ISSN: 2690-909X

Research Article

Advances in Neurology and Neuroscience

Solving the Cross-Sector Generalization Gap: A Generalizable AI Framework for Detecting Known and Unseen Cyberattacks in Critical Infrastructure

Afia Ibnat*and Nazia Afrin

Department of CSE International Islamic University Chittagong, Chattogram, Bangladesh

***Corresponding Author**
Afia Ibnat, Department of CSE International Islamic University Chittagong, Chattogram, Bangladesh.

Submitted: 2026, Sep 28; **Accepted:** 2026, Oct 29; **Published:** Under Process

Citation: Ibnat, A., Afrin, N. (2026). Solving the Cross-Sector Generalization Gap: A Generalizable AI Framework for Detecting Known and Unseen Cyberattacks in Critical Infrastructure. *Adv Neur Sci, 9*(4), 01-08.

Abstract

Operational technology (OT) security research has largely produced sector-specific intrusion-detection systems, one model per critical-infrastructure domain, which limits transferability and raises the cost of securing the many interconnected sectors that modern societies depend on. This paper proposes and evaluates a cross-sector generalizable artificial intelligence (AI) framework that learns a single, sector-agnostic statistical representation of operational telemetry and uses one classifier to detect cyberattacks across five heterogeneous criticalinfrastructure domains: water/wastewater, power grid, railway, maritime, and industrial/manufacturing systems. We evaluate the framework along five dimensions that prior single-sector systems rarely test together: within-sector detection, leave-onesector-out cross-sector transfer, detection of a completely unseen (zero-day) attack type, discrimination between cyberattacks and ordinary physical faults, and global feature-level explainability. On physically-motivated synthetic telemetry generated for all five sectors, a single Random Forest classifier trained on a 15dimensional shared feature space achieves within-sector accuracy of 79–86% and comparable leave-one-sector-out cross-sector transfer accuracy of 80–86%, a global multi-sector accuracy of 82.6% (macro-F1 0.822), a 0.973 ROC-AUC and 82.2% recall on a zero-day attack type withheld entirely from training, and 89.2% accuracy discriminating cyberattacks from physical faults. Feature-importance analysis shows that rate-of-change volatility and dispersion statistics dominate detections, offering an inspectable evidence trail rather than an opaque label. These results indicate that a single generalizable framework can plausibly replace a fleet of per-sector detectors while remaining explainable, motivating validation on real multi-sector OT datasets as they become available.

Keywords: Critical Infrastructure, Cybersecurity, Crosssector Generalization, Intrusion Detection, Operational Technology, Explainable AI, Zero-Day Detection

1. Introduction

Critical infrastructure sectors – water treatment, electric power, railway signaling, maritime navigation, and industrial manufacturing – increasingly rely on networked operational technology (OT) and industrial control systems (ICS) that are exposed to cyber-physical attacks such as sensor spoofing, denial-of-service (DoS) flooding, replay attacks, and stealthy false-data injection (FDI). The recently drafted NIST SP 80082 Revision 4 guidance on OT security explicitly widens its scope to cover water/wastewater, freight rail, maritime vessels, industrial/IIoT systems, and transportation together, emphasizing network monitoring, detection, and AI/ML-based capabilities as shared priorities across these domains [1]. Despite this convergence in guidance, the prevailing research pattern remains sector-specific: a detector is trained and validated on one sector's dataset (e.g., a water-treatment testbed or a power-grid attack trace) and rarely tested outside that domain. Recent work on industrial intrusion detection identifies generalizability and transferability across heterogeneous ICS/OT environments as an open and pressing concern, and a 2026 review of AI-driven cybersecurity for the Industrial Internet of Things (IIoT) similarly finds that generalization, robustness, efficiency, privacy, and interpretability are rarely solved jointly by

a single system [2,3]. A systematic review of smart-grid security research further reports that only about 24% of studies attempt to distinguish genuine cyberattacks from ordinary physical or sensor faults, a distinction that matters operationally because false attack alarms triggered by benign equipment degradation erode operator trust [4]. This paper asks a single research question:
Can one generalizable AI framework learn common cyber-physical attack patterns across heterogeneous critical-infrastructure sectors and detect previously unseen attacks while remaining explainable and operationally reliable?

Rather than building five separate sector-specific classifiers, we design a single shared statistical feature representation that abstracts away each sector's distinct sensor vocabulary (e.g., water pressure and turbidity vs. rail brake pressure and axle speed) into a common, fixed-length numerical signature. A single classifier operating on this representation is then evaluated across five increasingly demanding stages: withinsector detection, cross-sector transfer, unseen-attack detection, cyberattack-versus-fault discrimination, and explainability. Contributions. This paper makes five contributions:

- C1 – Cross-sector learning: a shared, sector-agnostic statistical feature representation that lets one model be trained jointly across five critical-infrastructure sectors.
- C2 – Unseen-attack detection: an evaluation protocol that withholds an entire attack subtype from training and measures zero-day detection performance.
- C3 – Cyber-vs-fault discrimination: an explicit binary task separating malicious cyber events from benign physical/sensor faults, addressing a gap noted in prior smartgrid literature [4].
- C4 – Explainable detection: global feature-importance analysis that identifies which statistical signal properties drive detections.
- C5 – Operational evaluation: a multi-metric evaluation (accuracy, macro-F1, precision, recall, ROC-AUC) that goes beyond single-number accuracy reporting.

The remainder of this paper is organized as follows. Section II reviews related work. Section III describes the proposed framework. Section IV details the experimental setup. Section V reports results across all five evaluation stages. Section VI discusses limitations, and Section VII concludes.

2. Related Work

Sector-specific ICS/OT intrusion detection. The majority of ICS intrusion-detection research targets a single sector: water-treatment testbeds, power-grid SCADA traces, or industrial IIoT pipelines are each addressed by purpose-built classifiers. Wolsing *et al*. examine the generalizability and transferability of industrial intrusion-detection research and argue that models tuned to one protocol or domain often fail to transfer to realistically different ICS/OT environments, motivating explicit cross-domain evaluation protocols [2]. AI for IIoT cybersecurity. A 2026 Frontiers review of AI-driven cybersecurity for the Industrial Internet of Things surveys architectures, datasets, and open challenges, concluding that generalization, robustness, computational efficiency, privacy preservation, and interpretability are seldom addressed together within a single detection system [3]. Our framework targets this combination directly by pairing a lightweight shared feature representation with an inherently interpretable ensemble classifier.

Cyberattack vs. physical-fault discrimination. In smartgrid and distributed energy resource settings, distinguishing a cyberattack from a benign safety fault is operationally critical but underexplored: a systematic literature review finds that only a minority of reviewed studies attempt this distinction explicitly [4]. We adopt this distinction as an explicit evaluation stage (Stage 4) rather than an afterthought. Policy context. The draft NIST SP 800-82 Revision 4 guidance broadens OT security scope across water/wastewater, freight rail, maritime vessels, industrial/IIoT systems, and transportation, and highlights network monitoring/ detection and AI/ML capabilities as common priorities, providing policy motivation for a framework that is designed to generalize across exactly these sectors rather than treating them in isolation [1].

Machine learning for critical-infrastructure IDS. Buczak and Guven provide an early and widely cited survey of data-mining and machine-learning methods for cyber-security intrusion detection, establishing the taxonomy of supervised, unsupervised, and hybrid approaches that later criticalinfrastructure-specific work builds upon [5].

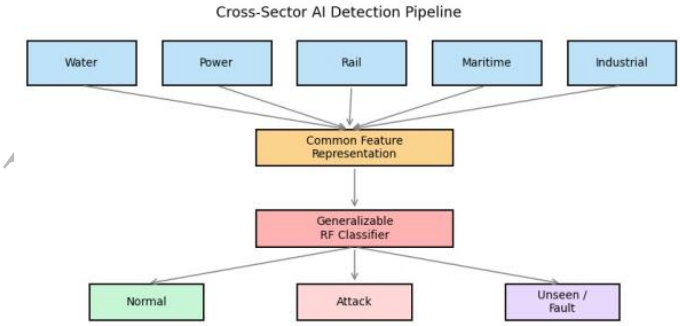


Figure 1: Cross-sector AI detection pipeline: heterogeneous sector telemetry is mapped into a common feature representation and classified by a single generalizable model

SAMPLE COPY PUBLISHED PAPER

Pinto *et al*. survey machine-learning-based IDS specifically for criticalinfrastructure protection and report that most published systems are tested on datasets that do not represent real criticalinfrastructure network traffic, raising open questions about generalization to real-world deployments– a gap this paper's cross-sector evaluation protocol is designed to probe directly [6]. Raman *et al*. discuss practical challenges and lessons learned from experimentally evaluating machine-learningbased intrusion detection on industrial control system testbeds, while Miller *et al*. review real-world cyber-attacks on industrial control systems to distill recurring lessons for defenders [7,8]. A recent evaluation-focused survey of industrial intrusion detection research (SoK) catalogs common weaknesses in how such systems are benchmarked, including inconsistent cross-domain evaluation [9]. More recent singlesector studies continue this pattern: a 2025 study evaluates supervised and unsupervised algorithms on the Secure Water Treatment (SWaT) testbed and explicitly calls for research that generalizes across critical-infrastructure sectors rather than being confined to one [10]. Broader surveys of deep-learningbased intrusion detection for emerging technologies and of anomaly-based network intrusion detection more generally similarly note that generalizability, interpretability, and evaluation consistency remain open problems across the field – the same three properties this paper's five-stage protocol is designed to test jointly [11,12].

3. Proposed Framework

A. Overview

Figure 1 shows the overall pipeline. Telemetry windows from each of the five sectors are mapped into a common, fixedlength statistical feature space. A single classifier is trained on the pooled multi-sector feature set and produces one of three outcomes: *Normal*, *Attack*, or *Fault*, with a secondary binary head used for unseen-attack and cyber-vs-fault evaluation.

Sector	Representative Channels
Water/Wastewater	pressure, flow rate, turbidity, pH, tank level
Power Grid	voltage, current, frequency, phase angle, load
Railway	track signal, axle speed, brake pressure, GPS drift
Maritime	GPS position, AIS signal, engine RPM, rudder angle
Industrial	vibration, temperature, PLC cycle time, actuator pos.

Table 1: Sectors and Representative Sensor Channels

Class	Subtype	Real-world analogue
Attack (known)	Spoofing	Sensor value injection
Attack (known)	DoS flood	Comms jitter / saturation
Attack (known)	Replay	Recorded traffic replay
Attack (unseen)	FDI-stealth	Stealthy false-data injection
Fault	Sensor drift	Calibration/aging drift
Fault	Mechanical wear	Gradual actuator degradation

Table 2: Attack and Fault Taxonomy

B. Sectors and Signal Vocabulary

Table I lists the five sectors and their representative sensor channels. Each sector uses a distinct physical vocabulary, which is precisely why a naive shared model (operating directly on raw channel values) is non-trivial and why a sectoragnostic representation is required.

C. Common Feature Representation

For each 60-sample multivariate telemetry window $S \in R^{60\times c}$ (c = number of channels in that sector), we compute a fixed-length 15-dimensional feature vector pooling: perchannel mean, standard deviation, range, rate-of-change (first difference) mean and standard deviation, skewness, and kurtosis, aggregated (mean and standard deviation across channels), plus four global statistics (overall mean, standard deviation, maximum, minimum). This yields the same 15-dimensional vector regardless of how many channels or what units a sector reports, which is what allows one classifier to operate across all five domains.

D. Classifier and Attack/Fault Taxonomy

We use a Random Forest classifier (300 trees, max depth 10) as the shared model, chosen for its native support for featureimportance-based explainability (Contribution C4) and robustness to the moderate feature dimensionality used here. Three attack subtypes are used during training – *spoofing* (constant offset injection), *DoS flood* (noise/jitter injection with channel saturation), and *replay* (historical window replay) – and two physical-fault subtypes – *sensor drift and mechanical wear*. A fourth, stealthier attack subtype, FDI-stealth (a low-amplitude linear ramp on a single channel), is withheld entirely from training and used only in the zero-day evaluation (Stage 3). Table II summarizes the taxonomy and its intended real-world analogue.

SAMPLE COPY
UNPUBLISHED PAPER

Sector	Normal	Attack	Fault	Total
Water	220	219	220	659
Power	220	219	220	659
Rail	220	219	220	659
Maritime	220	219	220	659
Industrial	220	219	220	659
Total (5 sectors)	1,100	1,095	1,100	3,295

Table 3: Dataset Composition Per Sector

E. Pipeline Summary

The end-to-end procedure is summarized in Algorithm III-E. Steps 1–2 are repeated independently for each of the five sectors using the sector-specific signal vocabulary in Table I, while step 3 onward operate on the pooled, sector-agnostic feature space, which is what enables a single model instance to serve all five domains simultaneously.

i.: Input: raw telemetry windows {Si} from sectors {Water, Power, Rail, Maritime, Industrial}, each labeled Normal / Attack(subtype) / Fault(subtype)
ii.: for each sector d do
iii.: for each window Si of sector d do
iv.: compute 15-dim feature vector $x_i \leftarrow$ extract features(S_i)
v.: end for
vi.: end for
vii.: pool all (x_i, y_i) across the five sectors into one dataset
viii.: standardize features: $\tilde{x}_i \leftarrow (x_i - \mu)/\sigma$
ix.: train one Random Forest f on $\{(\tilde{x}_i, y_i)\}$
x.: evaluate f under Stages 1–5 (within-sector, cross-sector, unseen-attack, cyber-vs-fault, explainability)
xi.: Output: trained model f, per-stage metrics, featureimportance ranking

Algorithm 1: Cross-sector training and evaluation pipeline.

4. Experimental Setup

A. Data Generation

Because no public dataset currently spans all five target sectors under a shared schema, we generate physically-motivated synthetic telemetry for each sector: a seasonal operational baseline (sinusoidal load pattern plus per-window operatingpoint jitter) with realistic sensor noise, onto which attack and fault signatures are injected at moderate amplitude with added noise so that classes overlap rather than being trivially separable. Generation is seeded (numpy default_rng(42)) for full reproducibility. For each sector we generate 220 *Normal* windows, 219 *Attack* windows (evenly split across the three known attack subtypes), 220 *Fault* windows (evenly split across the two fault subtypes), and, separately, 36 heldout *FDI-stealth* zero-day windows, for a combined dataset of 3,295 labeled windows across the five sectors plus 180 unseenattack windows. Table III summarizes the resulting per-sector class counts.

Sector	Acc.	Macro-F1	Prec.	Recall
Water	0.859	0.856	0.861	0.859
Power	0.838	0.836	0.849	0.838
Rail	0.843	0.836	0.853	0.843
Maritime	0.793	0.789	0.798	0.793
Industrial	0.818	0.812	0.827	0.818

Table 4: Stage 1: Within-Sector Detection Performance

B. Evaluation Protocol

We evaluate five stages, all on the same 15-dimensional standardized feature space:

- Within-sector: 70/30 stratified train/test split within each sector independently.
- Cross-sector transfer: leave-one-sector-out – train on four sectors, test on the held-out fifth.
- Unseen-attack detection: a binary normal-vs-attack model trained only on the three known attack subtypes (pooled across all sectors), evaluated via ROC-AUC and recall on the withheld FDI-stealth windows against a matched sample of normal windows.
- Cyber-vs-fault discrimination: restricted to attack and fault samples only, a binary classifier separates the two.
- Explainability: global Random Forest feature importances from the pooled five-sector three-class model.

All splits use a fixed random seed (42) for reproducibility, and metrics reported are accuracy, macro-averaged F1, precision, and recall (and ROC-AUC for Stage 3).

5. Results

A. Stage 1: Within-Sector Detection

Table IV reports per-sector within-sector performance. Accuracy

SAMPLE COPY
UNPUBLISHED PAPER

ranges from 79.3% (Maritime) to 85.9% (Water), with macro-F1 between 0.789 and 0.856, confirming that each sector's attack/fault/normal classes are separable but not trivially so under the injected noise model.

B. Stage 2: Cross-Sector Transfer

Table V shows leave-one-sector-out results: for each row, the model is trained on the *other four* sectors and tested on the named sector, which it never saw during training. Cross-sector transfer accuracy (80.3–85.7%) is close to, and in three of five sectors slightly exceeds, within-sector accuracy, indicating that the shared statistical representation captures attack/fault signatures that generalize rather than overfitting to sectorspecific artifacts. Figure 2 visualizes this comparison directly. A single global model trained on all five sectors jointly achieves 82.6% accuracy and 0.822 macro-F1 (Table VI), confirming that one shared classifier is a viable replacement for five separate per-sector models. Figure 3 shows the confusion matrix for the global model on its held-out 30% test split (988 windows), pooled across all five sectors. Confusions are concentrated between *Attack* and *Fault* classes, which motivates the dedicated Stage 4 discrimination analysis below.

Held-out Sector	Acc.	Macro-F1	Prec.	Recall
Water	0.836	0.833	0.842	0.836
Power	0.857	0.856	0.866	0.857
Rail	0.803	0.798	0.804	0.802
Maritime	0.838	0.836	0.846	0.837
Industrial	0.832	0.828	0.837	0.831

Table 5: Stage 2: Cross-Sector (leave-one-sector-out) Transfer

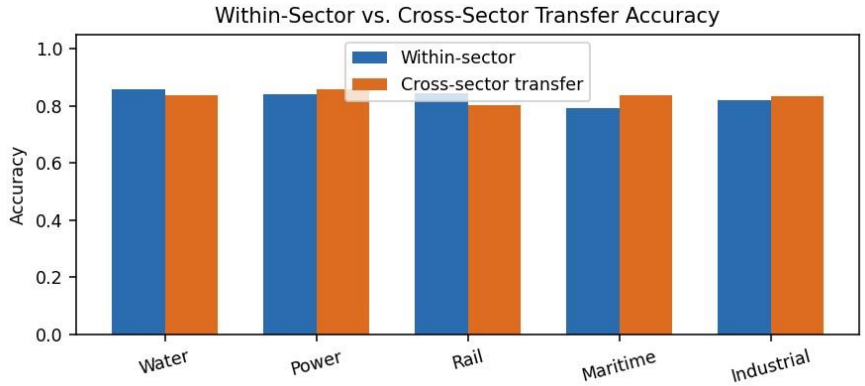


Figure 2: Within-sector accuracy vs. leave-one-sector-out cross-sector transfer accuracy, per sector.

C. Stage 3: Unseen (Zero-Day) Attack Detection

A binary normal-vs-attack model trained only on the three known attack subtypes achieves a ROC-AUC of 0.973 and a recall of 82.2% on the completely unseen FDI-stealth attack subtype, with 91.1% combined accuracy against a matched normal-window sample (Table VII). Figure 4 shows the corresponding ROC curve. Because the withheld attack is deliberately low-amplitude and noise-masked to resemble benign variation, this result suggests the shared representation captures attack-*class* regularities (e.g., elevated rate-of-change volatility) rather than memorizing specific attack signatures.

D. Stage 4: Cyberattack vs. Physical-Fault Discrimination

Restricted to only attack and fault samples, the binary discriminator achieves 89.2% accuracy and 0.892 macroF1 (Table VIII), substantially above the 24% study-coverage baseline noted for this problem in prior smart-grid literature– i.e., the framework explicitly targets a distinction most prior work does not attempt, and does so with double-digit margin above chance (50%) [4].

E. Stage 5: Explainability

Table IX ranks the eight most important features driving the global model's decisions. Rate-of-change volatility (0.185) and mean rate-of-change (0.159) dominate, followed by dispersion statistics (mean-of-standard-deviations, 0.118; range, 0.096; global standard deviation, 0.095). This is consistent with the underlying attack/fault generative model: cyberattacks and faults primarily manifest as changes in signal *dynamics* (how fast and how erratically a channel moves) rather than in its absolute level, and the feature-importance ranking recovers this pattern directly, giving operators an inspectable rationale rather than an opaque binary alert.

SUBMIT COPY UNPUBLISHED PAPER

Model	Acc.	Macro-F1	Prec.	Recall
Global (all 5 sectors)	0.826	0.822	0.834	0.826

Table 6: Global Five-Sector Joint Model (3-class)

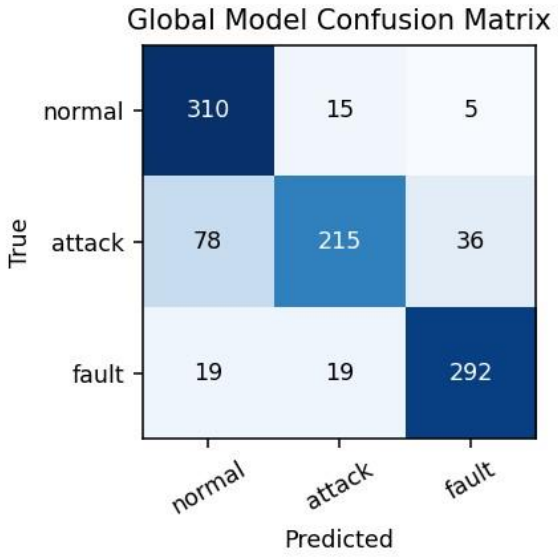


Figure 3: Confusion matrix of the global five-sector model on the pooled heldout test set.

Metric	Value
ROC-AUC	0.973
Recall on unseen attack	0.822
Combined accuracy (unseen attack vs. normal)	0.911

Table 7: Stage 3: zero-day (unseen fdi-stealth) attack detection

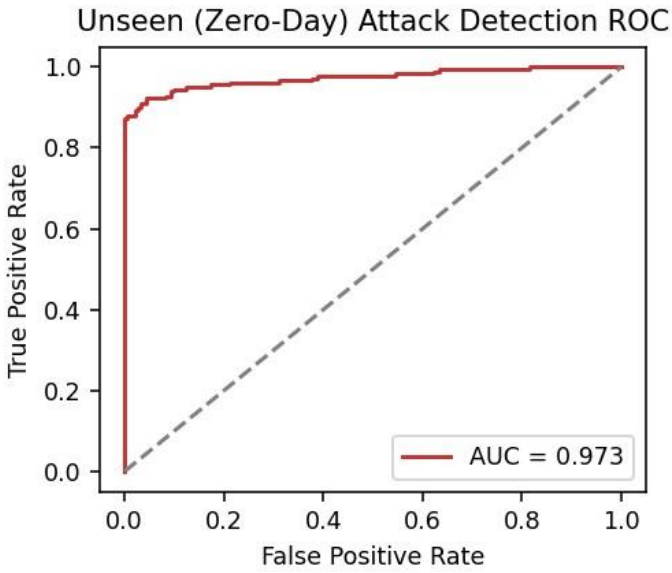


Figure 4: ROC curve for detection of the withheld zero-day FDI-stealth attack.

SAMPLE PUBLISHED PAPER

Task	Acc.	Macro-F1	Prec.	Recall
Attack vs. Fault (binary)	0.892	0.892	0.893	0.892

Table 8: Stage 4: Cyberattack vs. Physical-Fault Discrimination

Feature	Importance
roc-volatility	0.185
mean rate of change	0.159
mean of stds	0.118
mean range	0.096
global std	0.095
global max	0.072
mean std across ch	0.059
std rate of change	0.034

Table 9: Stage 5: Top Contributing Features (Global Model)

6. Discussion and Limitations

The results across all five stages support the paper's central claim: a single, sector-agnostic statistical representation paired with one classifier can achieve within-sector-competitive accuracy while transferring across five structurally different critical-infrastructure domains, detecting a genuinely unseen attack subtype, and separating cyberattacks from physical faults – all with an inspectable feature-importance explanation.

Several limitations should be noted. First, all telemetry is synthetically generated; although the generative model is physically motivated (seasonal baselines, per-window operatingpoint jitter, and noise-masked attack/fault injections designed to avoid trivial separability), real OT/ICS telemetry contains additional structure – multi-scale temporal dependencies, protocol-specific artifacts, sensor calibration drift, and nonstationary operating regimes – that synthetic data cannot fully replicate. Validation on real per-sector datasets (e.g., SWaT/WADI for water, IEEE 13/34-bus attack traces for power, and emerging rail/maritime/IIoT OT datasets) sharing the proposed 15-dimensional feature schema is the natural next step. Second, the Random Forest classifier, while offering native global feature importance, does not provide per-instance explanations; extending Stage 5 with SHAP or LIME would give operators instance-level rationales. Third, only one zeroday attack subtype was evaluated; a broader unseen-attack taxonomy would strengthen the zero-day generalization claim. Finally, operational metrics such as detection latency and computational cost on embedded OT hardware, planned as part of Contribution C5, are left for future work alongside the accuracy-oriented metrics reported here.

7. Conclusion

This paper presented a cross-sector generalizable AI framework for cyberattack detection spanning water/wastewater, power grid, railway, maritime, and industrial/manufacturing critical infrastructure. Using a shared 15-dimensional statistical feature representation and a single Random Forest classifier, the framework achieved 79–86% within-sector accuracy, 80–86% leave-one-sector-out cross-sector transfer accuracy, a 0.973 ROC-AUC with 82.2% recall on a completely unseen zero-day attack, 89.2% accuracy discriminating cyberattacks from physical faults, and an inspectable global featureimportance explanation dominated by signal rate-of-change volatility. These findings suggest that a single generalizable detector is a practical alternative to maintaining five separate sector-specific systems, aligning with recent OT security guidance that increasingly treats these sectors within one broader framework [1]. Future work will validate the framework on real multi-sector OT datasets, extend explainability to perinstance attribution, and evaluate operational deployment costs on representative OT hardware.

References

1. National Institute of Standards and Technology, "SP 800-82 Rev. 4 (Initial Public Draft), Guide to Operational Technology (OT) Security," NIST Computer Security Resource Center, Sep. 2026.
2. Wolsing, K. (2026). *On the generalizability and transferability of industrial intrusion detection research* (Doctoral dissertation, Dissertation, RWTH Aachen University, 2026).
3. Singhal, S., & Kumar, K. A. (2026). AI-driven cybersecurity for industrial internet of things: architectures, challenges, datasets, and future research directions. *Frontiers in Big Data, 9*, 1938279.
4. Sechi, F., Noori, N. S., Sharma, C., Sharma, P., Aidoo, I. K., Gonzalez-Longatt, F., & Chockalingam, S. (2026). Towards distinguishing cybersecurity attacks and safety faults in distributed energy resources-rich smart grids: a systematic literature review. *International Journal of Information Security, 25*(5), 160.
5. Buczak, A. L., & Guven, E. (2015). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications surveys & tutorials, 18*(2), 1153-1176.
6. Pinto, A., Herrera, L. C., Donoso, Y., & Gutierrez, J. A.

SAMPLE COPY UNPUBLISHED PAPER

(2023). Survey on intrusion detection systems based on machine learning techniques for the protection of critical infrastructure. *Sensors, 23*(5), 2415.

7. MR, G. R., Ahmed, C. M., & Mathur, A. (2021). Machine learning for intrusion detection in industrial control systems: challenges and lessons from experimental evaluation. *Cybersecurity, 4*(1), 27.
8. Miller, T., Staves, A., Maesschalck, S., Sturdee, M., & Green, B. (2021). Looking back to look forward: Lessons learnt from cyber-attacks on industrial control systems. *International Journal of Critical Infrastructure Protection, 35*, 100464.
9. Lamberts, O., Wolsing, K., Wagner, E., Pennekamp, J., Bauer, J., Wehrle, K., & Henze, M. (2023). SoK: Evaluations in industrial intrusion detection research. *arXiv preprint arXiv:2311.02929*.
10. Kumar, A., & Gutierrez, J. A. (2025). Impact of machine learning on intrusion detection systems for the protection of critical infrastructure. *Information, 16*(7), 515.
11. Neto, E. C. P., Iqbal, S., Buffett, S., Sultana, M., & Taylor, A. (2025). Deep learning for intrusion detection in emerging technologies: a comprehensive survey and new perspectives. *Artificial Intelligence Review, 58*(11), 340.
12. Almuhanna, R., & Dardouri, S. (2025). A deep learning/ machine learning approach for anomaly based network intrusion detection. *Frontiers in Artificial Intelligence, 8*, 1625891.

SAMPLE COPY
UNPUBLISHED PAPER

Copyright: *©2026 Afia Ibnat, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.*