumentation and Measurement*, *72*, Article 2502911. http://dx.doi.org/10.1109/TIM.2022.3223142.

Wang, Y., Wang, J., Fan, Y., Chai, Q., Zhang, H., Li, X., & Li, R. (2025). Multi-granularity contrastive zero-shot learning model based on attribute decomposition. *Information Processing & Management*, *62*(1), Article 103898. http://dx.doi.org/10.1016/j.ipm.2024.103898.

Wei, Y., Tang, Y., Zheng, X., & Ji, C. (2025). Patch is effective for multivariate time series classification. *Knowledge-Based Systems*, *332*, Article 114923. http://dx.doi.org/10.1016/j.knosys.2025.114923.

Wu, H., Hu, T., Liu, Y., Zhou, H., Wang, J., & Long, M. (2023). TimesNet: Temporal 2D-variation modeling for general time series analysis. In *International conference on learning representations*. https://openreview.net/forum?id=ju_Uqw384Oq.

Wu, X., Qiu, X., Li, Z., Wang, Y., Hu, J., Guo, C., Xiong, H., & Yang, B. (2025). CATCH: Channel-aware multivariate time series anomaly detection via frequency patching. In *International conference on learning representations*. https://openreview.net/forum?id=m08aK3xxdJ.

Xu, H., Chen, W., Zhao, N., Li, Z., Bu, J., Li, Z., Liu, Y., Zhao, Y., Pei, D., Feng, Y., Chen, J., Wang, Z., & Qiao, H. (2018). Unsupervised anomaly detection via variational auto-encoder for seasonal KPIs in web applications. In *Proceedings of the 2018 world wide web conference* (pp. 187–196). http://dx.doi.org/10.1145/3178876.3185996.

Yao, H., Liu, M., Yin, Z., Yan, Z., Hong, X., & Zuo, W. (2024). GLAD: Towards better reconstruction with global and local adaptive diffusion models for unsupervised anomaly detection. In *Proceedings of the 18th European conference on computer vision* (pp. 1–17). http://dx.doi.org/10.1007/978-3-031-73209-6_1.

Yao, W., Wang, J., Perc, M., Yao, W., Dai, J., Guo, D., & Yao, D. (2021). Time irreversibility and amplitude irreversibility measures for nonequilibrium processes. *Communications in Nonlinear Science and Numerical Simulation*, *96*, Article 105688. http://dx.doi.org/10.1016/j.cnsns.2020.105688.

Yu, Y., Chen, J., Gao, T., & Yu, M. (2019). DAG-GNN: DAG structure learning with graph neural networks. In *Proceedings of the 36th international conference on machine learning* (pp. 7154–7163). http://proceedings.mlr.press/v97/yu19a.html.

Yu, H., Guo, P., & Sano, A. (2024). AdaWaveNet: Adaptive wavelet network for time series analysis. *Transactions on Machine Learning Research*, https://openreview.net/forum?id=m4bE9Y9FlX.

Zhang, C., Song, D., Chen, Y., Feng, X., Lumezanu, C., Cheng, W., Ni, J., Zong, B., Chen, H., & Chawla, N. V. (2019). A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data. In *Proceedings of the 33rd AAAI conference on artificial intelligence* (pp. 1409–1416). http://dx.doi.org/10.1609/aaai.v33i01.33011409.

Zhao, K., Zhuang, Z., Guo, C., Miao, H., Jensen, C. S., Cheng, Y., & Yang, B. (2025). Unsupervised time series anomaly prediction with importance-based generative contrastive learning. In *Proceedings of the 31st ACM SIGKDD conference on knowledge discovery and data mining* (pp. 3945–3956). http://dx.doi.org/10.1145/3711896.3737174.

Zheng, X., Aragam, B., Ravikumar, P., & Xing, E. P. (2018). DAGs with NO TEARS: Continuous optimization for structure learning. In *Advances in neural information processing systems* (pp. 9492–9503). https://proceedings.neurips.cc/paper/2018/hash/e347c51419ffb23ca3fd5050202f9c3d-Abstract.html.

Zhong, G., Liu, F., Jiang, J., Wang, B., & Chen, C. L. P. (2024). Refining one-class representation: A unified transformer for unsupervised time-series anomaly detection. *Information Sciences*, *656*, Article 119914. http://dx.doi.org/10.1016/j.ins.2023.119914.

Zhou, Q., He, S., Liu, H., Chen, J., & Meng, W. (2024). Label-free multivariate time series anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, *36*(7), 3166–3179. http://dx.doi.org/10.1109/TKDE.2024.3349613.