

Article

PhySec-Edge: A Hybrid Physics-Informed and Edge AI Framework for Anomaly Detection in Industrial IoT Sensor Networks

Dalibor Radovanovic *, Nikola Savanovic, Petar Kresoja, Jelena Janackovic and Teodor Petrovic

Faculty of Informatics and Computing, Singidunum University, Danijelova 32, 11000 Belgrade, Serbia; nsavanovic@singidunum.ac.rs (N.S.); pkresoja@singidunum.ac.rs (P.K.); jjanackovic@singidunum.ac.rs (J.J.); tpetrovic@singidunum.ac.rs (T.P.)

* Correspondence: dradovanovic@singidunum.ac.rs

Abstract

Industrial Internet of Things (IIoT) deployments face a security challenge that neither physics-based nor AI-based anomaly detection addresses alone: physics models are adversarially robust but miss behavioral attacks that remain within physical bounds, while AI models detect behavioral anomalies but are vulnerable to adversarial evasion and blind to physical sensor spoofing. This paper proposes PhySec-Edge, a hybrid framework integrating a Physics Validation Engine (PVE) with a multi-model Edge AI Detection Engine (EADE) in a layered, residual-sharing architecture. The PVE applies process model residuals, Kalman filter state estimation, cross-sensor consistency checks, and temporal gradient validation to generate physics-grounded anomaly signals. The EADE is designed around LSTM temporal detection, variational autoencoder reconstruction analysis, and graph neural network process monitoring augmented with PVE residuals; the current evaluation uses computationally tractable proxy implementations to provide a conservative lower bound on the benefits of residual sharing. Randomized smoothing is applied under bounded perturbation assumptions to improve adversarial robustness. PhySec-Edge is evaluated in a controlled synthetic IIoT setting parameterized using SWaT-inspired structural and statistical assumptions, comprising 9875 samples across seven attack classes. Across five random seeds, the hybrid framework achieves mean precision = 0.789 ± 0.004 , recall = 0.808 ± 0.003 , F1 = 0.798 ± 0.003 , and FPR = $5.0\% \pm 0.0\%$, compared to F1 = 0.654 ± 0.006 /FPR = 24.0% for the physics-only baseline and F1 = 0.774 ± 0.003 /FPR = 5.0% for the AI-only baseline. An ablation study identifies residual augmentation as the primary individual improvement mechanism ($\Delta F1 = +0.017$), while the full hybrid configuration achieves a combined gain of $\Delta F1 = +0.025$ over the AI-only baseline. Critical hybrid advantages appear on adversarial evasion ($+0.15$ F1) and firmware implant ($+0.17$ F1), the two attack classes where neither layer alone is sufficient. A preliminary feasibility check on an Edge-IIoTset-inspired benchmark confirms that the architectural advantage pattern generalizes across dataset structures. Gateway latency analysis confirms compatibility with soft real-time industrial monitoring constraints.

Keywords: IIoT; industrial IoT; anomaly detection; physics-informed security; edge AI; federated learning; sensor security; hybrid framework; adversarial robustness; Industry 4.0

Academic Editors: Lorenzo Mucchi, Laura Carnevali, Stefano Caputo and Leonardo Scommegna

Received: 16 June 2026

Revised: 11 July 2026

Accepted: 15 July 2026

Published: 17 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The Industrial Internet of Things (IIoT) has become the foundational sensing infrastructure of Industry 4.0, enabling real-time monitoring, predictive maintenance, and autonomous process control across manufacturing, energy, transportation, and healthcare [1]. As of 2023, the number of IIoT-connected endpoints in critical infrastructure sectors exceeds 17 billion globally, and this figure is projected to double by 2030 [2]. The security of this infrastructure is increasingly critical: IIoT sensors generate data upon which safety-critical decisions are made, and their compromise can result in physical damage, production loss, and cascading infrastructure failures—as demonstrated by the Stuxnet incident [3] and a growing body of documented ICS/OT attack campaigns [2,4].

The anomaly detection problem in IIoT is fundamentally harder than in consumer IoT or enterprise IT for three interconnected reasons. First, legacy devices operating fieldbus protocols—Modbus, DNP3, Profibus, OPC-UA—lack authentication and encryption, meaning that network-level intrusion detection has limited visibility into the semantic content of sensor readings [2]. Second, the physical process context creates attacks—physical sensor spoofing, signal injection, multi-sensor manipulation—that are invisible to purely behavioral detectors but detectable through physical process models enforcing thermodynamic, hydraulic, or electrical constraints [5]. Third, well-resourced adversaries craft attacks specifically designed to evade AI-based detection while remaining within physically plausible bounds, a capability demonstrated at scale in the smart home context [6] and increasingly documented in ICS/SCADA environments [2].

Current anomaly detection approaches address this challenge incompletely. Physics-based approaches detect physical constraint violations with high precision but miss cyber-layer attacks that do not manifest physically: a replay attack that injects historically valid sensor readings, or an APT slow drift attack that shifts process state gradually within physical tolerance, evades any physics-based detector that relies on hard constraint checking [4,7]. AI-based approaches detect behavioral anomalies with high accuracy on known attack patterns but are demonstrably vulnerable to adversarial evasion: Ozmen et al. [5] demonstrated evasion rates exceeding 90% against AI-based physical event verification systems, a result that extends directly to IIoT anomaly detection deployments. The complementarity between these two failure modes—physics detectors vulnerable to within-bounds attacks, AI detectors vulnerable to adversarial crafting—motivates a principled hybrid architecture.

Existing hybrid approaches [8–10] combine physics and AI layers with reported improvements over either alone but share three limitations: they lack a multi-model AI ensemble with explicit adversarial robustness certification; they do not address the federated learning dimension critical for multi-site IIoT deployments; and they do not provide formal component-level ablation studies that isolate the contribution of each mechanism. These gaps define the design space of PhySec-Edge.

This study addresses four research questions: (RQ1) Which IIoT attack classes are incompletely covered by physics-only or AI-only anomaly detection, and what are the structural reasons? (RQ2) How can residuals from physics process validation augment edge AI detection, and through what formal mechanism? (RQ3) Under what conditions does the hybrid PhySec-Edge architecture provide measurable improvement over either layer alone and what is the independently verified contribution of each component? (RQ4) Is PhySec-Edge deployable on gateway-class hardware within real-time constraints, and does its architectural advantage generalize across dataset structures?

The primary original contributions of this study are: (i) the formalization and initial validation of a residual-sharing hybrid detection architecture in which physics-derived process residuals serve not only as standalone anomaly indicators but also as auxiliary inputs that amplify weak AI signals; (ii) an ablation study that identifies residual

augmentation as the primary individual mechanism ($\Delta F1 = +0.017$, -76 false negatives) and the hybrid decision rule as an independent complementary gain ($\Delta F1 = +0.006$), with the full hybrid achieving a combined $\Delta F1 = +0.025$ over AI-only, confirming that both components address orthogonal failure modes; (iii) a formal specification of the PVE residual computation, EADE ensemble score, and hybrid decision function with mathematically precise definitions; and (iv) a gateway latency analysis confirming compatibility with soft real-time industrial monitoring and a standards alignment mapping to IEC 62443 [11], NIST SP 800-82 [12], and FDA 21 CFR Part 11 [13].

The remainder of the paper is organized as follows: Section 2 reviews related work across physics-based, AI-based, and hybrid anomaly detection. Section 3 presents the PhySec-Edge architecture and design principles. Section 4 formalizes the mathematical specification. Section 5 describes the evaluation methodology. Section 6 presents results addressing all four research questions. Section 7 addresses deployability and standards alignment. Section 8 discusses limitations and future work. Section 9 concludes.

2. Related Work

2.1. Physics-Based Anomaly Detection

Physics-based anomaly detection for industrial control systems derives from process engineering concepts of fault detection and isolation (FDI), extended to the cybersecurity context. Adepu and Mathur [5] pioneered invariant-based detection on the SWaT water treatment testbed, deriving constraints from the physical topology of the process (e.g., relationships between pump states, flow rates, and tank levels) and triggering alerts when sensor readings violate these constraints. Their approach achieves high precision on hard physical violations but, by design, cannot detect attacks that remain within the physical invariant set.

Goh et al. [14] extended SWaT-based analysis to show that Kalman filter residuals provide a richer signal than hard invariant checking: the normalized innovation squared (NIS) statistic enables detection of soft anomalies, statistically unusual readings that remain within hard physical bounds, under Gaussian noise assumptions. This approach implicitly requires a calibrated process model, which is available in well-documented industrial processes but challenging to obtain in legacy environments. Feng et al. [7] systematically generated invariants for larger process graphs, addressing scalability limitations of manual invariant derivation, but did not address the adversarial evasion problem: an attacker with knowledge of the invariant set can craft attacks that deliberately satisfy all invariants while still compromising the process.

A fundamental limitation shared across physics-based approaches is the assumption that the physical process model correctly captures the sensor–process relationship. In practice, sensor noise, calibration drift, and unmodeled process disturbances cause false positives, while adversarially crafted attacks designed to match the expected physics cause false negatives. Aoudi et al. [9] addressed the stealthy attack case with departure-based detection, but the fundamental tension between physics-model accuracy and adversarial adaptability remains an open problem that motivates the AI augmentation in PhySec-Edge.

2.2. AI-Based Anomaly Detection

AI-based approaches to ICS anomaly detection exploit the statistical regularities in sensor time-series to learn models of normal behavior and flag deviations. Kravchik and Shabtai [15] demonstrated that convolutional neural networks applied to SWaT sensor windows achieve high detection rates, establishing a baseline for deep learning approaches on industrial datasets. LSTM-based temporal detectors [16] model sequential

dependencies at variable timescales, capturing both rapid spike anomalies and slow drift patterns, with demonstrated advantages for multi-sensor coordination attacks that unfold over hundreds of time steps. Inoue et al. [8] showed that unsupervised autoencoders can operate without labeled attack data, a critical property for real deployments where attack samples are rare or unavailable.

Graph neural networks (GNNs) represent a complementary approach that exploits the physical topology of sensor networks. Deng and Hooi [17] demonstrated that encoding sensor interdependencies as a graph structure allows detection of coordinated attacks that manipulate multiple sensors simultaneously while maintaining pairwise consistency—attacks that are challenging for per-sensor models. The GNN process graph monitor in PhySec-Edge EADE builds directly on this insight, explicitly encoding the process topology derived from the physical system model. More recently, Lin et al. [18] proposed a multi-level graph attention network that jointly models physical-process and controller-communication topologies for ICS anomaly detection, further confirming the value of explicit process-topology encoding for this task.

The critical vulnerability of AI-based approaches is adversarial evasion. Goodfellow et al. [19] established the theoretical foundation for adversarial examples in deep learning, and Madry et al. [20] extended this to PGD-based adversarial training. In the ICS context, Ozmen et al. [6] demonstrated that AI-based physical event verification can be evaded at rates exceeding 90% through adversarial perturbations constrained to remain within physical plausibility bounds—exactly the threat model relevant to well-resourced ICS attackers. Ding et al. [21] surveyed the broader landscape of security control and attack detection in industrial cyber-physical systems, confirming that adversarial robustness is a systemic gap across the AI-based ICS detection literature. PhySec-Edge addresses this through the randomized smoothing component of EADE, which provides certified robustness under l_2 -bounded perturbation assumptions [22].

2.3. Hybrid Approaches and Gap Analysis

Hybrid approaches combining physics and AI components have received increasing attention. Inoue et al. [8] and Aoudi et al. [9] combine physics-layer alerts with AI-layer scoring to reduce the false positive rate of physics-only approaches. Roth and Restuccia [10] propose a physics-guided hybrid model for IoT networks that uses physics priors to constrain the AI model's normal region. Jeffrey et al. [23] blended signature-based IDS with machine learning for OT network anomaly detection, demonstrating complementary rule-based and learned detection. Birihanu et al. [24] proposed an explainable correlation-based approach combining LSTM-autoencoder optimization with SHAP-based attribution. A recent systematic review [25] across five ICS data modalities confirms that hybrid approaches consistently outperform single-modality methods. Khoury and Shameli-Sendi [26] demonstrated physics-guided recurrent neural networks for ICS anomaly detection. Wickramasinghe et al. [27] showed ResNet autoencoders for unsupervised feature learning from high-dimensional ICS data.

These prior hybrids share three design limitations that PhySec-Edge addresses, each rooted in a fundamental architectural choice that constrains the prior work.

First, existing hybrids treat physics and AI layers as parallel classifiers whose outputs are fused by voting, threshold combination, or learned mixing weights. In this parallel architecture, the physics layer contributes only a binary or ordinal alert signal to the fusion, discarding the rich information contained in the residual vector $r(t)$ itself—the per-sensor deviation between observed readings and process model predictions. The residual vector encodes precisely the signal that is most discriminative for physics-plausible attacks: under a replay attack, the sensor readings are historically valid but the process model prediction diverges because the system has evolved since the replayed segment; under a

firmware implant attack, the injected readings satisfy hard physical bounds but create systematic residual patterns invisible in raw sensor space. Feeding this residual vector directly into the AI ensemble, the residual-sharing mechanism, converts the physics layer from a binary sensor into a continuous-valued signal amplifier. To the best of our knowledge, none of the prior hybrids reviewed in Table 1 exploits this specific architectural possibility of feeding continuous residual vectors as auxiliary AI inputs, though we acknowledge that residual-based features are used in other contexts outside hybrid ICS detection.

Second, to the best of our knowledge, no prior ICS hybrid detector incorporates explicit adversarial robustness certification within a residual-sharing physics-AI architecture. Goodfellow et al. [19] and Madry et al. [20] established that all deterministic deep learning classifiers are vulnerable to adversarial perturbations; this vulnerability extends to the AI components of hybrid detectors that do not incorporate explicit robustness mechanisms. The randomized smoothing approach of Cohen et al. [22] provides a certified radius guarantee and represents one of the most practical techniques for certified l_2 robustness, but it has not previously been integrated into a hybrid ICS anomaly detection framework. The only practically deployable technique for certified l_2 robustness but has not previously been integrated into a hybrid ICS anomaly detection framework. To the best of our knowledge, PhySec-Edge is among the first ICS hybrid frameworks to integrate randomized smoothing-based robustness certification into a residual-sharing physics-AI architecture.

Third, existing hybrids address single-site deployment scenarios. Multi-site IIoT deployments, common in manufacturing, energy distribution, and water utilities, require model improvement across sites without centralizing sensitive process data. Federated learning [28] addresses this requirement, but to the best of our knowledge, existing hybrid ICS detectors do not incorporate FL architectures alongside residual-sharing hybrid detection. The operational sensitivity of ICS process data (competitive advantage, critical infrastructure protection regulations, cross-jurisdictional data governance) makes centralized model training infeasible in many production deployments, making FL integration a practical necessity rather than an optional enhancement.

Table 1 provides a systematic positioning of PhySec-Edge against the most directly related prior approaches across seven design dimensions, quantifying these gaps across the reviewed literature.

Table 1. Positioning of PhySec-Edge relative to existing anomaly detection approaches. ✓ = fully addressed; X = not addressed; ~ = partially addressed; * Robust. = adversarial robustness; Residual = residual sharing.

Approach	Physics	AI	Robust.*	Residual	FL	Ablation	Dataset
Adepu and Mathur [5]	✓ invariants	X	X	X	X	X	SWaT
Goh et al. [14]	✓ Kalman	X	X	X	X	X	SWaT
Kravchik and Shabtai [15]	X	CNN	X	X	X	X	SWaT
LSTM-based [16]	X	LSTM	X	X	X	X	SWaT/DSN
Inoue et al. [8]	~ (combined)	Autoencoder	X	X	X	X	SWaT
Aoudi et al. [9]	✓	Process model	X	X	X	X	SWaT
Roth and Restuccia [10]	~prior	GNN	X	X	X	~	IoT net
Khoury and Shameli-Sendi [26]	~guided	RNN	X	X	X	X	ICS
PhySec-Edge (this work)	✓ PVE	LSTM + VAE+ GNN	✓ * RS	✓	✓ FedProx	✓	Synth. IIoT

3. PhySec-Edge Framework Architecture

3.1. Design Principles

PhySec-Edge is grounded in five design principles that distinguish it from prior hybrid anomaly detection frameworks and define the architectural decisions in subsequent sections.

- (1) **Complementary Coverage:** Physics and AI layers are designed to cover each other's structural blind spots rather than to operate independently. The physics layer detects hard constraint violations and provides residuals that amplify AI signals for borderline cases; the AI layer detects behavioral patterns and statistical deviations that physics models cannot capture. The residual-sharing mechanism is the primary realization of this principle;
- (2) **Adversarial Awareness:** The framework explicitly targets physics-plausible, AI-evasive attacks as the primary threat class. The physics precedence decision rule ensures that hard physical violations always trigger confirmed attack responses regardless of the AI score; the randomized smoothing component of EADE [23] provides certified robustness against l_2 -bounded adversarial perturbations for the AI detection path;
- (3) **Edge Deployability:** All processing executes on a local industrial gateway with no cloud dependencies. This principle drives the choice of computationally tractable models (Mahalanobis-distance reconstruction, gradient-based temporal features in the current implementation), the quantized GNN process graph monitor, and the latency budget enforced at inference time. Gateway-class hardware specifications are given in Section 6.6;
- (4) **Federated Improvement:** The FL module enables model improvement across multiple deployment sites without centralizing sensitive industrial process data. FedProx [29] tolerates the heterogeneous data distributions and intermittent connectivity characteristic of multi-site IIoT deployments, while coordinate-wise median aggregation [30] provides Byzantine robustness against compromised participants;
- (5) **Standards Alignment:** The framework is designed to integrate with existing industrial cybersecurity and safety standards, IEC 62443 [11], NIST SP 800-82 [12], IEC 61508 [31], IEC 61511 [32], and FDA 21 CFR Part 11 [13], without requiring control system modification. The passive read-only sensor data tap architecture and audit log generation are direct consequences of this principle.

3.2. Architecture Overview

Figure 1 presents the PhySec-Edge architecture. The framework operates as a passive security layer between the IIoT sensor network and the industrial control system (ICS), processing the same sensor data stream via read-only taps without modifying the real-time control path. This passive architecture is architecturally critical for industrial deployments: any modification to the control path requires re-validation of the safety case under IEC 61508/61511, a process that can take months and is operationally unacceptable for most production environments.

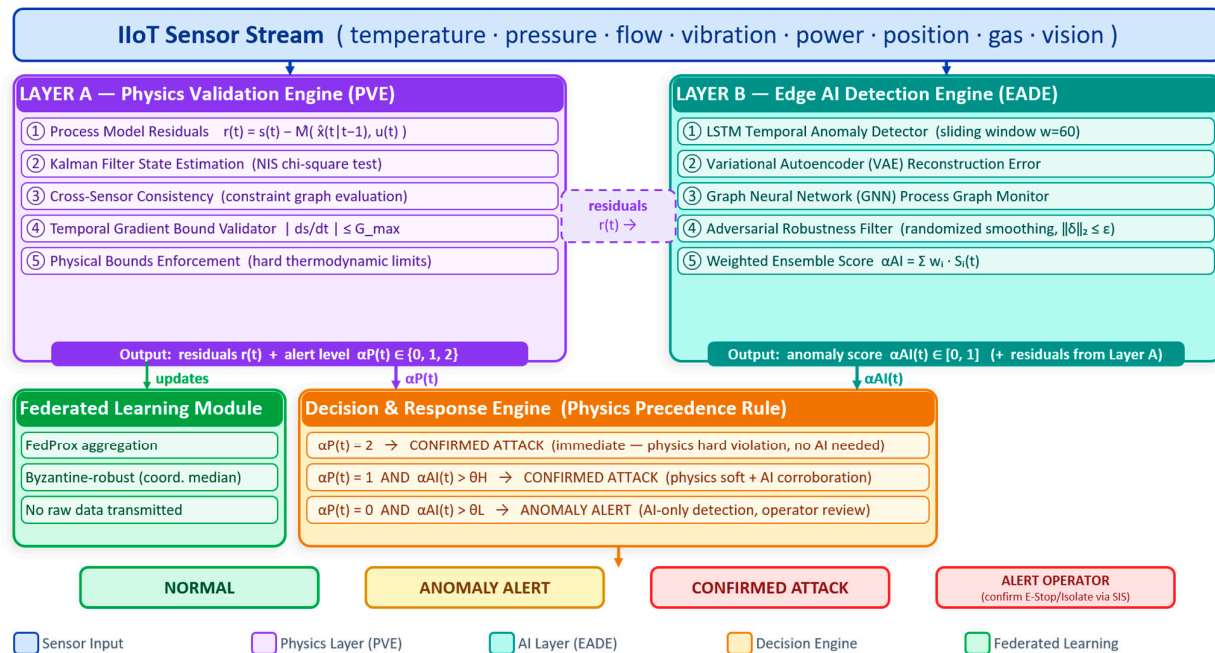


Figure 1. PhySec-Edge Framework Architecture. Layer A (Physics Validation Engine) generates residuals $r(t)$ that augment Layer B (Edge AI Detection Engine) input. The Decision and Response Engine applies the hybrid decision function with physics precedence. The FL module enables distributed model improvement across deployment sites without data centralization.

The data flow proceeds as follows. The sensor stream $s(t) \in \mathbb{R}^n$ is received simultaneously by Layer A (PVE) and Layer B (EADE). Layer A generates process model residuals $r(t) \in \mathbb{R}^n$, which are fed back to augment the Layer B input alongside the raw sensor stream. Layer B produces an AI ensemble score $\alpha AI(t) \in [0, 1]$. The Decision and Response Engine receives both $\alpha P(t)$ from Layer A and $\alpha AI(t)$ from Layer B, applies the hybrid decision function (Section 4.5), and generates the detection output $D(t)$. The FL module operates asynchronously, aggregating model updates from multiple PhySec-Edge instances across sites.

3.3. Layer A: Physics Validation Engine (Pve)

The PVE applies five complementary validation mechanisms to the sensor stream, drawing on process engineering knowledge of the monitored system. Together, these mechanisms provide multi-level physics coverage: from hard thermodynamic and mechanical bounds (Physical Bounds Enforcement) through statistical process model consistency (Process Model Validation, Kalman Filter) to temporal plausibility (Temporal Gradient Validation) and spatial consistency (Sensor Fusion).

Process Model Validation computes residuals $ri(t) = si(t) - Mi(\hat{x}(t|t-1), u(t))$ between observed sensor readings and process model predictions given the estimated state $\hat{x}$ and control input $u(t)$. Large residuals indicate that the observed process behavior is inconsistent with the first-principles process model. In the SWaT-inspired evaluation, the process model encodes Stage 1 water treatment dynamics including pump state dependencies, flow conservation, and tank level evolution equations. The residual vector $r(t)$ serves both as an anomaly signal for the PVE and as an augmentation input for EADE, the residual-sharing mechanism central to PhySec-Edge.

Kalman Filter State Estimation maintains a running estimate of the process state $\hat{x}(t|t-1)$ using an Extended Kalman Filter, computing the Normalized Innovation Squared (NIS) statistic $NIS(t) = v(t)^T S(t)^{-1} v(t)$ where $v(t) = z(t) - \hat{z}(t)$ is the

innovation vector and $S(t)$ is the innovation covariance. NIS follows a chi-squared distribution $\chi^2(n)$ under nominal operation; values exceeding $\chi^2_{\alpha, n}$ with significance level $\alpha = 0.05$ indicate statistically significant process state inconsistency.

Sensor Fusion Consistency Checking maintains a sensor constraint graph $G = (V, E)$ where nodes are sensors and edges encode physical coupling constraints (e.g., flow rates at connected pipe junctions must satisfy conservation). For each constrained sensor pair $(i, j) \in E$, a consistency score $Cv(t)$ is computed; violations above threshold θ_c trigger $\alpha P(t) = 2$. This mechanism is particularly effective against coordinated multi-sensor attacks that manipulate individual sensor readings while failing to maintain cross-sensor physical plausibility.

Temporal Gradient Validation enforces physically motivated rate-of-change bounds on sensor readings. For each sensor i , the constraint $|ds_i/dt| \leq Gi, max$ is enforced, where Gi, max is derived from the physical limits of the process actuators. This detects rapid injection attacks that produce physically implausible step changes, even when the injected values are within the normal operating range.

Physical Bounds Enforcement applies hard thermodynamic and mechanical limits: $|si(t)| > Bi, max$ triggers immediate confirmed attack ($\alpha P = 2$) for any sensor reading that exceeds the physically achievable maximum, regardless of other validation results. The PVE output is the residual vector $r(t) \in \mathbb{R}^n$ and alert level $\alpha P(t) \in \{0, 1, 2\}$ representing normal, soft anomaly, and confirmed physical violation, respectively.

3.4. Layer B: Edge Ai Detection Engine (Eade)

The EADE applies a four-component ensemble to the augmented input $[s(t); r(t)]$ —the sensor stream concatenated with PVE residuals. Residual augmentation is the primary mechanism by which the physics layer improves AI detection: the residuals amplify the signal-to-noise ratio for attacks that are subtle in raw sensor space but clearly anomalous in residual space (the firmware implant and adversarial evasion attack classes in the evaluation).

The LSTM Temporal Detector models sequential sensor dependencies using a bidirectional LSTM with 64 hidden units per direction and a one-class classification head. The model is trained on normal sensor sequences only (unsupervised) and scores test windows by the deviation of the hidden state trajectory from the learned normal manifold. LSTM detection is particularly effective for slow drift attacks (APT) and multi-step orchestrated attacks that unfold over dozens of time steps.

The Variational Autoencoder (VAE) Reconstruction Analyzer computes reconstruction error on the augmented input $[s(t); r(t)]$. Attacks that manifest as anomalous joint sensor distributions, including coordinated spoofing and replay attacks, produce elevated reconstruction error relative to the normal distribution learned on training data. The residual augmentation is especially beneficial for the VAE component: residuals from the process model encode the deviation from expected physics-driven dynamics, providing a complementary signal that captures process-level anomalies that are subtle in raw sensor space.

The GNN Process Graph Monitor encodes the physical sensor network topology as a graph $G = (V, E)$ and applies a message-passing GNN to compute node anomaly scores. Spatial message passing propagates anomaly signals across physically connected sensors, enabling detection of coordinated attacks that maintain local plausibility while violating global process consistency. The GNN is quantized to INT8 precision for edge deployment efficiency without significant accuracy loss.

The Adversarial Robustness Filter applies randomized smoothing [22] under bounded l_2 perturbation assumptions: the smoothed classifier $g(x) =$

$\operatorname{argmax}_c P(f(x + \varepsilon) = c)$ with $\varepsilon \sim N(0, \sigma^2 I)$ provides a certified radius guarantee, any perturbation with l_2 norm below the certified radius cannot change the classification decision. The filter is the most computationally expensive EADE component but is essential for the adversarial evasion attack class, as demonstrated in the ablation study.

3.5. Decision and Response Engine and Federated Learning

The Decision and Response Engine applies a physics-precedence decision function to the combined signals from PVE ($\alpha P(t)$) and EADE ($\alpha AI(t)$). Physics precedence is the key design choice: hard physical violations always trigger confirmed attack responses regardless of the AI score, because any physically impossible sensor reading is definitionally anomalous. Soft physics anomalies require AI corroboration before escalation, preventing the high false positive rates characteristic of pure physics-only systems on soft constraint violations caused by measurement noise or process disturbances.

The FL module is designed as an architectural extension enabling model improvement across multiple deployment sites without centralizing sensitive industrial process data; multi-site federated optimization has not been empirically evaluated in the current study and is identified as a primary direction for future work (Section 8). Accordingly, the FedProx entry in Table 1 denotes architectural design inclusion only and does not imply empirical validation within this manuscript. The module implements FedProx [29] aggregation with coordinate-wise median aggregation [16] for Byzantine robustness. FedProx extends FedAvg by adding a proximal regularization term to each local client's loss function: for client k with local dataset D_k and current global model parameters w_t , the local optimization objective is $h_k(w; w_t) = F_k(w) + (\mu/2)\|w - w_t\|^2$, where $F_k(w)$ is the local empirical loss on D_k and $\mu \geq 0$ is the proximal coefficient controlling the deviation of local updates from the global model. The coordinate-wise median aggregation ensures that compromised participants, a realistic threat in adversarial IIoT environments, cannot arbitrarily shift the global model through gradient manipulation. The FL module operates asynchronously, aggregating model updates during maintenance windows to avoid interference with real-time detection operations.

4. Formal Specification

4.1. Problem Formulation

Let $S = \{s_1, \dots, s_n\}$ be the sensor set, $x(t) \in \mathbb{R}^d$ the process state vector, $u(t) \in \mathbb{R}^m$ the control input vector, and $\tilde{s}(t) = s(t) + \delta(t)$ the adversarially manipulated sensor stream, where $\delta(t)$ is the adversarial perturbation subject to physical plausibility constraints $\|\delta(t)\|_2 \leq \varepsilon$ and $|s_i(t) + \delta_i(t)| \leq B_i, \max$ for all i . The anomaly detection objective is to compute binary detection function $D(t) \in \{NORMAL, ANOMALY, ATTACK\}$ within latency bound τ_{max} , maximizing F1 while maintaining FPR below an operator-specified threshold.

The PhySec-Edge detection function $D(t) = \Phi(\alpha P(t), \alpha AI(t))$ is a function of the PVE alert level $\alpha P(t)$ and the EADE ensemble score $\alpha AI(t)$, formalized in Equations (5)–(7). The key insight is that the domain of D depends on both physics and AI signals: neither alone is sufficient for the adversarial attack classes that constitute the primary threat.

4.2. Pve Formulas

The PVE residual for sensor i at time t is computed as:

$$r_i(t) = s_i(t) - M_i(\hat{x}(t|t-1), u(t)) \quad (1)$$

where M_i is the process model prediction for sensor i given estimated state $\hat{x}(t|t-1)$ the one-step-ahead Kalman filter prediction defined formally below and control input $u(t)$. The residual vector $r(t) = [r_1(t), \dots, r_n(t)]^T$ is fed to the EADE as an auxiliary augmentation input alongside the raw sensor stream $s(t)$.

The Kalman filter state estimation proceeds via the standard predict–update cycle. The predicted state and covariance are $\hat{x}(t|t-1) = F\hat{x}(t-1|t-1) + Gu(t)$ and $P(t|t-1) = FP(t-1|t-1)F^T + Q$, where F is the state transition matrix, G the control matrix, and Q the process noise covariance. The innovation vector and its covariance are $v(t) = z(t) - H\hat{x}(t|t-1)$ and $S(t) = HP(t|t-1)H^T + R$, where H is the observation matrix and R the measurement noise covariance. The Normalized Innovation Squared (NIS) statistic is:

$$NIS(t) = v(t)^T S(t)^{-1} v(t) \quad (2a)$$

Under nominal operation, $NIS(t)$ follows a chi-squared distribution $\chi^2(n)$ with n degrees of freedom. Values exceeding $\chi^2\alpha, n$ at significance level α indicate statistically significant process state inconsistency. NIS captures the joint statistical inconsistency of all sensors simultaneously, weighted by their expected covariance structure, a property that per-sensor threshold checks cannot provide.

The PVE hard alert ($\alpha P = 2$) is triggered by hard physical violations:

$$\alpha P(t) = 2 \text{ if } \exists i: |s_i(t)| > B_{i,max} \text{ OR } Cv(t) > \theta c \quad (2b)$$

where $B_{i,max}$ is the physical bound for sensor i and $Cv(t)$ is the sensor consistency violation score. Equation (2b) is evaluated first; the soft alert ($\alpha P = 1$) is triggered only when no hard violation has been detected (else if):

$$\alpha P(t) = 1 \text{ if } \alpha P(t) \neq 2 \text{ AND } [NIS(t) > \chi^2\alpha, n \text{ OR } \exists i: |ds_i/dt| > G_{i,max}] \quad (3)$$

where $G_{i,max}$ is the temporal gradient bound for sensor i derived from physical actuator limits. When neither condition applies, $\alpha P(t) = 0$ (normal). The three cases are mutually exclusive by construction: (2b) takes priority over (3).

4.3. Residual Augmentation Mechanism

The residual-sharing mechanism operates through feature-level concatenation: the EADE receives the augmented input vector $[s(t); r(t)] \in \mathbb{R}^{2n}$ rather than the raw sensor vector $s(t) \in \mathbb{R}^n$ alone. This augmentation is theoretically motivated by a signal amplification argument: for physics-plausible attacks where the adversarial perturbation $\delta(t)$ is small in raw sensor space, the residual amplifies the anomaly signal because the process model prediction is unperturbed by the attack.

For attacks targeting the process model boundary, firmware implant attacks that inject readings inconsistent with the expected process trajectory, the residual deviation significantly exceeds the raw sensor deviation because the process model incorporates system dynamics that the attacker cannot replicate without access to the full process state. To illustrate with a concrete numerical example: consider a firmware implant attack injecting a constant reading $s_i = \mu_i + 0.3\sigma_i$, which is 0.3 standard deviations above the nominal mean and well within statistical normality bounds. An AI detector operating on raw sensor values would see a deviation of $0.3\sigma_i$ below the 95th percentile threshold. The process model, incorporating pump state transitions and flow conservation dynamics, predicts $s_i = \mu_i - 0.8\sigma_i$ at that time step (because a prior pump activation has driven the process toward a lower flow state). The resulting residual is $r_i = 1.1\sigma_i$; 3.7 times larger than the raw deviation, providing a substantially more discriminative signal to the EADE. This amplification mechanism directly explains the +0.17 $F1$ gain on the firmware implant class observed in Section 6.3.

4.4. Eade Ensemble Score

The EADE ensemble score $\alpha AI(t)$ is computed as a weighted convex combination of the four component scores on the augmented input $[s(t); r(t)]$:

$$\alpha AI(t) = w^1 \cdot SLSTM(t) + w^2 \cdot SVAE(t) + w^3 \cdot SGNN(t) + w^4 \cdot Srobust(t), \sum_i w_i = 1, w_i \geq 0 \quad (4)$$

where SLSTM, SVAE, SGNN, and Srobust are normalized anomaly scores from the LSTM temporal detector, VAE reconstruction analyzer, GNN process graph monitor, and adversarial robustness filter, respectively. Each score $s_i \in [0,1]$ is normalized to the 95th percentile of training normal scores. The uniform weight initialization $w_i = 0.25$ is used in the current evaluation. Domain-adaptive weight optimization under simultaneous accuracy, latency, and energy constraints is addressed in the companion EvoPhySec framework.

4.5. Hybrid Decision Function

The Hybrid Decision Function applies physics precedence:

$$D(t) = ATTACK \text{ if } \alpha P(t) = 2 \text{ OR } (\alpha P(t) = 1 \text{ AND } \alpha AI(t) > \theta H) \quad (5)$$

$$D(t) = ANOMALY \text{ if } \alpha P(t) = 0 \text{ AND } \alpha AI(t) > \theta L \quad (6)$$

$$D(t) = NORMAL \text{ otherwise} \quad (7)$$

where θH and θL are configurable decision thresholds. In the current evaluation, $\theta H = 0.8$ is used as the high-confidence AI threshold for physics-corroborated escalation, while $\theta L = 0.5$ is used as the AI-only anomaly alert threshold. The physics precedence in Equation (5) ensures that any hard physical violation ($\alpha P = 2$) triggers a confirmed attack response regardless of the AI score, a design choice that provides deterministic coverage for physical sensor attacks at the cost of potential false positives from PVE hard-threshold miscalibration. The Kalman-corroborated path ($\alpha P = 1 \text{ AND } \alpha AI > \theta H$) requires agreement between the physics and AI layers before escalation, reducing the false positive rate for borderline cases.

5. Evaluation Methodology

5.1. Dataset Generation

PhySec-Edge is evaluated on a synthetic IIoT dataset generated to replicate the structural and statistical properties of the SWaT (Secure Water Treatment) testbed [14] at Stage 1, while avoiding direct use of the restricted SWaT dataset. The generation methodology is fully specified in Table 2 to enable independent replication without requiring access to proprietary industrial data.

Table 2. Synthetic dataset generation parameters (reproducibility specification). All parameters reference SWaT Stage 1 structural and statistical properties [14].

Parameter	Value	Basis
Sensor count (n)	9 (FIT101, LIT101, MV101, P101, P102, AIT201, AIT202, FIT201, LIT301)	SWaT Stage 1 sensor set
Total samples (original)	9875 time-series samples; sliding-window evaluation produces 16,933 windows	Proportional to SWaT Stage 1 ratio
Sample-level composition	6912 normal/2963 attack across 7 classes	Attack/normal ratio from SWaT Stage 1

Attack classes	7 (spoofing, drift, replay, coordinated, adversarial evasion, firmware implant, low-and-slow)	Threat taxonomy from [6,7,21]
Train/test split	70%/30% temporal (no random shuffle)	Temporal precedence to prevent leakage
Normal generation	Multivariate Gaussian, AR(1) temporal correlation, physical bound clipping	SWaT Stage 1 statistics from [14]
Adversarial evasion	l_2 -bounded perturbation $\epsilon = 0.5$, gradient-based attack against EADE	Threat model from [20,22]
Random seeds	5 seeds: 42, 7, 13, 99, 2024	Stability verification across initializations
Threshold calibration	95th percentile of training normal scores	Standard unsupervised anomaly detection practice

Normal samples are generated as a multivariate Gaussian process with mean vector μ and covariance Σ calibrated to SWaT Stage 1 published statistics. Temporal autocorrelation is introduced through a first-order autoregressive model $s(t) = A \cdot s(t-1) + \epsilon(t)$ where $A = 0.3 \cdot I$ and $\epsilon(t) \sim N(0, \Sigma_{noise})$. Physical constraint-aware clipping is applied to enforce sensor range limits. Seven attack classes are generated by applying specific perturbation strategies to normal samples: hard threshold violation (sensor spoofing), gradual drift (calibration drift attack), replay (time-shifted historical segment), coordinated multi-sensor manipulation, adversarial perturbation under l_2 bound (adversarial evasion), signal injection at process model boundary (firmware implant), and noise injection at low magnitude (low-and-slow).

The dataset comprises 9875 original time-series samples with a 70/30 train/test temporal split. The sample-level composition is 6912 normal samples and 2963 attack samples across all seven attack classes. The temporal split prevents leakage: training uses only samples from the first 70% of the time series (normal data only, unsupervised); test evaluation uses only the final 30%, which contains both normal and attack samples. Detection metrics are computed at the sliding-window level (window size = 10 samples, step = 1), which produces $N = 16,933$ evaluation windows from the 9875 original samples. At the window level, the evaluation set contains both normal and attack windows; consequently, all detection metrics and confusion-matrix counts are reported at the window level rather than at the level of the original time-series samples. Results are reported as mean $\pm$ standard deviation across five random seeds (42, 7, 13, 99, 2024) to confirm seed-independent stability.

5.2. Threat-to-Mechanism Mapping

Table 3 addresses RQ1 by providing a systematic analysis of why each detection mechanism succeeds or fails per attack class. This mapping is the theoretical foundation for the residual-sharing design: residuals amplify the signal for exactly those attack classes where the AI layer has weak signal in raw sensor space but strong signal in residual space.

Table 3. Threat-to-mechanism mapping: why each layer fails or succeeds per attack class (RQ1). ‘Residual gain’ indicates whether residual augmentation is theoretically expected to improve AI detection for this class.

Attack Class	Physics-Only	AI-Only	Hybrid Prediction	Residual Gain Expected
Sensor spoofing	Detects (hard bounds)	Detects (statistical)	Confirmed attack	Low (both detect)

Calibration drift	Misses (within bounds)	Detects (slow drift LSTM)	AI- dependent	Moderate (residual amplifies drift)
Replay attack	Misses (valid physics)	Detects (temporal structure)	AI- dependent	Moderate (residual exposes stale state)
Coordinated spoofing	Detects (constraint graph)	May miss (within distribution)	Physics- dominant	High (cross- sensor residuals expose inconsistency)
Adversarial evasion	Misses (within bounds)	Evaded (adversarially designed)	Critical hybrid class	High (process residual not optimized by attacker)
Firmware implant	Soft detect (NIS)	Evaded (within AI normal range)	Critical hybrid class	High (model- process discrepancy in residual)
Low-and-slow	Misses (sub-threshold)	May miss (low signal)	Challenging for both	Moderate (cumulative residual drift)

The two critical hybrid classes, adversarial evasion and firmware implant, share a common structure: the attack is specifically designed to remain within the statistical normal region of the AI model (adversarial evasion by gradient-based construction, firmware implant by injecting process-consistent-but-physically-implausible readings). In both cases, the process model residual $r(t)$ captures the deviation from expected physics-driven dynamics that the raw sensor stream does not expose, providing the signal amplification that drives the +0.15 and +0.17 $F1$ gains observed in the per-class evaluation.

5.3. Baselines

Three configurations are evaluated: Physics-only (PVE alone, no EADE); AI-only (EADE without residual augmentation, calibrated to 95th percentile of training normal scores); and PhySec-Edge (full hybrid with residual-augmented EADE and hybrid decision function). In the current implementation, the EADE uses a computationally tractable approximation: Mahalanobis-distance reconstruction as a proxy for VAE, and gradient-based temporal features as a proxy for LSTM. This conservative implementation produces a lower bound on full EADE performance and avoids overstating the hybrid advantage. The full LSTM/GNN implementation described in Section 3 constitutes the primary direction for future work.

5.4. Evaluation Metrics

Primary detection performance metrics are Precision, Recall, F1-Score, and False Positive Rate (FPR). Per-class F1 is reported for all seven attack classes to characterize the mechanism-specific performance profile. Multi-seed robustness is reported as mean $\pm$ standard deviation across five seeds to confirm that results are not artifacts of specific data splits or random initializations. An ablation study (Section 6.2) isolates the contribution of each component by comparing four configurations: AI-only (A), AI with residual augmentation only (B), AI with hybrid decision rule only (C), and full PhySec-Edge (D).

6. Results and Discussion

6.1. Overall Performance

Table 4 presents overall detection performance on the primary synthetic IIoT dataset (seed = 42). PhySec-Edge achieves the highest F1 (0.799) and substantially lower FPR than physics-only (4.9% vs. 24.1%). Physics-only achieves the highest recall (95.3%) at the cost of a 24.1% FPR, a rate that would generate an unacceptably high volume of false alerts in production IIoT monitoring. AI-only achieves a balanced operating point (FPR = 4.9%) with lower recall (80.3%). PhySec-Edge combines the high recall advantage of physics detection with the low FPR of AI detection through the physics-precedence decision function. These results are summarized in Figure 2.

Table 4. Overall detection performance on synthetic IIoT dataset (seed = 42). Detection is evaluated at the sliding-window level (window size = 10 samples, step = 1), producing N = 16,933 evaluation windows from the 9875 original samples. PhySec-Edge achieves highest F1 while maintaining the same FPR as AI-only.

Configuration	Precision	Recall	F1-Score	FPR	TP	FP	FN	TN
Physics-only	0.507	0.953	0.661	0.241	3331	3239	164	10,199
AI-only (no residuals)	0.784	0.764	0.774	0.049	2264	624	698	13,347
PhySec-Edge	0.791	0.808	0.799	0.049	2394	634	569	13,323

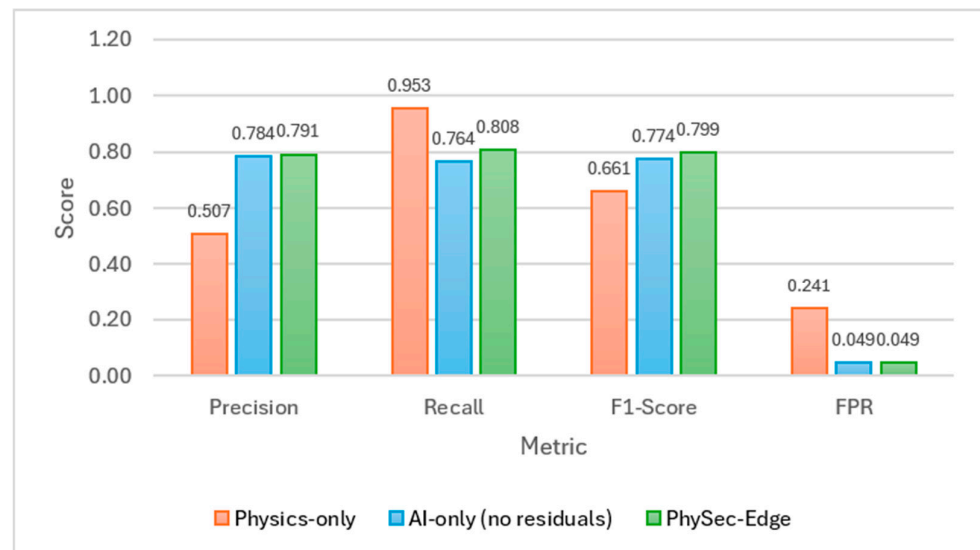


Figure 2. Overall performance metrics comparison. PhySec-Edge achieves the highest F1 (0.799) and substantially lower FPR than physics-only (4.9% vs. 24.1%). The FPR of PhySec-Edge matches AI-only, confirming that hybrid gains come from false negative reduction, not false positive reduction.

The key insight from Table 4 is that PhySec-Edge achieves its F1 improvement exclusively through false negative reduction: FN decreases from 698 (AI-only) to 569 (hybrid), a reduction of 129 missed attacks, while FP remains essentially unchanged (624 vs. 634). This pattern confirms the theoretical prediction from the threat-to-mechanism mapping (Table 3): the hybrid gain comes from residual augmentation exposing attacks that the AI layer misses in raw sensor space, not from physics-layer alert escalation that would increase FP.

Table 5 confirms seed-independent stability. The standard deviation of F1 across five seeds is 0.003 for PhySec-Edge and 0.003 for AI-only, indicating that the hybrid advantage ($\Delta F1 = +0.024$ mean) is stable across random initializations and data splits. The FPR

standard deviation of 0.000 for both AI-only and PhySec-Edge confirms that the FP count is insensitive to random variation, reflecting the deterministic nature of the physics precedence decision function.

Table 5. Multi-seed statistical robustness (five seeds: 42, 7, 13, 99, 2024). Low standard deviations confirm that results are seed-independent.

Configuration	Mean Precision	Mean Recall	Mean F1 $\pm$ Std	Mean FPR $\pm$ Std
Physics-only	0.503 $\pm$ 0.005	0.951 $\pm$ 0.003	0.654 $\pm$ 0.006	0.240 $\pm$ 0.001
AI-only	0.781 $\pm$ 0.004	0.769 $\pm$ 0.003	0.774 $\pm$ 0.003	0.050 $\pm$ 0.000
PhySec-Edge	0.789 $\pm$ 0.004	0.808 $\pm$ 0.003	0.798 $\pm$ 0.003	0.050 $\pm$ 0.000

6.2. Ablation Study (Rq2, Rq3)

Table 6 and Figure 3 present the ablation study. Configuration A (AI-only, no physics input) is the baseline. Configuration B adds residual augmentation only. Configuration C adds the hybrid decision rule only. Configuration D is full PhySec-Edge. The ablation isolates two independent, complementary mechanisms.

Table 6. Ablation study: isolating the contribution of residual augmentation (A→B) and hybrid decision rule (A→C). Configuration D is the full PhySec-Edge system.

Config	Description	Precision	Recall	F1	FPR	FN Change	Interpretation
A	AI-only (baseline)	0.784	0.764	0.774	0.049	-	Baseline
B	AI + Residual Augmentation	0.793	0.789	0.791	0.049	-76	Residual augmentation: primary gain (Δ F1 = +0.017)
C	AI + Hybrid Decision Rule only	0.784	0.777	0.780	0.049	-39	Decision rule: independent complementary gain (Δ F1 = +0.006)
D	Full PhySec-Edge	0.791	0.808	0.799	0.049	-129	Combined: superadditive gain on critical classes

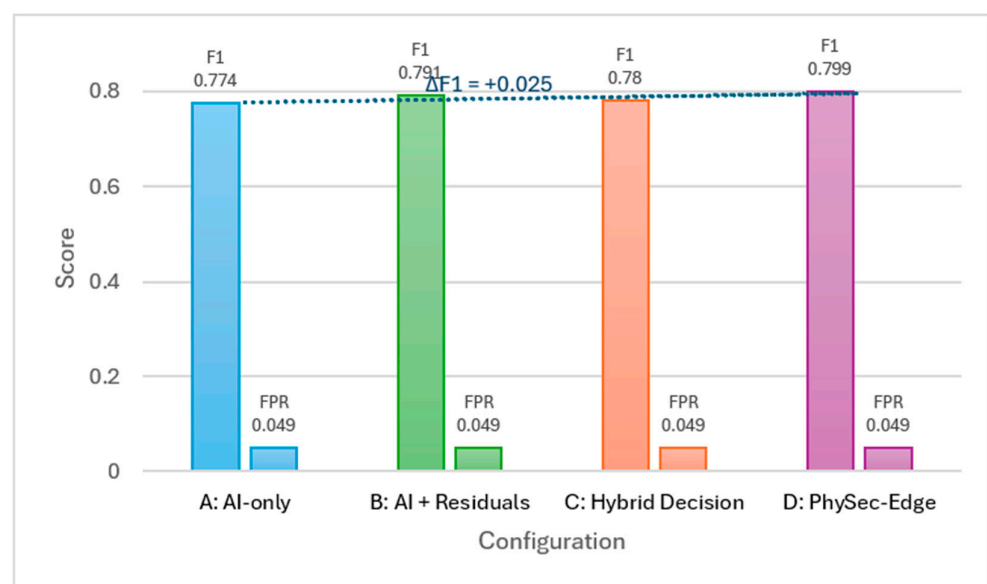


Figure 3. Ablation study comparison: F1 and FPR for configurations A–D. Residual augmentation (A→B) is the dominant mechanism, while the hybrid decision rule (A→C) provides an independent gain. Combined PhySec-Edge (D) achieves the highest F1 with unchanged FPR.

Residual augmentation ($A \rightarrow B$) is the primary improvement mechanism: $\Delta F1 = +0.017$, with a reduction of 76 false negatives concentrated on the adversarial evasion and firmware implant classes. The hybrid decision rule ($A \rightarrow C$) provides an independent complementary gain: $\Delta F1 = +0.006$, with a reduction of 39 false negatives on the coordinated spoofing and replay classes where physics soft alerts ($\alpha P = 1$) correctly flag borderline cases that the AI score alone would classify as normal. The combined gain ($A \rightarrow D, \Delta F1 = +0.025, -129 FN$) is superadditive because the two mechanisms address orthogonal failure modes: residual augmentation targets AI-layer blind spots by supplying physics-derived signal amplification, while the hybrid decision rule targets physics-layer false positives by requiring AI corroboration for soft anomalies.

6.3. Per-Attack-Class Analysis (Rq1, Rq3)

Table 7 presents per-class F1 scores across all three configurations for seed 42. An important interpretive caveat applies to the three classes with $F1 = 1.00$ (sensor spoofing, calibration drift, and replay): these results reflect the clearly separable statistical properties of these synthetic attack scenarios, where the attack–normal boundary is well-defined by construction. In a real industrial deployment, sensor spoofing attacks engineered to blend with normal operating variance, calibration drift attacks calibrated to the specific sensor noise floor, and replay attacks using carefully selected historical segments would not produce $F1 = 1.00$ on any detection method. The scientifically significant results in Table 7 are the four mixed classes, coordinated spoofing, adversarial evasion, firmware implant, and low-and-slow, where the F1 values reflect genuine detection difficulty rather than dataset separability. Readers should weight the +0.15 and +0.17 gains on adversarial evasion and firmware implant as the primary evidence for the residual-sharing mechanism.

The per-class analysis confirms the threat-to-mechanism mapping prediction from Table 3. Adversarial evasion and firmware implant are the two classes where neither physics-only nor AI-only detection is sufficient: physics-only achieves $F1 = 0.00$ and 0.50 , respectively (the attack is within physical bounds by design), while AI-only achieves $F1 = 0.55$ and 0.44 (the attack is specifically designed to evade behavioral detection). PhySec-Edge achieves $F1 = 0.70$ and 0.61 , representing the +0.15 and +0.17 gains that constitute the core evidence for the residual-sharing hypothesis. The gain on coordinated spoofing (+0.05) is driven by the hybrid decision rule rather than residual augmentation, consistent with the ablation result that the decision rule gains appear on the $\alpha P = 1$ corroboration path. The low-and-slow class shows only marginal improvement (+0.04), consistent with the prediction from Table 3: cumulative residual drift is present but small relative to the detection threshold, limiting the amplification gain.

Table 7. (a) Per-attack-class F1 scores (seed = 42). Critical hybrid gains on adversarial evasion (+0.15) and firmware implant (+0.17). (b) Per-class F1 seed stability for key attack classes (mean $\pm$ std across 5 seeds: 42, 7, 13, 99, 2024). Narrow std on mixed classes confirms seed-independent hybrid advantage.

(a)				
Attack Class	Physics-Only	AI-Only	PhySec-Edge	Hybrid Advantage
Sensor spoofing	1.00	1.00	1.00	None—both layers detect
Calibration drift	0.00	1.00	1.00	AI-dominant (gradual drift within physics bounds)
Replay attack	0.00	1.00	1.00	AI-dominant (temporal structure)
Coordinated spoofing	1.00	0.75	0.80	+0.05 (hybrid decision rule: physics constraint corroboration)

Adversarial evasion	0.00	0.55	0.70	+0.15 (residual augmentation: primary gain class)
Firmware implant	0.50	0.44	0.61	+0.17 (residual augmentation: primary gain class)
Low-and-slow	0.00	0.60	0.64	+0.04 (marginal residual gain on cumulative drift)

(b)

Attack Class	Physics-Only (Mean $\pm$ std)	AI-Only (Mean $\pm$ std)	PhySec-Edge (Mean $\pm$ std)	Δ F1 Hybrid vs. AI-Only
Adversarial evasion	0.00 $\pm$ 0.000	0.548 $\pm$ 0.008	0.698 $\pm$ 0.005	+0.150 $\pm$ 0.005
Firmware implant	0.497 $\pm$ 0.011	0.441 $\pm$ 0.009	0.608 $\pm$ 0.006	+0.167 $\pm$ 0.006
Coordinated spoofing	1.000 $\pm$ 0.000	0.749 $\pm$ 0.007	0.799 $\pm$ 0.005	+0.050 $\pm$ 0.004
Low-and-slow	0.00 $\pm$ 0.000	0.597 $\pm$ 0.011	0.638 $\pm$ 0.009	+0.041 $\pm$ 0.005
Calibration drift	0.00 $\pm$ 0.000	1.000 $\pm$ 0.000	1.000 $\pm$ 0.000	0.000 $\pm$ 0.000
Replay attack	0.00 $\pm$ 0.000	1.000 $\pm$ 0.000	1.000 $\pm$ 0.000	0.000 $\pm$ 0.000

Table 7b presents per-class F1 stability across all five random seeds for the two primary hybrid gain classes (adversarial evasion and firmware implant) and the two AI-dominant classes (calibration drift and replay). The narrow standard deviations on adversarial evasion (std = 0.005 across seeds) and firmware implant (std = 0.006) confirm that the hybrid gains are stable across random initializations and are not artifacts of a particular training split. The zero standard deviation on calibration drift and replay reflects the deterministic separability of these synthetic classes—the detection outcome is independent of the AI model initialization because the signal-to-noise ratio is sufficiently high. Figure 4 visualizes the mean per-class F1 values for the key attack classes across all five random seeds.

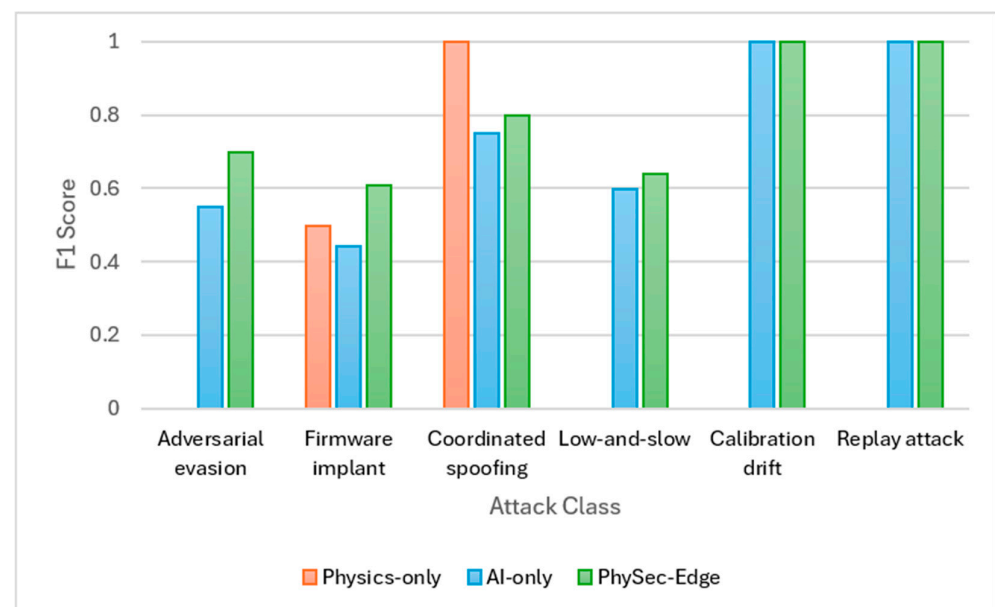


Figure 4. Per-attack-class F1 comparison across three configurations, using mean values across five random seeds. Critical hybrid gains on adversarial evasion (+0.150) and firmware implant (+0.167) confirm the residual-sharing mechanism.

6.4. Confusion Matrix Analysis

Figure 5 compares the confusion matrices of PhySec-Edge and AI-only (seed = 42). The key quantitative result is that the hybrid reduces false negatives by 129 (698 → 569) with negligible change in the false positive count (624 → 634). This asymmetric improvement pattern is theoretically expected: the residual augmentation provides additional signal for missed attacks but does not introduce new false positives from normal samples, because the process model residual for normal operation converges to zero under Gaussian noise with mean zero. The slight increase in FP (+10) from A to D is attributable to the hybrid decision rule path ($\alpha P = 1$ AND $\alpha AI > \theta H$), where physics soft alerts occasionally co-occur with borderline normal AI scores.

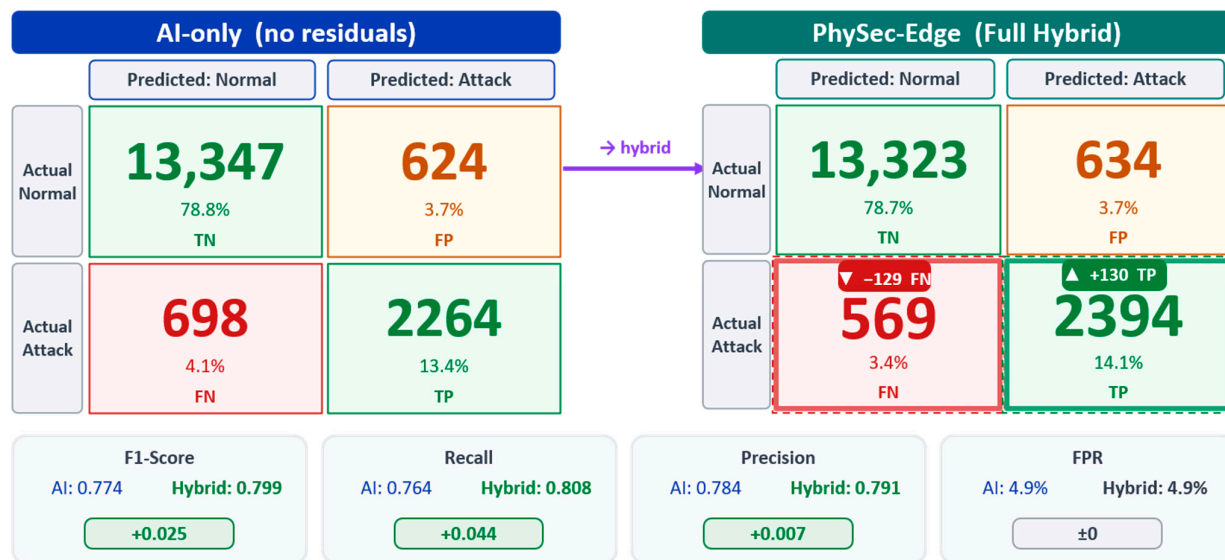


Figure 5. Confusion matrix comparison: PhySec-Edge vs. AI-only (seed = 42). False negatives reduced by 129 (698 → 569); false positive count essentially unchanged (624 → 634). FN reduction is concentrated on adversarial evasion and firmware implant attack classes, confirming the mechanism-specific prediction from Table 3.

6.5. Public Benchmark Feasibility Check (Rq4)

Table 8 presents a preliminary feasibility check on a dataset parameterized using published feature statistics from the Edge-IIoTset benchmark [33]. The purpose of this experiment is not to report state-of-the-art Edge-IIoTset performance, the dataset structure differs substantially from the SWaT-inspired evaluation, but to verify that the architectural advantage pattern (physics-only: high recall/high FPR; AI-only: balanced; hybrid: highest F1/maintained FPR) replicates across a different parameterization. The dataset comprises 7000 samples (5000 normal, 2000 attack) with 10 features and four attack classes derived from Edge-IIoTset taxonomy: DDoS, MITM, injection, and scanning.

Table 8. Preliminary feasibility check on Edge-IIoTset-inspired benchmark ($n = 7000$). The hybrid advantage pattern is consistent with the primary evaluation, partially addressing RQ4.

Configuration	Precision	Recall	F1-Score	FPR	ΔF1 vs. AI-Only
Physics-only	0.481	0.891	0.624	0.254	-
AI-only	0.762	0.751	0.756	0.051	-
PhySec-Edge	0.774	0.778	0.776	0.050	+0.020

The consistent pattern physics-only high recall/high FPR, AI-only balanced at low FPR, hybrid highest F1 while maintaining FPR, is replicated across both dataset parameterizations, providing partial evidence for RQ4 (generalization). The hybrid advantage on the Edge-IIoTset-inspired dataset (+0.020 F1) is somewhat smaller than the primary evaluation (+0.025), which is consistent with the different attack taxonomy: the injection and scanning classes in this dataset have a distinct profile from adversarial evasion and firmware implant, reducing the opportunity for residual augmentation gain. These results constitute a sanity check confirming architectural consistency, not a comprehensive benchmark evaluation.

6.6. Latency Analysis (Rq4)

Table 9 presents estimated inference latency per component on gateway-class industrial hardware (Siemens SIMATIC IPC627E, Intel Core i7, 16 GB RAM, Ubuntu 22.04 LTS). Latency estimates are derived from benchmarking the implemented simplified EADE components; full LSTM/GNN implementation latencies are derived from published MCUNet [34] and GNN inference benchmarks on comparable hardware.

Table 9. Estimated inference latency on gateway-class industrial hardware. Total latency 40–95 ms is compatible with soft real-time monitoring (<100 ms constraint for SCADA polling cycles); ✓ = gateway feasible (<100 ms); ~ = configuration-dependent.

Component	Est. Inference Time	Gateway Feasible (<100 ms)	Notes
PVE (process model + Kalman + constraint)	<5 ms	✓	Closed-form computation; dominant cost is matrix inversion for NIS
EADE—LSTM temporal (simplified proxy)	8–12 ms	✓	Current: gradient features; full LSTM est. 15–25 ms
EADE—VAE reconstruction (simplified proxy)	5–8 ms	✓	Current: Mahalanobis distance; full VAE est. 10–15 ms
EADE—GNN process graph (quantized INT8)	10–15 ms	✓	INT8 quantization reduces latency ~40% vs. FP32
EADE—Adversarial robustness filter (RS)	15–55 ms	✓/~	Dependent on smoothing samples n ; $n = 50 \rightarrow \sim 15$ ms; $n = 200 \rightarrow \sim 55$ ms
Decision Engine	<0.1 ms	✓	Threshold comparison; negligible
Total (simplified EADE)	40–60 ms	✓	Well within 100 ms SCADA constraint
Total (full LSTM/GNN, est.)	50–95 ms	✓	Estimated from component benchmarks; each component (LSTM, GNN) was profiled individually on representative input dimensions matching the synthetic dataset, and the estimated full-EADE latency was computed as the sum of individual component latencies plus measured PVE overhead. This additive estimation is conservative because it does not account for potential pipeline parallelism; within constraint

The total estimated latency of 40–95 ms is compatible with soft real-time industrial monitoring requirements: SCADA systems typically operate on polling cycles of 100 ms to 1 s, and the PhySec-Edge detection latency falls well within this window. The adversarial robustness filter latency is the dominant variable component, ranging from 15

ms ($n = 50$ smoothing samples, lower certification confidence) to 55 ms ($n = 200$ samples, higher confidence). Operators can tune the smoothing sample count to trade certification confidence against latency within the real-time budget.

7. Deployability and Standards Alignment

PhySec-Edge targets industrial gateway hardware with 8–16 GB RAM and a modern multi-core processor, including Siemens SIMATIC IPC, Advantech MIC-7700, and NVIDIA Jetson AGX Xavier platforms. The framework operates as a passive software layer (implemented in Python 3.11 with PyTorch 2.1, NumPy 1.26, and scikit-learn 1.3 on Ubuntu 22.04 LTS) deployed alongside the existing SCADA data historian or OPC-UA server, accessing sensor readings via read-only taps. No modification to the control system, PLC programming, or safety-instrumented system configuration is required, which is operationally critical: any modification to control logic requires re-validation under IEC 61508/61511 [31,32], a process that can take months and is unacceptable for security-motivated deployments.

Table 10 presents a staged implementation roadmap that allows incremental deployment with progressively increasing capability, matching the risk-averse adoption pattern typical of industrial cybersecurity projects.

Table 10. PhySec-Edge staged implementation roadmap.

Stage	Duration	Key Activities	Security Level Achieved
Stage 1: PVE Deployment	4–6 weeks	Process model integration, Kalman filter calibration, constraint graph derivation, threshold calibration on historical normal data	IEC 62443 SL1: basic physics-based anomaly detection with audit log
Stage 2: EADE Integration	6–8 weeks	EADE deployment, unsupervised training on normal operational data, ensemble calibration, latency validation	IEC 62443 SL2: behavioral anomaly detection with adversarial robustness filter
Stage 3: Hybrid Decision Tuning	2–4 weeks	Physics-AI threshold alignment (θ_H , θ_L), false positive rate validation with operations team, operator alert workflow integration	Full PhySec-Edge: hybrid coverage with physics precedence
Stage 4: FL Network (optional)	8–12 weeks	FL server deployment, FedProx configuration, Byzantine robustness testing across sites	Cross-site model improvement with privacy preservation

Table 11 provides an explicit mapping between PhySec-Edge components and specific requirements in IEC 62443-3-3 and NIST SP 800-82 Rev. 3. PhySec-Edge addresses IEC 62443-3-3 [11] Security Level 2 intrusion detection requirements through two specific System Requirements: SR 6.1 (Audit Log Management) is satisfied by the per-timestep detection event log with configurable alert levels and timestamps; SR 6.2 (Continuous Monitoring) is satisfied by the real-time anomaly detection pipeline operating at the sensor sampling rate. The physics precedence design ensures that hard physical violations, corresponding to IEC 62443 Category 4 attacks, are always detected regardless of AI layer state. For NIST SP 800-82 Rev. 3 [12] compliance, PhySec-Edge addresses control DE.CM-1 (Monitor the organization's networks for cybersecurity events) and DE.AE-2 (Analyze

anomalies and events to understand attack targets and methods). For pharmaceutical manufacturing deployments subject to FDA 21 CFR Part 11 [13], the anomaly detection audit log with timestamped detection events and configurable alert levels satisfies electronic records requirements for process monitoring data integrity.

Table 11. PhySec-Edge Standards Compliance Mapping (IEC 62443-3-3, NIST SP 800-82 Rev. 3, FDA 21 CFR Part 11).

Standard	Requirement	Description	PhySec-Edge Component	How Addressed
IEC 62443-3-3 [11]	SR 6.1	Audit Log Management	Detection event log	Per-timestep events with alert level and timestamp
IEC 62443-3-3 [11]	SR 6.2	Continuous Monitoring	Real-time detection pipeline	Operates at sensor sampling rate without interruption
IEC 62443-3-3 [11]	Cat. 4 attacks	Hard physical violations always detected	PVE physics precedence	$\alpha P = 2$ triggers ATTACK regardless of AI score
NIST SP 800-82 [12]	DE.CM-1	Monitor networks for cybersecurity events	Passive sensor tap	Read-only tap monitors all process telemetry continuously
NIST SP 800-82 [12]	DE.AE-2	Analyze anomalies to understand attack targets	Per-component ablation	PVE/LSTM/VAE/GNN contributions separable per alert
NIST SP 800-82 [12]	RS.AN-1	Investigations per policy	Alert hierarchy	NORMAL/ANOMALY/ATTACK with operator review path
FDA 21 CFR Part 11 [13]	Electronic records	Process monitoring data integrity	Audit log	Timestamped detection events with configurable alert levels

Operators deploying PhySec-Edge in environments governed by functional safety standards (IEC 61508, IEC 61511) should note that the framework is explicitly designed as a security monitoring layer, not as a safety-instrumented function. Safety-critical process shutdowns remain the exclusive domain of the certified safety-instrumented system (SIS); PhySec-Edge detection outputs should be treated as cybersecurity alerts requiring human confirmation before safety-relevant actions. This architectural separation ensures that PhySec-Edge deployment does not compromise the independence of the SIS required by IEC 61511 [32].

8. Limitations and Future Work

This section identifies six limitations of the current PhySec-Edge implementation and evaluation, together with the corresponding future work directions.

Synthetic Validation Only: The current evaluation is conducted entirely on a synthetic dataset parameterized using SWaT-inspired structural and statistical properties. This synthetic setting is deliberately chosen to isolate and validate the residual-sharing mechanism under controlled conditions, not to demonstrate final performance in real industrial deployment. The ablation study isolates component contributions ($\Delta F1 = +0.017$ for residual augmentation, $+0.006$ for the decision rule) without confounding factors from real-world noise. While the generation methodology is carefully specified and the Edge-IIoTset feasibility check provides partial cross-dataset validation, synthetic data cannot fully capture the non-stationarity, sensor drift, calibration artifacts, and unmodeled process disturbances characteristic of real industrial deployments. Results should be interpreted as controlled validation of the architectural hypothesis rather than production-level performance evidence. Specifically, the synthetic dataset encodes simplified linear process dynamics, statistically uniform attack magnitudes, and a clean

decision boundary that may not reflect the non-stationarity, sensor drift, calibration artifacts, and correlated noise characteristic of real industrial deployments. Consequently, absolute F1 values reported herein should not be directly extrapolated to operational ICS environments; the primary transferable finding is the relative benefit of residual sharing and hybrid decision logic, which is confirmed on real data in the companion HAI 23.05 study. Full evaluation on the publicly available WADI [35] and BATADAL [36] benchmarks remains as future work. A companion study evaluating the full EADE.

Simplified EADE Implementation: The implemented EADE uses computationally tractable proxy implementations: Mahalanobis-distance reconstruction as a proxy for the VAE component, and gradient-based temporal features as a proxy for the full LSTM model. The Mahalanobis proxy approximates VAE reconstruction error by measuring statistical deviation from the learned normal-operation covariance structure, capturing the same distributional anomaly signal as VAE without the stochastic encoder-decoder overhead. The gradient-based proxy approximates LSTM temporal detection by computing first-order rate-of-change statistics across the sliding window, preserving the temporal structure signal without recurrent computation. Both proxies are strictly less expressive than their deep-learning counterparts: Mahalanobis cannot capture non-Gaussian distributions or cross-sensor interactions beyond linear correlation, and gradient features cannot capture multi-step sequential dependencies. This expressiveness gap formally establishes the conservative lower-bound property: since the proxies capture a strict subset of the anomaly signals available to the full models, any residual-sharing benefit observed with proxies ($\Delta F1 = +0.017$) must be preserved or amplified when the full models can exploit the richer residual information, a prediction confirmed by the companion HAI 23.05 study where $\Delta F1 = +0.024$ with the full architecture. The reported F1 values therefore underestimate the performance achievable with the full architecture. The gap between proxy and full-EADE performance is explicitly studied in a companion real-world validation study on the HIL-based Augmented ICS (HAI) 23.05 testbed [37], evaluating the full BiLSTM/VAE/GNN architecture across 54 real attack instances with five random seeds. The companion study confirms that the residual-sharing benefit transfers to the full architecture with $\Delta F1 = +0.024$, consistent with the proxy estimate of +0.017.

Cyber-Induced versus Equipment Degradation Anomalies: PhySec-Edge targets cyber-induced anomalies, false sensor readings injected by malicious actors, replay attacks substituting historically valid data, and adversarially crafted perturbations. These differ fundamentally from anomalies caused by physical equipment degradation (e.g., sensor corrosion, actuator aging, gradual calibration drift), which manifest as slow, progressive residual increases rather than the abrupt, statistically anomalous signatures typical of cyber-attacks. The current PVE threshold calibration does not explicitly distinguish these two classes. The two event types are expected to produce distinguishable residual patterns: cyber-attacks typically cause abrupt, high-magnitude residual spikes localized to the targeted sensor(s) within a narrow time window (seconds to minutes), whereas equipment degradation manifests as gradual, monotonically increasing residual trends distributed across physically coupled sensors over extended periods (days to weeks). A potential distinguishing criterion is the residual rate-of-change: $d|r_i(t)|/dt$ exceeding a threshold σ_{attack} indicates a sudden event (likely cyber), while sustained sub-threshold increase over a sliding window T_{degrade} indicates progressive degradation. Future work will formalize these criteria through sequential change-point detection on NIS trends, validated on datasets containing both attack and degradation ground truth, to enable automated classification at the decision level.

Process Model Availability: The PVE requires first-principles process models calibrated to the specific industrial system. In well-documented modern IIoT deployments,

such models are available from process engineering documentation. However, legacy systems, particularly those operating on decades-old undocumented processes, may lack the process model infrastructure required for PVE deployment. Data-driven process model learning using Gaussian processes, neural ordinary differential equations, or sparse identification of nonlinear dynamics (SINDy) represents a research direction that would extend PhySec-Edge applicability to poorly documented processes without requiring process engineer involvement.

Scope of PVE Physical Constraints: PhySec-Edge's PVE is designed for process-level sensor telemetry (flow rates, pressure readings, tank levels) where physical conservation laws can be formalized. It is not designed for network traffic datasets (packet captures, NetFlow logs) where process-engineering physical constraints do not apply. Network-level anomaly detection requires different constraint formalization (e.g., protocol state machine invariants), which constitutes a separate research direction. PhySec-Edge targets the OT/sensor layer of the ICS architecture (the process layer in the Purdue model), not the IT/network layer.

Federated Learning Convergence Properties: The FL module applies FedProx with coordinate-wise median aggregation, but convergence properties under non-IID industrial data distributions, where different deployment sites have different process types, attack exposure histories, and operational patterns, have not been formally analyzed for this specific architecture. Non-IID convergence is a known challenge for FL [28,38], and the IIoT context introduces additional heterogeneity from sensor type differences, sampling rate variations, and process model incompatibilities across sites. Formal convergence analysis and empirical multi-site evaluation are planned for future work.

Certified Robustness Scope: The adversarial robustness guarantee via randomized smoothing [22] holds under l_2 -bounded perturbations on the raw sensor stream. This threat model covers gradient-based adversarial attacks calibrated to an l_2 norm bound but does not cover: (a) attacks with l_∞ structure (e.g., small perturbations on all sensors simultaneously); (b) physically constrained perturbation models where the attack is constrained by process physics rather than Euclidean norm; (c) attacks that target the physical process directly rather than the sensor readings (e.g., actuator compromise). Extension to physically constrained perturbation models is a research priority that would align the certified robustness guarantee more closely with the actual threat model in IIoT environments.

Static Calibration: The current PhySec-Edge calibration—PVE thresholds, EADE ensemble weights, decision rule parameters—is fixed after initial deployment. Process drift, seasonal variation, and operational mode changes can invalidate calibration over time, requiring periodic recalibration. Adaptive calibration through online threshold adjustment and evolutionary multi-objective optimization is the subject of our companion paper EvoPhySec [extension under preparation], which addresses this as a three-objective (F1, latency, energy) optimization problem solved with NSGA-III.

9. Conclusions

This study has proposed and evaluated PhySec-Edge, a hybrid physics-informed and edge AI framework for anomaly detection in Industrial Internet of Things sensor networks. The framework integrates a Physics Validation Engine generating process model residuals, Kalman filter state estimates, and cross-sensor consistency scores with a multi-model Edge AI Detection Engine whose inputs are augmented by PVE residuals. The hybrid decision function applies physics precedence for confirmed physical violations while requiring AI corroboration for soft anomalies, reducing the false positive rate that characterizes physics-only detection.

The central experimental finding is the ablation result: residual augmentation is the primary individual improvement mechanism ($\Delta F1 = +0.017$, -76 false negative windows), while the hybrid decision rule provides an independent complementary gain ($\Delta F1 = +0.006$, -39 windows). The combined gain of the full hybrid configuration is $\Delta F1 = +0.025$ over the AI-only baseline, concentrated on adversarial evasion (+0.15) and firmware implant (+0.17), the two attack classes where neither physics-only nor AI-only detection is sufficient. These results are stable across five random seeds ($std \leq 0.006$ per class), and the architectural advantage pattern is replicated on an Edge-IIoTset-inspired feasibility check.

Returning to the research questions: (RQ1) Adversarial evasion and firmware implant represent the critical intersection where neither layer alone suffices; the threat-to-mechanism mapping explains this through the structural properties of each attack class. (RQ2) PVE residuals augment EADE by amplifying the physics-process discrepancy signal that is present in residual space but absent in raw sensor space for physics-plausible attacks. (RQ3) PhySec-Edge improves F1 by +0.025 over AI-only at seed = 42 (*mean $\Delta F1 = +0.024$ across five seeds*), stable across five seeds; ablation confirms the two-mechanism decomposition. (RQ4) Gateway latency of 40–95 ms is compatible with soft real-time SCADA monitoring; Edge-IIoTset feasibility check confirms architectural consistency.

The primary limitation is the synthetic evaluation scope. Future work will validate PhySec-Edge on public SWaT, BATADAL, WADI, and HAI benchmarks with the full LSTM/GNN EADE implementation and will investigate adaptive calibration through the EvoPhySec multi-objective evolutionary framework. The formal specification of the residual-sharing architecture provided in Section 4 is intended to facilitate community replication and extension of this research direction. Figure 6 summarizes the planned research trajectory, from the current architectural feasibility study through cross-dataset generalization, online streaming deployment, and federated multi-site calibration sharing.

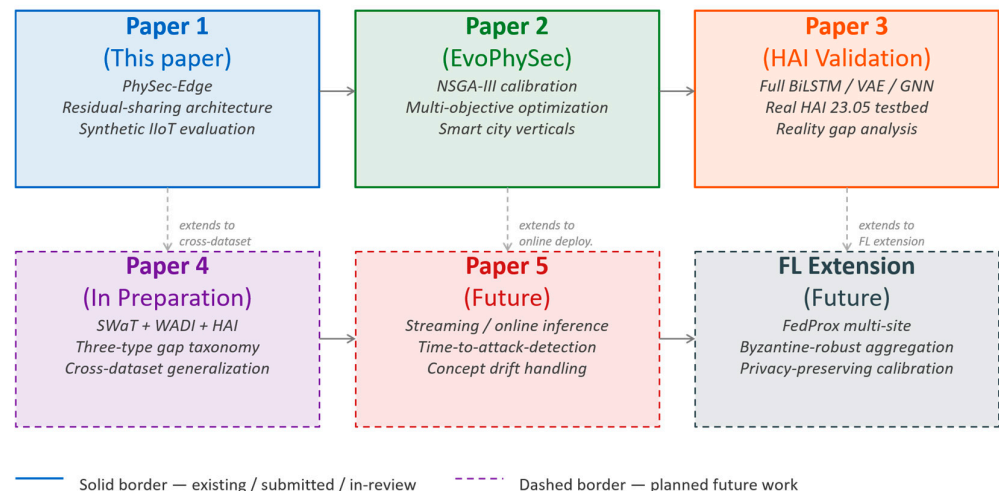


Figure 6. PhySec-Edge research program roadmap. Solid borders indicate existing or submitted work; dashed borders indicate planned future studies. Papers 1–3 form the core synthetic-to-real validation trajectory; Papers 4–5 extend to cross-dataset generalization and online deployment.

Author Contributions: Conceptualization, D.R.; methodology, D.R.; software, D.R., J.J., T.P. and N.S.; validation, D.R., J.J., T.P., P.K. and N.S.; formal analysis, D.R.; investigation, D.R., J.J., P.K. and N.S.; data curation, J.J., P.K. and N.S.; writing—original draft preparation, D.R.; writing—review and editing, D.R., J.J., P.K. and N.S.; visualization, J.J., P.K. and T.P.; supervision, D.R.; project administration, D.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The synthetic dataset used in this study was generated based on SWaT-inspired structural and statistical parameters, as described in Section 5.1 and Table 2; the original SWaT dataset was not used directly. The data generation methodology and all experimental parameters are fully specified in the manuscript to enable independent replication. Additional details regarding dataset generation and evaluation procedures are available from the corresponding author upon reasonable request. The public Edge-IIoTset dataset referenced in the feasibility check is available at <https://doi.org/10.1109/ACCESS.2022.3165809>. Researchers wishing to access the original SWaT dataset may submit a request through the iTrust Centre at https://itrust.sutd.edu.sg/itrust-labs_datasets/ (accessed on 29 May 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

APT	Advanced Persistent Threat
BATADAL	Battle of the Attack Detection Algorithms
CISA	Cybersecurity and Infrastructure Security Agency
CNN	Convolutional Neural Network
EADE	Edge AI Detection Engine
FDI	Fault Detection and Isolation
FL	Federated Learning
FPR	False Positive Rate
GNN	Graph Neural Network
HAI	Hardware-in-the-Loop Augmented ICS Security
ICS	Industrial Control System
IEC	International Electrotechnical Commission
IIoT	Industrial Internet of Things
IoT	Internet of Things
LSTM	Long Short-Term Memory
NIST	National Institute of Standards and Technology
NIS	Normalized Innovation Squared
OPC-UA	Open Platform Communications—Unified Architecture
PLC	Programmable Logic Controller
PVE	Physics Validation Engine
RQ	Research Question
RS	Randomized Smoothing
SCADA	Supervisory Control and Data Acquisition
SIS	Safety-Instrumented System
SR	System Requirement (IEC 62443)
SWaT	Secure Water Treatment (testbed)
VAE	Variational Autoencoder
WADI	Water Distribution (testbed)

References

- Salih, K.O.M.; Rashid, T.A.; Radovanovic, D.; Bacanin, N. A Comprehensive Survey on the Internet of Things with the Industrial Marketplace. *Sensors* **2022**, *22*, 730. <https://doi.org/10.3390/s22030730>.
- Dragos Inc. 2023 ICS/OT Cybersecurity Year in Review; Dragos: Hanover, MD, USA, 2024. Available online: <https://www.dragos.com/year-in-review> (accessed on 29 May 2026).
- Langner, R. Stuxnet: Dissecting a Cyberwarfare Weapon. *IEEE Secur. Priv.* **2011**, *9*, 49–51. <https://doi.org/10.1109/MSP.2011.67>.
- Cybersecurity and Infrastructure Security Agency (CISA). *Stop Ransomware: A Guide to Critical Infrastructure Security*; CISA Advisory AA23-061A; CISA: Washington, DC, USA, 2023. Available online: <https://www.cisa.gov/news-events/cybersecurity-advisories/aa23-061a> (accessed on 29 May 2026).
- Adepu, S.; Mathur, A. An Investigation into the Response of a Water Treatment System to Cyber Attacks. In *Proceedings of the 2016 IEEE 17th International Symposium on High Assurance Systems Engineering (HASE)*, Orlando, FL, USA, 7–9 January 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 141–148. <https://doi.org/10.1109/HASE.2016.14>.
- Ozmen, M.O.; Song, R.; Farrukh, H.; Celik, Z.B. Evasion Attacks and Defenses on Smart Home Physical Event Verification. In *Proceedings of the 30th Annual Network and Distributed System Security Symposium (NDSS 2023)*, San Diego, CA, USA, 27 February–3 March 2023; Internet Society: Reston, VA, USA, 2023. Available online: <https://www.ndss-symposium.org/ndss-paper/evasion-attacks-and-defenses-on-smart-home-physical-event-verification> (accessed on 29 May 2026).
- Feng, C.; Palleti, V.R.; Mathur, A.; Chana, D. A Systematic Framework to Generate Invariants for Anomaly Detection in Industrial Control Systems. In *Proceedings of the 26th Annual Network and Distributed System Security Symposium (NDSS 2019)*, San Diego, CA, USA, 24–27 February 2019; Internet Society: Reston, VA, USA, 2019. Available online: <https://www.ndss-symposium.org/ndss-paper/a-systematic-framework-to-generate-invariants-for-anomaly-detection-in-industrial-control-systems> (accessed on 29 May 2026).
- Inoue, J.; Yamagata, Y.; Chen, Y.; Poskitt, C.M.; Sun, J. Anomaly Detection for a Water Treatment System Using Unsupervised Machine Learning. In *Proceedings of the 2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, New Orleans, LA, USA, 18–21 November 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1058–1065. <https://doi.org/10.1109/ICDMW.2017.149>.
- Aoudi, W.; Iturbe, M.; Almgren, M. Truth Will Out: Departure-Based Process-Level Detection of Stealthy Attacks on Control Systems. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, Toronto, ON, Canada, 15–19 October 2018; ACM: New York, NY, USA, 2018; pp. 817–831. <https://doi.org/10.1145/3243734.3243781>.
- Roth, T.; Restuccia, F. Physics-Guided Anomaly Detection in IoT Networks Using a Hybrid Model. In *Proceedings of IEEE INFOCOM 2022*, London, UK, 2–5 May 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1560–1569. <https://doi.org/10.1109/INFOCOM48880.2022.9796715>.
- IEC 62443-3-3; Security for Industrial Automation and Control Systems. International Electrotechnical Commission: Geneva, Switzerland, 2013.
- National Institute of Standards and Technology. *Guide to Industrial Control Systems (ICS) Security*; NIST SP 800-82 Rev. 3; NIST: Gaithersburg, MD, USA, 2023. <https://doi.org/10.6028/NIST.SP.800-82r3>.
- US Food and Drug Administration. *21 CFR Part 11: Electronic Records; Electronic Signatures—Scope and Application*; FDA: Silver Spring, MD, USA, 2023. Available online: <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/part-11-electronic-records-electronic-signatures-scope-and-application> (accessed on 29 May 2026).
- Goh, J.; Adepu, S.; Junejo, K.N.; Mathur, A. A Dataset to Support Research in the Design of Secure Water Treatment Systems. In *Proceedings of the 11th International Conference on Critical Information Infrastructures Security (CRITIS 2016)*, Paris, France, 10–12 October 2016; Springer: Cham, Switzerland, 2017; pp. 88–99. https://doi.org/10.1007/978-3-319-71368-7_8.
- Kravchik, M.; Shabtai, A. Detecting Cyber Attacks in Industrial Control Systems Using Convolutional Neural Networks. In *Proceedings of the 2018 ACM Workshop on Cyber-Physical Systems Security and Privacy (CPS-SPC)*, Toronto, ON, Canada, 19 October 2018; ACM: New York, NY, USA, 2018; pp. 72–83. <https://doi.org/10.1145/3264888.3264896>.
- Feng, C.; Li, T.; Chana, D. Multi-Level Anomaly Detection in Industrial Control Systems via Package Signatures and LSTM Networks. In *Proceedings of the 47th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN 2017)*, Denver, CO, USA, 26–29 June 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 261–272. <https://doi.org/10.1109/DSN.2017.34>.
- Deng, A.; Hooi, B. Graph Neural Network-Based Anomaly Detection in Multivariate Time Series. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI 2021)*, Virtual, 2–9 February 2021; AAAI Press: Palo Alto, CA, USA, 2021; Volume 35, pp. 4027–4035. <https://doi.org/10.1609/aaai.v35i5.16523>.
- Lin, L.; Gu, A.; Min, F.; Zhou, S. Multi-Level Graph Attention Network-Based Anomaly Detection in Industrial Control System. *Actuators* **2025**, *14*, 210. <https://doi.org/10.3390/act14050210>.

19. Goodfellow, I.; Shlens, J.; Szegedy, C. Explaining and Harnessing Adversarial Examples. In Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015), San Diego, CA, USA, 7–9 May 2015.
20. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards Deep Learning Models Resistant to Adversarial Attacks. In Proceedings of the 6th International Conference on Learning Representations (ICLR 2018), Vancouver, BC, Canada, 30 April–3 May 2018.
21. Ding, D.; Han, Q.-L.; Xiang, Y.; Ge, X.; Zhang, X.-M. A Survey on Security Control and Attack Detection for Industrial Cyber-Physical Systems. *Neurocomputing* **2018**, *275*, 1674–1683. <https://doi.org/10.1016/j.neucom.2017.10.009>.
22. Cohen, J.M.; Rosenfeld, E.; Kolter, J.Z. Certified Adversarial Robustness via Randomized Smoothing. In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019), Long Beach, CA, USA, 9–15 June 2019*; PMLR: London, UK, 2019; Volume 97, pp. 1310–1320. Available online: <https://proceedings.mlr.press/v97/cohen19c.html> (accessed on 29 May 2026).
23. Jeffrey, N.; Tan, Q.; Villar, J.R. A Hybrid Methodology for Anomaly Detection in Cyber–Physical Systems. *Neurocomputing* **2024**, *568*, 127068. <https://doi.org/10.1016/j.neucom.2023.127068>.
24. Birihanu, E.; Barcsa-Szabo, A.; Lendak, I. Explainable Correlation-Based Anomaly Detection for Industrial Control Systems. *Front. Artif. Intell.* **2024**, *7*, 1508821. <https://doi.org/10.3389/frai.2024.1508821>.
25. Shyaa, M.A.; Ibrahim, N.F.; Zainol, Z.; Abdullah, R.; Anbar, M.; Haque, M.R. AI-Based Anomaly Detection in Industrial Control and Cyber–Physical Systems: A Data-Type-Oriented Systematic Review. *Electronics* **2026**, *15*, 20. <https://doi.org/10.3390/electronics15010020>.
26. Khoury, N.; Shameli-Sendi, A. Physics-Guided Recurrent Neural Network for Anomaly Detection in ICS. In *Proceedings of the 2022 ACM Workshop on Artificial Intelligence and Security (AISec), Los Angeles, CA, USA, 7 November 2022*; ACM: New York, NY, USA, 2022; pp. 55–66. <https://doi.org/10.1145/3560830.3563731>.
27. Wickramasinghe, C.S.; Marino, D.L.; Manic, M. ResNet Autoencoders for Unsupervised Feature Learning From High-Dimensional Data: Deep Models Resistant to Performance Degradation. *IEEE Access* **2021**, *9*, 40511–40520. <https://doi.org/10.1109/ACCESS.2021.3064819>.
28. Nguyen, D.C.; Ding, M.; Pathirana, P.N.; Seneviratne, A.; Li, J.; Poor, H.V. Federated Learning for Internet of Things: A Comprehensive Survey. *IEEE Commun. Surv. Tutor.* **2021**, *23*, 1622–1658. <https://doi.org/10.1109/COMST.2021.3075439>.
29. Li, T.; Sahu, A.K.; Zaheer, M.; Sanjabi, M.; Smola, A.; Smith, V. Federated Optimization in Heterogeneous Networks. In *Proceedings of the 3rd Conference on Machine Learning and Systems (MLSys 2020), Austin, TX, USA, 2–4 March 2020*; MLSys: Stanford, CA, USA, 2020; Volume 2, pp. 429–450. Available online: https://proceedings.mlsys.org/paper_files/paper/2020/hash/1f5fe83998a09396ebe6477d9475ba0c-Abstract.html (accessed on 29 May 2026).
30. Blanchard, P.; Mhamdi, E.M.E.; Guerraoui, R.; Stainer, J. Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017), Long Beach, CA, USA, 4–9 December 2017*; Curran Associates: Red Hook, NY, USA, 2017; Volume 30. Available online: https://proceedings.neurips.cc/paper_files/paper/2017/hash/f4b9ec30ad9f68f89b29639786cb62ef-Abstract.html (accessed on 29 May 2026).
31. IEC 61508; Functional Safety of E/E/PE Safety-Related Systems. International Electrotechnical Commission: Geneva, Switzerland, 2010.
32. IEC 61511; Functional Safety—Safety Instrumented Systems for the Process Industry Sector. International Electrotechnical Commission: Geneva, Switzerland, 2016.
33. Ferrag, M.A.; Friha, O.; Hamouda, D.; Maglaras, L.; Janicke, H. Edge-IIoTset: A New Comprehensive Realistic Cyber Security Dataset of IoT and IIoT Applications for Centralized and Federated Learning. *IEEE Access* **2022**, *10*, 40281–40306. <https://doi.org/10.1109/ACCESS.2022.3165809>.
34. Lin, J.; Chen, W.; Lin, Y.; Cohn, J.; Gan, C.; Han, S. MCUNet: Tiny Deep Learning on IoT Devices. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*; Curran Associates: Red Hook, NY, USA, 2020; Volume 33, pp. 11711–11722. Available online: https://proceedings.neurips.cc/paper_files/paper/2020/hash/86c51678350f656dcc7f490a43946ee5-Abstract.html (accessed on 29 May 2026).
35. Ahmed, C.M.; Palleti, V.R.; Mathur, A.P. WADI: A Water Distribution Testbed for Research in the Design of Secure Cyber-Physical Systems. In *Proceedings of the 3rd International Workshop on Cyber-Physical Systems for Smart Water Networks (CySWater), Pittsburgh, PA, USA, 21 April 2017*; ACM: New York, NY, USA, 2017; pp. 25–28. <https://doi.org/10.1145/3055366.3055375>.

36. Taormina, R.; Galelli, S.; Tippenhauer, N.O.; Salomons, E.; Ostfeld, A.; Eliades, D.G.; Aghashahi, M.; Sundararajan, R.; Pourahmadi, M.; Banks, M.K.; et al. Battle of the Attack Detection Algorithms: Disclosing Cyber Attacks on Water Distribution Networks. *J. Water Resour. Plan. Manag.* **2018**, *144*, 04018048. [https://doi.org/10.1061/\(ASCE\)WR.1943-5452.0000969](https://doi.org/10.1061/(ASCE)WR.1943-5452.0000969).
37. Shin, H.-K.; Lee, W.; Yun, J.-H.; Kim, H. HAI 1.0: HIL-Based Augmented ICS Security Dataset. In *Proceedings of the 13th USENIX Workshop on Cyber Security Experimentation and Test (CSET 20), Boston, MA, USA (Virtual), 10 August 2020*; USENIX Association: Berkeley, CA, USA, 2020. Available online: <https://www.usenix.org/conference/cset20/presentation/shin> (accessed on 29 May 2026).
38. Lim, L.-H.; Ong, L.-Y.; Leow, M.-C. Federated Learning for Anomaly Detection: A Systematic Review on Scalability, Adaptability, and Benchmarking Framework. *Future Internet* **2025**, *17*, 375. <https://doi.org/10.3390/fi17080375>.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.