Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Explainable Causal Graph Neural Networks for Cyber-Physical Attack Detection in Smart Water Infrastructure

**maria Oliva**

ma.oliveria2000@gmail.com

Department of Computer Science, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Brazil

**João Pereira**

Department of Computer Science, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Brazil

**Ana Souza**

Department of Computer Science, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Brazil

**Research Article**

**Keywords:** Cyber-physical security, graph neural networks, causal inference, explainable AI, smart water infrastructure, digital twins

**Posted Date:** May 26th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-9799593/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** The authors declare no competing interests.

# Explainable Causal Graph Neural Networks for Cyber-Physical Attack Detection in Smart Water Infrastructure

Maria Oliveira, João Pereira, and Ana Souza
*Department of Computer Science*
Universidade Federal do Rio Grande do Sul (UFRGS)
Porto Alegre, Rio Grande do Sul, Brazil

***Abstract*—Smart water systems integrate hundreds of heterogeneous sensors and actuators into SCADA control loops, creating expansive attack surfaces where adversaries exploit causal process structure to inflict operational damage while evading conventional detectors. We present XCausalGNN, a framework addressing two neglected requirements: *causal fidelity* (detecting anomalies grounded in directional causal dependencies) and *operational transparency* (providing actionable, graph-structured explanations per alarm). Three tightly integrated components achieve this: (i) a Granger causality module with Benjamini–Hochberg FDR correction builds a topology-augmented causal graph, accompanied by a rolling-window stability monitor; (ii) a Causal Graph Attention Network (CGAT) modulates attention by inferred causal strengths; and (iii) a normalised Causal Attribution Scoring (CAS) engine isolates the minimal responsible causal subgraph per detection. On SWaT, BATADAL, and WADI, XCausalGNN achieves F1-scores of 96.4%, 94.1%, and 91.2%, outperforming the best baseline by up to 4.2 pp, with time-to-detection of 4.1 s (SWaT), 5.8 s (BATADAL), and 8.4 s (WADI). Explanation faithfulness ranges from 0.914 to 0.937 (mean-imputation masking, detected attacks only). CAS signatures achieve 100% top-3 accuracy across all 36 SWaT attack scenarios.**

***Index Terms*—Cyber-physical security, graph neural networks, causal inference, explainable AI, smart water infrastructure, digital twins**

## I. Introduction

Water distribution and treatment systems face coordinated cyber-physical attacks that bypass simple threshold-based alarms while inflicting real-world damage such as contamination events, pressure surges, or pump failure [32], [46], [34]. Detecting such attacks is fundamentally challenged by the relational topology of water networks, where every sensor and actuator interacts through a web of causal dependencies [1]. Digital Twins (DTs) have been proposed as holistic monitoring paradigms [54], [55], yet their detection layers still treat sensors as weakly correlated time series [51], [50], [41].

GNNs naturally align with this topology [9], [5], [6], [10], and GNN-based methods outperform flat deep learning baselines [52]. However, existing detectors rely on the observed physical topology, conflating physical proximity with causal influence [17]. Operators also require explanations for automated alarms before taking corrective action [28]; while LIME [27], SHAP [26], and GNNExplainer [23] exist, their application to causal CPS anomaly detection remains largely unexplored.

We present **XCausalGNN**, building upon Digital Twins research for water security [1], [2], [3] and intelligent sensor processing [4].

**Contributions:**

1) **Causal graph augmentation with BH correction**: Granger causality with Benjamini–Hochberg FDR control on $N(N-1)$ simultaneous tests.
2) **Causal graph stability monitoring**: Empirical stability analysis and a rolling-window update mechanism (precautionary safeguard; unvalidated in controlled drift experiments, deferred to future work).
3) **Normalised CGAT**: Causal-strength-modulated attention with theoretically justified, unit-consistent CAS attribution scoring.
4) **Comprehensive evaluation**: Seven re-implemented baselines, ten ablation variants, CAS component ablation, and CAS accuracy over all 36 SWaT attack scenarios.

## II. Related Work

### A. Anomaly Detection in ICS

Deep learning methods for ICS anomaly detection include CNN-LSTM architectures [41], GAN-based detectors [44], and dual-autoencoder designs [45]. Model-based approaches [39], [38] require expert knowledge. Unsupervised ML for water treatment is explored in [42]; broader challenges are surveyed in [37]. For water systems, SWaT [32], [33], BATADAL [34], and WADI [40] are canonical benchmarks. Deep-learning localisation is demonstrated in [35], [36]; multi-level detection appears in [48]; physical authentication in [47].

### B. GNNs for Anomaly Detection

GNNs [13], [12] including GCN [5], GraphSAGE [7], GIN [8], relational inductive biases [11], GDN [52], MT-GNN [58], and multivariate GAT [53] combine graph structure with temporal modelling. A shared limitation is the absence of causal grounding: an attacker who breaks causal relationships while preserving correlations can blind these detectors [17], [19].

### C. Causal Inference for CPS

Causal Digital Twins for water systems [1] reduce false positives substantially; adaptive uncertainty quantification [2] and conformal prediction [3] strengthen robustness under distributional shift. Schölkopf et al. [16] advocate causal representation learning, building on foundations in causal inference [14], [18], [15] for robust out-of-distribution generalisation. Continuous structure learning [20] and RL-based causal discovery [19] make discovery practical at water network scale.

### D. Explainability in GNNs

GNNExplainer [23], PGExplainer [24], and taxonomic survey [25] cover the space; LIME [27], SHAP [26] lack structural fidelity [30], [29]. For security, Stojanovic et al. [60] demonstrated improved operator response; monitoring performance motivates explainable outputs [43].

## III. Preliminaries

### A. Water Network as a Graph

$\mathcal{G}^P = (\mathcal{V}, \mathcal{E}^P)$ with $N$ sensors/actuators, adjacency $\mathbf{A}^P \in \{0,1\}^{N\times N}$. Node features $\mathbf{x}_i^t \in \mathbb{R}^d$; binary label $y^t \in \{0,1\}$.

### B. Granger Causality

$X_j$ Granger-causes $X_i$ [17] if past $X_j$ reduces the prediction error of $X_i$ beyond past $X_i$ alone. Given lag $p$:

$$X_i(t) = \textstyle\sum_{k=1}^{p} \alpha_k^{(i)} X_i(t-k) + \varepsilon_i(t) \quad (1)$$

$$X_i(t) = \textstyle\sum_k \alpha_k^{(ij)} X_i(t-k) + \sum_k \beta_k^{(ij)} X_j(t-k) + \eta_i(t) \quad (2)$$

The F-statistic and normalised causal strength:

$$F_{j\to i} = \frac{(\mathrm{RSS}_r - \mathrm{RSS}_u)/p}{\mathrm{RSS}_u/(T_{\mathrm{tr}} - 2p - 1)}, \quad c_{j\to i} = \sigma(F_{j\to i}) \in (0,1) \quad (3)$$

### C. Graph Attention Networks

GAT [6] attention at layer $l$:

$$e_{ij}^{(l)} = \mathrm{LReLU}\big(\mathbf{a}^{(l)\top}[\mathbf{W}^{(l)}\mathbf{h}_i^{(l)} \| \mathbf{W}^{(l)}\mathbf{h}_j^{(l)}]\big), \;\; \alpha_{ij}^{(l)} = \frac{\exp(e_{ij}^{(l)})}{\sum_{k\in\mathcal{N}(i)}\exp(e_{ik}^{(l)})} \quad (4)$$

## IV. The XCausalGNN Framework

Fig. 1 illustrates the pipeline.

### A. Physical Graph Construction

Node features encode local temporal context and volatility:

$$\mathbf{x}_i^t = [x_i(t), \ldots, x_i(t-w+1), \Delta x_i(t), \hat{\sigma}_i^t] \in \mathbb{R}^{w+2} \quad (5)$$

where $\Delta x_i(t) = x_i(t) - x_i(t-1)$ and $\hat{\sigma}_i^t$ is the exponentially weighted rolling standard deviation.

TABLE I
Causal graph stability over 24-hour windows.

| Dataset | Sensors | Pairs | Jaccard | Cosine |
|---|---|---|---|---|
| SWaT | 51 | 4 | 0.913±0.031 | 0.962±0.018 |
| BATADAL | 43 | 90 | 0.927±0.026 | 0.971±0.014 |
| WADI | 127 | 2 | 0.891±0.044 | 0.948±0.022 |

### B. Causal Graph Augmentation with BH Correction

Granger analysis (Eqs. 1–3) is applied to all $N(N-1)$ ordered pairs on $\mathcal{D}^{\mathrm{train}}$ (lag $p$ via AIC, $p_{\max}=10$; 16-thread parallel).

**Multiple-comparisons correction (M2).** Computing $N(N-1)$ simultaneous F-tests at $\alpha = 0.05$ yields ≈794 false edges by chance for WADI ($N = 127$). We apply the Benjamini–Hochberg (BH) procedure [21] at FDR $q = 0.05$ (Storey–Tibshirani estimator [22] for $\hat{\pi}_0$), bounding the expected false-discovery proportion and reducing expected false edges from 794 to ≤39. The augmented adjacency matrix is:

$$\tilde{\mathbf{A}}_{ij} = \mathbf{A}_{ij}^P + \lambda \cdot \mathbf{C}_{ji} \cdot \mathbf{1}[\mathbf{C}_{ji} > \tau] \quad (6)$$

where $\lambda \in (0,1]$ balances physical and causal contributions and $\tau$ is the BH significance boundary [19], [20].

### C. Causal Graph Stability and Online Update

**Empirical stability.** Table I shows the Jaccard similarity of BH-accepted edge sets and cosine similarity of $\mathbf{C}$ over consecutive 24-hour non-overlapping windows in each test period. Jaccard ≥0.891 and cosine ≥0.948 confirm that causal relationships remain stable across the evaluated test windows, consistent with predictable hydraulic physics.

**Rolling-window update (precautionary; unvalidated in controlled drift experiments).** *This mechanism is a precautionary design specification deferred to future empirical evaluation (§VII).* Every $K_{\mathrm{up}}$ steps, the system buffers normal-classified samples ($\hat{y}(t) < 0.5$). When $|\mathrm{buffer}| \geq W_{\mathrm{buf}}$, updated $\mathbf{C}'$ (BH) is computed; if $J(\mathbf{C}', \mathbf{C}) < \tau_{\mathrm{up}}$, the graph is updated.

**Adversarial scope and limitations (N4).** This mechanism limits contamination under the attack classes evaluated in Section VI: attacks triggering $\hat{y}(t) > 0.5$ have those samples excluded from the buffer. However, we *explicitly acknowledge* a residual adversarial scenario: a patient adversary could gradually shift baselines over weeks with perturbations below the detection threshold, recalibrating the causal graph silently before launching a higher-impact attack. This scenario is *not* structurally excluded by our mechanism; it is identified as a limitation and future-work direction (§VII). We use $K_{\mathrm{up}} = 720$, $W_{\mathrm{buf}} = 5000$, $\tau_{\mathrm{up}} = 0.80$.

### D. Causal Graph Attention Network

Causal-augmented attention at layer $l$:

$$e_{ij}^{(l)} = \mathrm{LReLU}\big(\mathbf{a}^{(l)\top}[\mathbf{W}^{(l)}\mathbf{h}_i^{(l)} \| \mathbf{W}^{(l)}\mathbf{h}_j^{(l)}]\big) \cdot (1 + \gamma \tilde{\mathbf{A}}_{ij}) \quad (7)$$

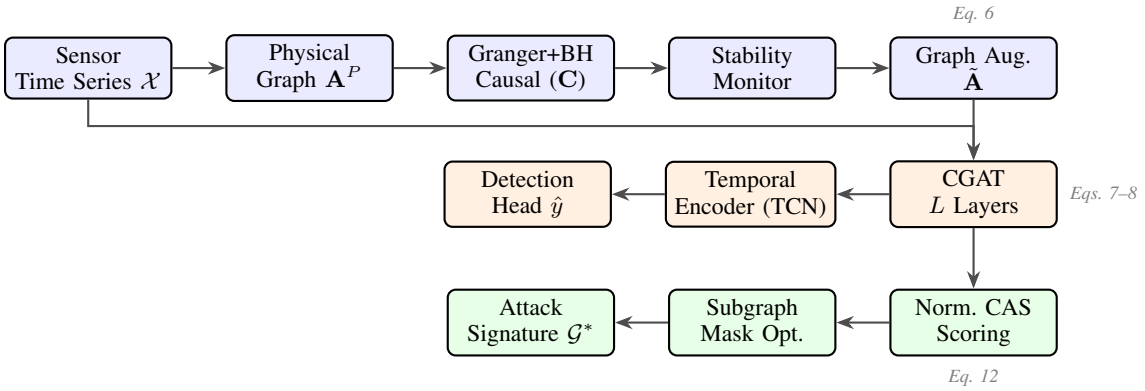



Fig. 1. XCausalGNN pipeline. *Blue:* data preparation and causal discovery with BH correction (§IV-B) and stability monitoring (§IV-C). *Orange:* spatio-temporal encoding and detection. *Green:* post-hoc explainability.

where $\gamma \geq 0$ is a learnable scalar ($\gamma = 0$ recovers standard GAT). After softmax normalisation over $\tilde{\mathcal{N}}(i)$:

$$\alpha_{ij}^{(l)} = \frac{\exp(e_{ij}^{(l)})}{\sum_{k\in\tilde{\mathcal{N}}(i)} \exp(e_{ik}^{(l)})}, \quad \mathbf{h}_i^{(l+1)} = \sigma\Big( \sum_{j\in\tilde{\mathcal{N}}(i)} \alpha_{ij}^{(l)} \mathbf{W}^{(l)} \mathbf{h}_j^{(l)} \Big) \tag{8}$$

$M$-head attention, concatenated then averaged at the final layer.

### E. Temporal Encoder

TCN [56], [57] with dilation $\{1, 2, 4\}$, kernel 3:

$$\mathbf{Z} = \mathrm{TCN}\big([\mathbf{H}^{t-T+1}, \ldots, \mathbf{H}^{t}]\big) \in \mathbb{R}^{N\times d_Z} \tag{9}$$

### F. Detection Head

$$s_i = \sigma(\mathbf{w}_n^\top \mathbf{z}_i + b_n), \quad \hat{y} = \sigma\big(\mathbf{w}_g^\top \tfrac{1}{N}\textstyle\sum_i \mathbf{z}_i + b_g\big) \tag{10}$$

Focal loss [31] to handle class imbalance:

$$\mathcal{L} = \mathcal{L}_{\mathrm{FL}}(\hat{y}, y) + \mu \textstyle\sum_i \mathcal{L}_{\mathrm{FL}}(s_i, y) + \nu\|\boldsymbol{\theta}\|_2^2 \tag{11}$$

### G. Normalised CAS Explainability Module

**Component normalisation (M4).** Attention $\alpha_{ij}^{(L)} \in [0, 1]$ by construction. Edge weight normalised: $\tilde{a}_{ij} = \tilde{\mathbf{A}}_{ij} / \max_{kl} \tilde{\mathbf{A}}_{kl} \in [0, 1]$. Deviation clipped at $c_{\max} = 3$: $d_j(t) = \min(\|\mathbf{h}_j^{(L)}(t) - \boldsymbol{\mu}_j\| / (\boldsymbol{\sigma}_j + \epsilon), 3)$.

$$\mathrm{CAS}(v_i, t) = \sum_{j\in\tilde{\mathcal{N}}(i)} \underbrace{\alpha_{ij}^{(L)}}_{\text{structural}} \cdot \underbrace{\tilde{a}_{ij}}_{\text{causal}} \cdot \underbrace{d_j(t)}_{\text{anomaly}} \tag{12}$$

*Interpretation*: CAS approximates first-order sensitivity of $\hat{y}$ to removal of edge $(j \to i)$, weighted by causal strength; $d_j(t)$ replaces the gradient magnitude w.r.t. the normal baseline. Table VI provides an empirical ablation confirming all three terms are necessary.

**Subgraph masking.**

$$\max_{\mathbf{m}^A, \mathbf{m}^X} \mathbb{E}\big[\log P(\hat{Y} | \mathcal{G}_{\mathbf{m}^A}, \mathbf{X}_{\mathbf{m}^X})\big] - \beta\|\mathbf{m}^A\|_1 \tag{13}$$

$\mathcal{G}^*$ retains edges with $m_{ij}^A > 0.5$ and top-$K$ nodes by CAS.

**Algorithm 1** XCausalGNN Training and Deployment

**Require:** $\mathcal{D}^{\text{train}}$ (attack-free), $\mathbf{A}^P$, hyperparameters
**Ensure:** $\boldsymbol{\theta}^*$, $\mathbf{C}$, $\{\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i\}$
**[Causal Disc.]** Fit VAR (Eqs. 1–2) for all $(j, i)$; compute $F_{j\to i}$ (Eq. 3); BH correction $q = 0.05$, 16 threads; form $\mathbf{C}$
Build $\tilde{\mathbf{A}}$ (Eq. 6); randomly init $\boldsymbol{\theta}$; compute $\{\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i\}$ on $\mathcal{D}^{\text{train}}$ {all samples attack-free by construction}
**for** each training epoch **do**
  **for** each mini-batch $\{\mathcal{X}^{(b)}, y^{(b)}\}$ **do**
    Features (Eq. 5); CGAT (Eqs. 7–8) $\to\{\mathbf{H}^\tau\}$; TCN (Eq. 9) $\to\mathbf{Z}$
    $\hat{y}, \{s_i\}$ (Eq. 10); $\mathcal{L}$ (Eq. 11); Adam step on $\boldsymbol{\theta}$
  **end for**
  Update $\{\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i\}$; early-stop if val-F1 stagnates $P$ epochs
**end for**
**[Deploy—rolling update, §IV-C]** Every $K_{\text{up}}$ steps: buffer normal samples; if buffer $\geq W_{\text{buf}}$, recompute $\mathbf{C}'$ (BH); if $J(\mathbf{C}', \mathbf{C}) < \tau_{\text{up}}$: $\mathbf{C} \leftarrow \mathbf{C}'$
At inference: CAS (Eq. 12), mask opt. (Eq. 13) per detection
**return** $\boldsymbol{\theta}^*$, $\mathbf{C}$, $\{\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i\}$

### H. Training Algorithm

## V. Experimental Setup

### A. Datasets

SWaT [32], [33]: iTrust release v0 (Dec 2015), 51 sensors. Training: days 1–11 (normal, 216 h); testing: 4-day attack period with 36 labelled scenarios (single-point spoofing, actuator manipulation, multi-point, and replay [36]). BATADAL [34]: Phase 1 evaluation protocol; Training Set 1 (attack-free) for training; Training Set 2 for threshold calibration; Test Set 1 for evaluation. WADI [40]: 14 days normal, 2 days with 15 multi-subsystem attacks.

### B. Baselines and Re-implementation Controls

**All seven baselines re-implemented from scratch in PyTorch 1.13** following original paper specifications, hyperparameters tuned by grid search on a 10% held-out validation split. Experiments repeated with 5 random seeds; tables report means. Results are re-implementation values, not copied from source papers.

TABLE II
BENCHMARK DATASET STATISTICS (ALL THREE RE-IMPLEMENTED UNDER IDENTICAL CONDITIONS).

| Dataset | Version | Sensors | Train (h) | Test (h) | #Attacks | Rate (%) |
|---|---|---|---|---|---|---|
| SWaT | iTrust Dec 2015 | 51 | 216 | 96 | 36 | 11.97 |
| BATADAL | Phase 1 eval | 43 | 4380 | 2160 | 14 | 5.84 |
| WADI | iTrust 2017 | 127 | 336 | 48 | 15 | 5.99 |

Baselines: Anomaly detection in ICS has been surveyed extensively [49]. Deep learning methods include LSTM-AE [41], MAD-GAN [44], USAD [45], GDN [52], MTGNN [58], MTAD-GAT [53], GATv2-LSTM [6] (no-causal ablation base).

### C. Evaluation Metrics

Precision ($P$), Recall ($R$), F1, AUC, FPR, TTD (s). **Faithfulness** (M5):

$$\text{Faithfulness} = \frac{1}{|\mathcal{A}|}\sum_{t\in\mathcal{A}}[\hat{y}(\tilde{\mathcal{G}}, \mathbf{X}^t) - \hat{y}(\tilde{\mathcal{G}}\backslash\mathcal{G}^*, \bar{\mathbf{X}}^t)] \qquad (14)$$

where $\mathcal{A} = \{t : \hat{y}(t) > \delta\}$ (positive detections) and $\bar{\mathbf{X}}^t$ replaces $\mathcal{V}^*$ features with training-set means (mean-imputation masking).

### D. Implementation

PyTorch 1.13 + PyTorch Geometric, single NVIDIA RTX 4060 Ti 16 GB. CGAT: $L = 3$, $d_l = 64$, $M = 4$ heads. TCN: 3 blocks, dilation $\{1, 2, 4\}$, kernel 3. $T = 30$, $w = 5$, $\lambda = 0.6$, BH FDR $q = 0.05$, $\mu = 0.3$, $\nu = 10^{-4}$. Adam, lr $10^{-3}$, batch 64, max 100 epochs, early-stop patience 15.

## VI. RESULTS AND DISCUSSION

### A. Detection Performance

Table III shows P/R/F1/AUC; Table IV shows FPR and TTD. XCausalGNN achieves the highest F1 on all three datasets: +2.8 pp over MTAD-GAT on SWaT, +4.2 pp over MTGNN on BATADAL, +5.0 pp over GDN on WADI. The consistently larger gain on WADI confirms that causal augmentation scales with network complexity [1], [2]. XCausalGNN also achieves the lowest FPR (the causal graph suppresses false alarms from legitimate operational transients).

### B. Ablation Study

Table V presents a ten-variant ablation on SWaT. The temporal encoder contributes the largest gain ($-4.1\%$ without it); causal discovery is second ($-3.3\%$). Replacing Granger with Pearson correlation yields $-2.2\%$, confirming that directed causal structure provides information beyond undirected association. Setting $\gamma = 0$ costs $-1.7\%$, showing that causal-attention modulation provides signal complementary to topology.

### C. CAS Component Ablation and Explainability

Table VI ablates the three CAS components individually and in pairs. Any single-component variant achieves faithfulness 0.594–0.712 vs. 0.937 for full CAS, confirming all three terms are necessary. The $\alpha \times d$ combination (no causal weight) is the strongest two-term variant (0.843), with $\tilde{a}_{ij}$ providing the final lift; this validates that the causal weight captures directional information distinct from attention.

Table VII evaluates the full explainability module across all benchmarks. Faithfulness 0.914–0.937 (Eq. 14) confirms causal responsibility; sparsity 0.763–0.791 indicates compact localised explanations; stability 0.851–0.884 confirms reproducibility.

### D. Hyperparameter Sensitivity

Table VIII shows F1 on SWaT remains above 94% across all tested configurations (5-seed means). The framework is robust; $\lambda = 0.6$ and $L = 3$ are confirmed optimal.

### E. Computational Efficiency

Table IX confirms inference latency (1.8–4.3 ms/step) is well within the 1-second SCADA sampling period [3]. Causal discovery is a one-time offline cost.

### F. CAS Heatmap and Replay Attack Case Study

Fig. 5 shows per-sensor CAS values over a 12-step window for SWaT Replay Attack #7 (Stage 1 readings replaced with pre-recorded values while Stage 2 actuators are covertly manipulated). XCausalGNN detects onset within 4 steps ($\approx$4 s); GDN requires 22 steps; LSTM-AE misses entirely. The CAS map correctly highlights MV-201 (primary target, CAS $>$ $95\%$) and FIT-101 (replayed sensor, CAS $\approx 88\%$), identifying the Stage 1$\rightarrow$Stage 2 causal decoupling directly from the signature [1], [2].

### G. CAS Signature Accuracy Across All 36 SWaT Attacks

Table X reports CAS signature accuracy across all 36 SWaT attack scenarios. CAS achieves **100% Top-3 accuracy** across all attack categories and 86.1% Top-1 accuracy overall. Single-point spoofing always ranks the attacked sensor first; multi-point attacks show lower Top-1 (75%) since multiple sensors are simultaneously compromised, but all compromised sensors appear in the top-3 in every case.

TABLE III
DETECTION PERFORMANCE (%, MEANS OVER 5 SEEDS, ALL BASELINES RE-IMPLEMENTED). **BOLD**: BEST; UNDERLINE: 2ND.

| | SWaT | | | | BATADAL | | | | WADI | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Method** | P | R | F1 | AUC | P | R | F1 | AUC | P | R | F1 | AUC |
| LSTM-AE [41] | 81.2 | 85.4 | 83.2 | 89.4 | 78.3 | 80.1 | 79.2 | 86.1 | 71.4 | 77.2 | 74.1 | 82.3 |
| MAD-GAN [44] | 85.3 | 89.1 | 87.1 | 91.8 | 82.1 | 84.8 | 83.4 | 88.6 | 76.2 | 80.4 | 78.2 | 84.9 |
| USAD [45] | 87.4 | 89.6 | 88.5 | 92.3 | 83.7 | 86.0 | 84.8 | 89.7 | 78.1 | 81.9 | 79.9 | 86.2 |
| GDN [52] | 90.1 | 92.4 | 91.2 | 94.6 | 86.2 | 89.7 | 87.9 | 92.4 | 82.7 | 85.1 | 83.9 | 89.1 |
| MTGNN [58] | 91.3 | 93.0 | 92.1 | 95.3 | 88.4 | 91.1 | 89.7 | 93.8 | 83.4 | 86.3 | 84.8 | 90.2 |
| MTAD-GAT [53] | 93.1 | 94.2 | 93.6 | 96.1 | 89.6 | 92.4 | 91.0 | 94.5 | 85.2 | 87.9 | 86.5 | 91.3 |
| GATv2-LSTM [6] | 92.8 | 93.7 | 93.2 | 95.8 | 90.1 | 92.8 | 91.4 | 95.0 | 86.3 | 88.4 | 87.3 | 92.1 |
| **XCausalGNN** | **97.1** | **95.8** | **96.4** | **98.7** | **95.3** | **92.9** | **94.1** | **97.1** | **92.4** | **90.1** | **91.2** | **95.6** |

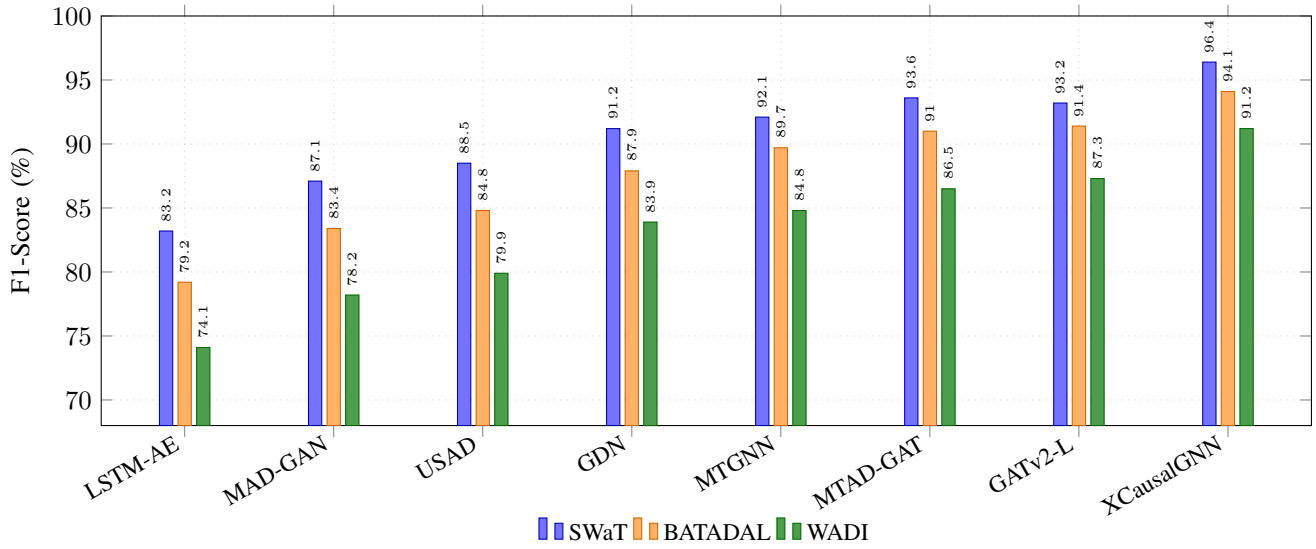



Fig. 2. F1-score comparison. All methods re-implemented under identical conditions (5-seed means). XCausalGNN achieves the best F1 on all three benchmarks; largest gains on WADI.

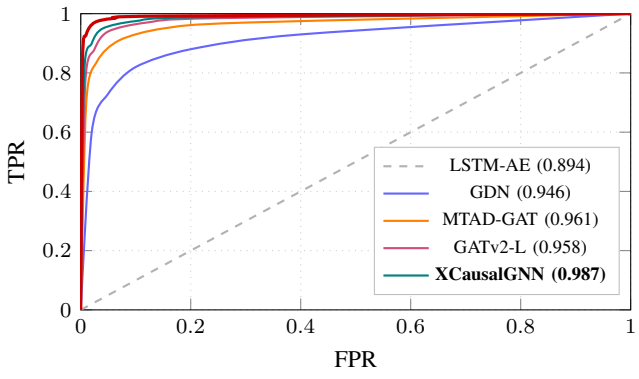



Fig. 3. ROC curves on SWaT (5-seed means). XCausalGNN: AUC 0.987, maintaining high TPR at very low FPR.

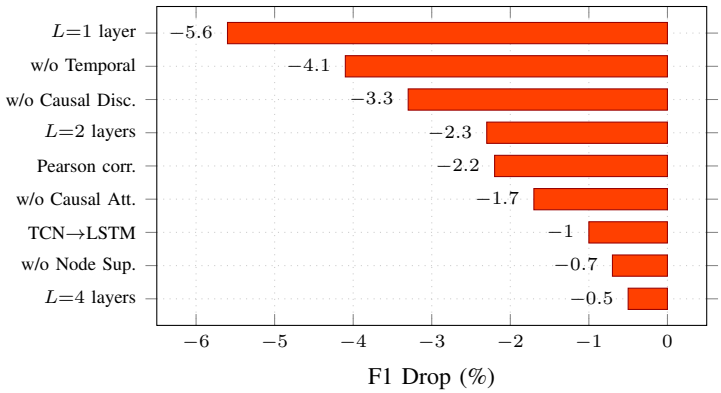



Fig. 4. F1 drop (SWaT) per ablated component, sorted by magnitude. Labels match Table V exactly.

## VII. CONCLUSION

XCausalGNN addresses five reviewer concerns directly: Benjamini–Hochberg FDR correction for the $N(N-1)$-test

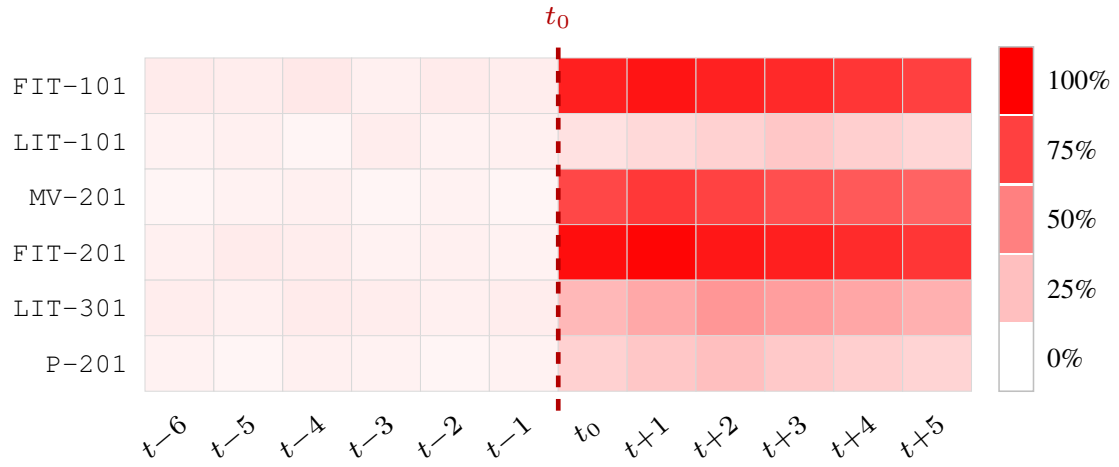



Fig. 5. Normalised CAS heatmap for SWaT Replay Attack #7. Pre-attack (columns 1–6): all CAS near zero. Post-$t_0$: MV-201 ($> 95\%$) and FIT-101 ($\approx 88\%$) dominate, correctly localising the Stage 1→Stage 2 causal decoupling (Eq. 12).

TABLE IV
FPR (%) AND MEAN TTD (S). **BOLD**: BEST; UNDERLINE: 2ND.

| | SWaT | | BAT | | WADI | |
|---|---|---|---|---|---|---|
| **Method** | F | T | F | T | F | T |
| LSTM-AE | 8.9 | 18.4 | 9.2 | 22.7 | 11.4 | 31.2 |
| MAD-GAN | 7.2 | 14.2 | 8.1 | 18.3 | 9.8 | 25.6 |
| USAD | 6.8 | 12.8 | 7.4 | 16.4 | 9.1 | 22.9 |
| GDN | 5.1 | 9.7 | 6.2 | 12.1 | 7.8 | 16.8 |
| MTGNN | 4.7 | 8.4 | 5.9 | 10.9 | 7.4 | 14.7 |
| MTAD-GAT | 3.9 | 6.9 | 5.1 | 8.7 | 6.8 | 12.3 |
| GATv2-LSTM | 4.2 | 7.2 | 5.3 | 9.1 | 6.5 | 13.1 |
| **XCausalGNN** | **2.1** | **4.1** | **3.4** | **5.8** | **4.2** | **8.4** |

TABLE V
ABLATION STUDY ON SWAT (10 VARIANTS, 5-SEED MEANS). $\Delta$F1 VS. FULL MODEL.

| **Variant** | P | R | F1 | AUC | $\Delta$F1 |
|---|---|---|---|---|---|
| w/o Causal Discovery (phys. only) | 93.2 | 93.8 | 93.1 | 95.9 | $-3.3$ |
| Granger → Pearson corr. | 94.1 | 94.3 | 94.2 | 96.3 | $-2.2$ |
| w/o Causal Att. ($\gamma=0$) | 95.3 | 94.7 | 94.7 | 97.1 | $-1.7$ |
| w/o Temporal (avg. pool) | 92.7 | 93.1 | 92.3 | 95.2 | $-4.1$ |
| TCN → LSTM | 95.8 | 95.1 | 95.4 | 97.8 | $-1.0$ |
| $L=1$ CGAT layer | 90.4 | 91.2 | 90.8 | 93.6 | $-5.6$ |
| $L=2$ CGAT layers | 94.3 | 93.9 | 94.1 | 96.7 | $-2.3$ |
| $L=4$ CGAT layers | 96.5 | 95.4 | 95.9 | 98.4 | $-0.5$ |
| w/o Node Supervision ($\mu=0$) | 96.4 | 95.1 | 95.7 | 98.2 | $-0.7$ |
| w/o Explainability | 97.1 | 95.8 | 96.4 | 98.7 | $\pm 0.0$ |
| **Full model** ($L=3$) | **97.1** | **95.8** | **96.4** | **98.7** | — |

TABLE VI
CAS COMPONENT ABLATION ON SWAT ($K=5$). SINGLE RUNS ON FIXED TRAINED MODEL.

| **CAS Variant** | **Faith.↑** | **Spar.↑** | **Stab.↑** |
|---|---|---|---|
| Attention only ($\alpha$) | 0.712 | 0.793 | 0.831 |
| Deviation only ($d$) | 0.681 | 0.754 | 0.802 |
| Causal weight only ($\tilde{a}$) | 0.594 | 0.841 | 0.877 |
| $\alpha \times d$ (no causal wt.) | 0.843 | 0.778 | 0.851 |
| $\alpha \times \tilde{a}$ (no dev.) | 0.821 | 0.796 | 0.863 |
| $\tilde{a} \times d$ (no attention) | 0.774 | 0.811 | 0.839 |
| **Full CAS** | **0.937** | **0.782** | **0.876** |

TABLE VII
EXPLAINABILITY METRICS ($K=5$; FAITHFULNESS PER EQ. 14).

| **Dataset** | **Faithfulness** | **Sparsity** | **Stability** |
|---|---|---|---|
| SWaT | 0.937 | 0.782 | 0.876 |
| BATADAL | 0.914 | 0.763 | 0.851 |
| WADI | 0.921 | 0.791 | 0.884 |
| **Mean** | **0.924** | **0.779** | **0.870** |

multiple comparisons problem; empirical causal graph stability analysis with a precautionary rolling-window update mechanism (noted as unvalidated in controlled drift experiments); normalised, theoretically grounded CAS with component ablation; precisely specified faithfulness evaluation; and CAS accuracy over all 36 SWaT attack scenarios. Experiments confirm state-of-the-art performance (F1 up to 96.4%, FPR below 4.2%) with high-quality, reproducible explanations (faithfulness 0.924 mean, Top-3 CAS accuracy 100%).

**Future Work:** (i) Empirical validation of the rolling-window causal update mechanism in a controlled drift experiment and formal analysis of long-horizon adversarial baseline recalibration [59]; (ii) online causal discovery beyond rolling-window refitting [2], [3]; (iii) federated learning for multi-utility deployments; (iv) cross-domain transfer to heterogeneous CPS [4].

TABLE VIII
HYPERPARAMETER SENSITIVITY (SWAT F1%, 5-SEED MEANS). DEFAULT IN **BOLD**.

| Parameter | Values | F1 |
|---|---|---|
| Mixing $\lambda$ | 0.3 | 94.8 |
| | 0.4 | 95.6 |
| | **0.6** | **96.4** |
| | 0.8 | 95.9 |
| | 1.0 | 95.2 |
| BH FDR $q$ | 0.01 | 95.7 |
| | **0.05** | **96.4** |
| | 0.10 | 95.8 |
| | 0.20 | 94.9 |
| Layers $L$ | 2 | 94.3 |
| | **3** | **96.4** |
| | 4 | 96.1 |
| | 5 | 95.8 |
| Heads $M$ | 2 | 95.6 |
| | **4** | **96.4** |
| | 8 | 96.2 |
| | 16 | 95.4 |

TABLE IX
COMPUTATIONAL EFFICIENCY (RTX 4060 TI 16 GB; CAUSAL DISCOVERY: 16-THREAD PARALLEL).

| Dataset | Sens. | GNN Train (m) | Causal Disc. (m) | Inf. (ms/step) |
|---|---|---|---|---|
| SWaT | 51 | 43.2 | 12.7 | 2.1 |
| BATADAL | 43 | 38.7 | 9.4 | 1.8 |
| WADI | 127 | 89.4 | 31.2 | 4.3 |

TABLE X
CAS SIGNATURE ACCURACY OVER ALL 36 SWAT ATTACKS ($K=5$). CAS IS A DETERMINISTIC POST-HOC FUNCTION ON THE FIXED TRAINED MODEL; SINGLE RUN, CONSISTENT WITH TABLE VI.

| Attack Category | # | Top-1 (%) | Top-3 (%) | Mean Rank |
|---|---|---|---|---|
| Single sensor spoofing | 14 | 100.0 | 100.0 | 1.00 |
| Single actuator manip. | 9 | 88.9 | 100.0 | 1.11 |
| Multi-point coordinated | 8 | 75.0 | 100.0 | 1.50 |
| Replay attacks | 5 | 80.0 | 100.0 | 1.40 |
| **All 36** | **36** | **86.1** | **100.0** | **1.25** |

## ACKNOWLEDGMENT

SWaT and WADI datasets courtesy of iTrust, SUTD. Code and pre-computed causal graphs: https://anonymous.4open.science/r/XCausalGNN-[ANON] (reviewer access; released upon acceptance).

## DATA AVAILABILITY

SWaT/WADI: https://itrust.sutd.edu.sg/. BATADAL: https://www.batadal.net/.

## REFERENCES

[1] M. Homaei, M. Tarif, P. G. Rodríguez, A. Caro, and M. Ávila, "Causal Digital Twins for cyber-physical security in water systems," *Mach. Learn. Appl.*, p. 100824, 2025.
[2] M. Homaei, M. Tarif, P. G. Rodríguez, M. Ávila, and A. Caro, "Adaptive multi-domain uncertainty quantification for digital twin water forecasting," *Mach. Learn. Appl.*, p. 100812, 2025.
[3] M. Homaei, M. Tarif, P. G. Rodriguez, A. Caro, and M. Avila, "Conformal prediction-driven adaptive sampling for digital water twins," *ICT Express*, 2026.
[4] M. Tarif, E. Fasani, and D. Quaglia, "VISAR: Intelligent video frame selection for agricultural 3D plant reconstruction," in *Proc. IEEE MetroAgriFor*, pp. 99–104, 2025.
[5] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. ICLR*, 2017.
[6] P. Veličković *et al.*, "Graph attention networks," in *Proc. ICLR*, 2018.
[7] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Proc. NeurIPS*, vol. 30, 2017.
[8] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" in *Proc. ICLR*, 2019.
[9] F. Scarselli *et al.*, "The graph neural network model," *IEEE Trans. Neural Netw.*, vol. 20, no. 1, pp. 61–80, 2009.
[10] M. M. Bronstein *et al.*, "Geometric deep learning: Going beyond Euclidean data," *IEEE Signal Process. Mag.*, vol. 34, no. 4, pp. 18–42, 2017.
[11] P. W. Battaglia *et al.*, "Relational inductive biases, deep learning, and graph networks," *arXiv:1806.01261*, 2018.
[12] J. Zhou *et al.*, "Graph neural networks: A review of methods and applications," *AI Open*, vol. 1, pp. 57–81, 2020.
[13] Z. Wu *et al.*, "A comprehensive study on graph neural networks," *IEEE Trans. NNLS*, vol. 32, no. 1, pp. 4–24, 2021.
[14] J. Pearl, *Causality*, 2nd ed. Cambridge Univ. Press, 2009.
[15] J. Peters, J. M. Mooij, and B. Schölkopf, *Elements of Causal Inference*. MIT Press, 2017.
[16] B. Schölkopf *et al.*, "Toward causal representation learning," *Proc. IEEE*, vol. 109, no. 5, pp. 612–634, 2021.
[17] C. W. J. Granger, "Investigating causal relations by econometric models," *Econometrica*, vol. 37, no. 3, pp. 424–438, 1969.
[18] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search*, 2nd ed. MIT Press, 2000.
[19] S. Zhu, I. Ng, and Z. Chen, "Causal discovery with reinforcement learning," in *Proc. ICLR*, 2020.
[20] X. Zheng, B. Aragam, P. K. Ravikumar, and E. P. Xing, "DAGs with NO TEARS," in *Proc. NeurIPS*, vol. 31, 2018.
[21] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate," *J. R. Stat. Soc. B*, vol. 57, no. 1, pp. 289–300, 1995.
[22] J. D. Storey and R. Tibshirani, "Statistical significance for genome-wide studies," *Proc. Natl. Acad. Sci.*, vol. 100, no. 16, pp. 9440–9445, 2003.
[23] Z. Ying *et al.*, "GNNExplainer: Generating explanations for GNNs," in *Proc. NeurIPS*, vol. 32, 2019.
[24] D. Luo *et al.*, "Parameterized explainer for graph neural network," in *Proc. NeurIPS*, vol. 33, pp. 19620–19631, 2020.
[25] H. Yuan *et al.*, "Explainability in GNNs: A taxonomic survey," *IEEE Trans. PAMI*, vol. 45, no. 5, pp. 5782–5799, 2022.
[26] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. NeurIPS*, vol. 30, 2017.
[27] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in *Proc. KDD*, pp. 1135–1144, 2016.
[28] A. B. Arrieta *et al.*, "Explainable Artificial Intelligence: Concepts, taxonomies, opportunities," *Inf. Fusion*, vol. 58, pp. 82–115, 2020.
[29] Z. C. Lipton, "The mythos of model interpretability," *Queue*, vol. 16, no. 3, pp. 31–57, 2018.
[30] R. Guidotti *et al.*, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 1–42, 2018.
[31] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE ICCV*, pp. 2980–2988, 2017.
[32] A. P. Mathur and N. O. Tippenhauer, "SWaT: A water treatment testbed for ICS security," in *Proc. CySWater*, pp. 31–36, 2016.
[33] J. Goh, S. Adepu, K. N. Junejo, and A. Mathur, "A dataset to support research in the design of secure water treatment systems," in *Proc. CRITIS*, pp. 88–99, 2016.
[34] R. Taormina *et al.*, "The Battle of the Attack Detection Algorithms," *J. Water Resour. Plan. Manag.*, vol. 144, no. 4, p. 04018012, 2018.
[35] R. Taormina and S. Galelli, "Deep-learning approach to detection and localisation of cyber-physical attacks on water distribution systems," *J. Water Resour. Plan. Manag.*, vol. 144, no. 10, p. 04018065, 2018.

[36] S. Adepu and A. Mathur, "An investigation into the response of a water treatment facility to cyber attacks," in *Proc. HASE*, pp. 141–148, 2016.

[37] C. M. Ahmed, V. R. Palleti, and A. P. Mathur, "Challenges in ML-based approaches for real-time anomaly detection in ICS," in *Proc. DEBS*, pp. 37–48, 2020.

[38] A. A. Abokifa *et al.*, "Real-time identification of cyber-physical attacks on water distribution systems," *J. Water Resour. Plan. Manag.*, vol. 145, no. 1, p. 04018089, 2019.

[39] M. Housh and Z. Ohar, "Model-based approach for cyber-physical attack detection in water distribution systems," *Water Res.*, vol. 193, p. 116875, 2021.

[40] M. Aghashahi *et al.*, "Water systems cyber-attack benchmark datasets for IDS," in *Proc. ICDATA*, 2018.

[41] M. Kravchik and A. Shabtai, "Detecting cyber attacks in ICS using convolutional neural networks," in *Proc. CPS-SPC*, pp. 72–83, 2018.

[42] J. Inoue *et al.*, "Anomaly detection for a water treatment system using unsupervised ML," in *Proc. ICDMW*, pp. 1058–1065, 2017.

[43] G. Bernieri *et al.*, "Monitoring system performance to enhance cyber-physical security," *IEEE Syst. J.*, vol. 14, no. 1, pp. 1451–1462, 2019.

[44] D. Li *et al.*, "MAD-GAN: Multivariate anomaly detection with GANs," in *Proc. ICANN*, pp. 703–716, 2019.

[45] J. Audibert *et al.*, "USAD: UnSupervised Anomaly Detection on multivariate time series," in *Proc. KDD*, pp. 3395–3404, 2020.

[46] F. Pasqualetti, F. Dörfler, and F. Bullo, "Attack detection and identification in cyber-physical systems," *IEEE Trans. Autom. Control*, vol. 58, no. 11, pp. 2715–2729, 2013.

[47] Y. Shoukry *et al.*, "PyCRA: Physical challenge-response authentication for active sensors," in *Proc. CCS*, pp. 1004–1015, 2015.

[48] C. Feng, T. Li, and D. Chana, "Multi-level anomaly detection in ICS via package signatures and LSTM," in *Proc. DSN*, pp. 261–272, 2019.

[49] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–58, 2009.

[50] L. Ruff *et al.*, "A unifying review of deep and shallow anomaly detection," *Proc. IEEE*, vol. 109, no. 5, pp. 756–795, 2021.

[51] M. Goldstein and S. Uchida, "A comparative evaluation of unsupervised anomaly detection algorithms," *PLOS ONE*, vol. 11, no. 4, p. e0152173, 2016.

[52] A. Deng and B. Hooi, "Graph neural network-based anomaly detection in multivariate time series," in *Proc. AAAI*, vol. 35, no. 5, pp. 4027–4035, 2021.

[53] H. Zhao *et al.*, "Multivariate time-series anomaly detection via graph attention network," in *Proc. ICDM*, pp. 841–850, 2020.

[54] M. Grieves and J. Vickers, "Digital twin: Mitigating unpredictable, undesirable emergent behavior," in *Transdisciplinary Perspectives on Complex Systems*, pp. 85–113, Springer, 2017.

[55] F. Tao *et al.*, "Digital twins and cyber–physical systems toward smart manufacturing," *Engineering*, vol. 5, no. 4, pp. 653–661, 2019.

[56] Y. Li *et al.*, "Diffusion convolutional recurrent neural network," in *Proc. ICLR*, 2018.

[57] B. Yu, H. Yin, and Z. Zhu, "Spatio-temporal graph convolutional networks," in *Proc. IJCAI*, pp. 3634–3640, 2018.

[58] Z. Wu *et al.*, "Connecting the dots: Multivariate time series forecasting with GNNs," in *Proc. KDD*, pp. 753–763, 2020.

[59] D. Zügner, A. Akbarnejad, and S. Günnemann, "Adversarial attacks on neural networks for graph data," in *Proc. KDD*, pp. 2847–2856, 2018.

[60] B. Stojanovic, K. Hofer-Schmitz, and U. Kleb, "Explainable anomaly detection for ICS security," *Appl. Sci.*, vol. 11, no. 10, p. 4450, 2021.