

Cyber-Physical System Security through Anomaly-Aware Intelligent Monitoring

Ravichandran S^{1*}, Kalyanasundaram H² and Elangovan H³

¹Department of Artificial Intelligence and Data Science, Amaravathi Institute of Cyber-Physical Systems, Vietnam
Email: sneha.ravichandran@amaravathi.edu

ORCID: <https://orcid.org/5278-2799-7447-8878>

²Department of Information Technology, Sharavati Center for Robotics and Automation, Indonesia Email:
harish.kalyanasundaram@sharavati.edu

ORCID: <https://orcid.org/9186-5772-1074-1932>

³Department of Robotics and Automation, Chaliyar Institute of Quantum Computing, Philippines Email:
harish.elangovan@chaliyar.edu

ORCID: <https://orcid.org/7443-3689-8825-782X>

ABSTRACT

Cyber-physical systems increasingly govern water treatment, power distribution, and industrial control processes whose sensors and actuators were rewired onto networked infrastructure faster than their security models matured, a gap the Stuxnet attack exposed starkly by falsifying monitored sensor readings while physically sabotaging a real process. This article synthesizes the literature on anomaly-aware intelligent monitoring as a response to that gap: learning a statistical or graph-structured model of normal multivariate physical-process dynamics, from recurrent autoencoders and convolutional ensembles through graph deviation networks and generative adversarial architectures, and flagging departures from it regardless of whether they originate from cyberattack, sensor fault, or equipment failure. We trace this pattern from early physical-invariant and shallow-learning baselines through the deep sequence and graph-based models now dominant on shared industrial testbeds, and extend the discussion to distributed, federated, and digital-twin-based monitoring architectures designed for privacy-aware, multi-site deployment. Read together, the evidence supports intelligent monitoring as a coherent and rapidly maturing architectural pattern for cyber-physical system security, while leaving the precision-recall tradeoff on subtle and gradual attacks, adversarial robustness of learned detectors, and validated privacy-preserving distributed deployment as substantially open research questions.

Keywords: cyber-physical systems; anomaly detection; industrial control systems; intrusion detection; intelligent monitoring

Citation: Ravichandran S, Kalyanasundaram H and Elangovan H [2023]. Cyber-Physical System Security through Anomaly-Aware Intelligent Monitoring. DOI: <https://doi.org/10.5281/zenodo.22843862>

Copyright: © 2023 by the authors. Open access under [CC BY 4.0 license](https://creativecommons.org/licenses/by/4.0/).

Article Information: Received: 3 December 2022 Accepted: 10 March 2023 Published: 16 June 2023

Article Class: Research Article

1. Introduction

A cyber-physical system (CPS) is, at bottom, a control loop wearing a network stack: sensors measure a physical process, a computation decides what to do about it, and actuators push that decision back into the physical world. Water treatment plants, power substations, chemical reactors, and building automation systems all share this shape, and all of them have, over the last fifteen years, been quietly rewired to run that loop over Ethernet, Wi-Fi, or cellular links rather than dedicated analog wiring. The convenience is real — remote diagnostics, predictive maintenance, centralized supervisory control — and so is the exposure it creates. Stuxnet remains the reference case for what happens when that exposure is exploited with intent: a worm that reprogrammed Siemens programmable logic controllers to intermittently over-spin uranium enrichment centrifuges while feeding operators falsified, entirely normal-looking sensor readings, so that the physical damage accumulated for months before anyone suspected sabotage rather than mechanical failure [1]. That case is now over a decade old, but the structural lesson it taught — that a CPS attacker's goal is usually to manipulate the physical process while keeping the monitoring layer quiet — has only become more relevant as control systems have grown more networked, more heterogeneous, and more dependent on software they did not originate.

Conventional information-security tooling is a poor match for this problem. Signature-based intrusion detection, firewall rules, and access control lists are built around the premise that an attack looks like anomalous *traffic*: a scan, an unfamiliar payload, a protocol violation. A well-executed CPS attack, by contrast, can traverse the network using entirely legitimate commands — a real engineer's workstation, a real vendor protocol, a real setpoint change — while the actual malicious effect is a physical one: a valve held open too long, a temperature setpoint nudged outside a safe band, a voltage phasor quietly drifting from its true value. Detecting that kind of attack requires watching the *physical* signal, not just the network packets that carry the commands, and asking whether the process is behaving the way physics says it should. This is the premise behind anomaly-aware intelligent monitoring: instrument the physical process itself, learn what its normal dynamics look like, and flag departures from that learned model regardless of whether they originate from

a cyberattack, sensor fault, or genuine equipment failure. Mitchell and Chen's early survey of CPS intrusion detection techniques already identified this physical-process orientation as the field's defining methodological split from conventional network-centric intrusion detection, and every survey that has followed it has, in one way or another, organized itself around the same distinction [2].

This article synthesizes roughly a decade of research on that premise, tracing it from its statistical and control-theoretic origins through the deep learning architectures that now dominate the literature, and asking, throughout, what the underlying papers actually report rather than what the pattern *should* achieve in principle. We draw on foundational and survey material characterizing the CPS threat landscape [3,2,4,5], the industrial control systems and water-treatment testbed literature that has become the field's de facto shared benchmark [6,7,8,9,10,11], the multivariate time-series anomaly detection literature developed largely outside the security community but rapidly adopted by it [12,13,14], adjacent work on power-grid false data injection [15], and the generative, federated, and digital-twin architectures now being proposed to extend intelligent monitoring toward distributed, privacy-aware, and simulation-augmented deployments [16,17,18,19,20,21]. We also situate the discussion against the older data-mining tradition of network intrusion detection, whose benchmark datasets and evaluation habits — for better and for worse — still shape how CPS anomaly detectors are validated today [22,23].

Our organizing claim is a modest one: intelligent monitoring for CPS security has matured from a set of scattered proposals into a recognizable architectural pattern — learn a model of normal multivariate process dynamics, score deviations from it, and increasingly do so under constraints of distributed deployment and limited labeled attack data — but the evidence base behind that pattern remains narrower than its growing popularity might suggest, concentrated on a small number of shared testbeds and simulated attack scenarios. We try to be precise about which claims in the literature are well-supported by repeated, independent measurement and which rest on a single paper's reported numbers, because that distinction matters enormously to anyone deciding whether to trust this

technology in a real plant.

2. Foundations of Cyber-Physical System Security

CPS security differs from conventional IT security along three axes that recur throughout the literature. The first is consequence: a compromised database leaks information, but a compromised control loop can cause physical harm — equipment damage, environmental release, or, in the worst documented cases, risk to human safety — which changes the acceptable false-negative rate in ways that pure information-security thinking does not always anticipate. The second is longevity: industrial control hardware and the protocols it speaks (Modbus, DNP3, various proprietary fieldbus variants) were frequently designed decades before networked operation was a design consideration, and many such devices remain in service for twenty or thirty years, meaning security has to be retrofitted onto plant equipment that cannot always be patched, rebooted, or even taken offline for testing. The third is the availability-over-confidentiality priority ordering: where enterprise IT security conventionally ranks confidentiality first, CPS operators overwhelmingly prioritize availability and safety of the physical process, which shapes what kinds of defensive measures — active blocking versus passive monitoring, for instance — are operationally acceptable at all. Humayed et al.'s survey of CPS security lays out this taxonomy in detail, cataloguing the space along the traditional confidentiality-integrity-availability axes crossed against the cyber, physical, and cyber-physical attack surfaces a CPS presents, and remains a useful map of the vocabulary the rest of this field inherits [3].

Stuxnet is worth dwelling on a little further because it previewed nearly every theme this article returns to. The attack did not merely disable the centrifuges' control system; it recorded roughly three weeks of legitimate sensor data and replayed that recording to the operator's monitoring displays while the physical process was actually being sabotaged, defeating any monitoring approach that trusted the reported sensor values without independently modeling the physical dynamics those values were supposed to represent [1]. That specific tactic — falsify what the monitor sees rather than merely block what the operator can do — is precisely the scenario that motivates physics-aware anomaly detection: if a monitoring system learns the

joint statistical structure across many correlated physical measurements (flow, pressure, level, current, vibration) rather than trusting any single reported value in isolation, a falsified subset of sensors becomes detectable as an inconsistency in the overall pattern rather than a plausible individual reading. Mitchell and Chen's taxonomy of CPS intrusion detection techniques was already organizing the field along exactly this "audit source" dimension in 2014 — network-based, host-based, or physical-process-based detection — and their observation that physical-process-based techniques were, at that point, comparatively underdeveloped relative to network- and host-based IDS techniques is a useful marker of how much the field has shifted since [2].

Luo et al.'s more recent survey of deep-learning-based anomaly detection in cyber-physical systems captures that shift directly, cataloguing the rapid growth in papers applying recurrent, convolutional, graph, and generative architectures to physical-process monitoring across domains including industrial control, smart grids, autonomous vehicles, and medical devices [4]. Abshari and Sridhar's 2025 review covers substantially the same ground with the benefit of several additional years of literature, and is candid about a recurring critique that runs through this article as well: the great majority of reported results come from a small number of shared public testbeds evaluated under simulated attack scenarios, rather than from field deployments on operational infrastructure, which limits how confidently detection-rate numbers reported in one paper generalize to a genuinely different plant [5].

3. Anomaly Detection Paradigms for Intelligent Monitoring

Statistical and Shallow Machine Learning Baselines

Before the deep learning wave, and still very much in active use as a baseline and, in resource-constrained deployments, as a production technique in its own right, CPS anomaly detection drew heavily on the broader data-mining and machine-learning-for-cybersecurity literature. Buczak and Guven's widely cited survey of data mining and machine learning methods for cyber-security intrusion detection catalogued classification, clustering, and hybrid approaches — decision trees, support vector machines, Bayesian networks, genetic algorithms — that predate the deep learning era but established the evaluation conventions

(precision, recall, F-measure, ROC analysis against labeled benchmark traffic) that much of the later CPS literature still uses [22]. That evaluation tradition inherited its benchmark datasets from network intrusion detection specifically: Tavallaee et al.'s detailed statistical analysis of the KDD Cup 99 dataset identified serious redundancy and distributional problems in that decades-old benchmark and proposed the NSL-KDD variant to address them, and the paper's broader methodological point — that a benchmark's quirks can silently determine which detection method looks best — is a caution that later CPS-specific benchmarks have not always fully escaped [23].

Within CPS proper, Inoue et al.'s early application of unsupervised machine learning to the Secure Water Treatment (SWaT) testbed is a useful baseline case study precisely because it is honest about shallow methods' limitations. Comparing a deep neural network autoencoder against a one-class support vector machine on SWaT's labeled attack scenarios, the authors reported an F-measure of 0.80281 for the DNN (precision 0.983, recall 0.678) against 0.79628 for the SVM (precision 0.925, recall 0.699), with the DNN's higher precision offset almost exactly by the SVM's higher recall [8]. Neither method, notably, reliably caught every attack type: the paper reports that both approaches detected attacks producing constant anomalous sensor values fairly well but struggled with attacks involving gradual drift or anomalous actuator movements, a limitation that recurs across nearly every anomaly-detection paper surveyed for this article regardless of model architecture, and which points to a genuine open problem rather than a solved one.

Deep Sequence Models and Autoencoders

The deep learning generation of CPS anomaly detectors is organized around a shared intuition: physical processes are temporally correlated, multivariate, and governed by physical constraints that a sufficiently expressive model can learn implicitly from historical operating data, without requiring an engineer to hand-specify what "normal" looks like. Recurrent architectures, and LSTM-based autoencoders and variational autoencoders in particular, were the first deep architectures to be widely applied to this setting, learning to reconstruct or forecast a window of sensor readings and flagging large reconstruction or prediction error as anomalous. Su et al.'s OmniAnomaly is a

representative and influential instance of this family, combining a stochastic recurrent neural network with a planar normalizing flow to model the temporal dependence and the non-Gaussian structure of multivariate device metrics jointly, and reporting strong F1 performance (above 0.90 on several of the server-monitoring datasets evaluated in the original paper) alongside an interpretability mechanism that attributes an anomaly score back to the individual metrics that contributed most to it — a feature that matters a great deal operationally, because a plant operator investigating an alarm needs to know *which* sensor looks wrong, not just that something, somewhere, does [12].

Kravchik and Shabtai's convolutional approach to the same problem, evaluated directly on the SWaT testbed's forty-one physical sensors and actuators, reported that a four-layer 1D CNN ensembled across the plant's physical process stages achieved an F1 score of 0.9206 (precision 1.0, recall 0.8529), correctly flagging 32 of the 36 labeled attacks in the SWaT test set, and that this outperformed the DNN, one-class SVM, and TABOR baselines evaluated in the same paper [9]. That precision-recall tradeoff is worth pausing on: a detector with perfect precision and 85% recall is operationally attractive because it never cries wolf, but it also means roughly one in seven attacks in this evaluation went undetected, a gap the paper is explicit about rather than glossing over. Read alongside Inoue et al.'s DNN and SVM results on essentially the same testbed, a consistent picture emerges across this line of work: no single architecture evaluated to date achieves both very high precision and very high recall simultaneously on the full range of attack types SWaT's designers built into the benchmark, and different architectures trade off differently between the two.

4. Graph- and Generative-Model Architectures for Multivariate Monitoring

A newer generation of architectures has tried to close that gap by modeling the relationships *among* sensors explicitly, rather than treating a multivariate time series as an undifferentiated vector at each timestep. Deng and Hooi's Graph Deviation Network (GDN) learns a graph structure over sensor channels — which sensors are meaningfully correlated with which others under normal operation — jointly with a graph attention forecasting model, and flags anomalies where a sensor's observed

value deviates from what the model, conditioned on its learned neighbors' current values, predicts it should be. Evaluated on both SWaT and the WADI (Water Distribution) testbed, GDN reported an F1 score of 0.81 on SWaT and 0.57 on WADI, compared with 0.77 and 0.37 respectively for the MAD-GAN baseline, 0.74 and 0.25 for an LSTM-VAE baseline, and 0.39 and 0.36 for DAGMM; on WADI specifically, the authors report GDN achieving 54% higher F-measure than the next-best baseline evaluated in the paper [14]. The gap between the SWaT and WADI numbers for every method in that comparison — every architecture performs noticeably worse on WADI than on SWaT — is itself an informative data point: WADI is a larger, more complex distribution network with substantially fewer and more subtle labeled attacks relative to its sensor count, and the consistent cross-method degradation suggests the difficulty is intrinsic to the benchmark's attack design rather than an artifact of any one model.

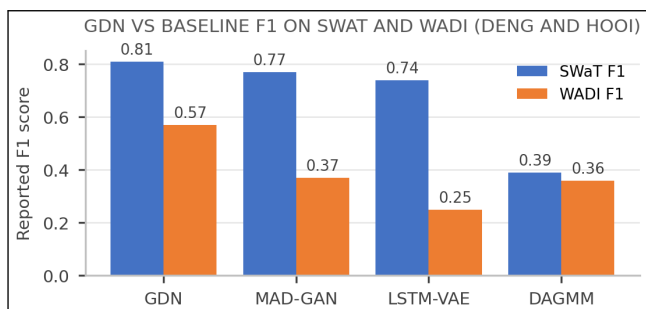


Figure 1: F1 scores reported by Deng and Hooi for their Graph Deviation Network (GDN) against three baseline multivariate anomaly detectors on the SWaT and WADI industrial water-testbed benchmarks, illustrating both GDN's reported advantage and the consistent cross-method performance drop from SWaT to the larger, sparser-labeled WADI testbed [14].

Generative adversarial architectures represent a second major branch of this newer generation. Li et al.'s MAD-GAN trains a generator to produce plausible multivariate sensor sequences and a discriminator to distinguish real from generated sequences, then repurposes the trained discriminator (and a reconstruction-based scoring signal derived from mapping a real test sequence back into the generator's latent space) as the anomaly detector at test time, on the premise that a GAN trained exclusively on attack-free operating data will struggle to both generate and reconstruct sequences that depart from the learned normal manifold [13]. Lim et al.'s review of GANs for network security anomaly detection situates this kind of approach within the broader GAN literature, cataloguing not just detection applications but the parallel and less

comfortable observation that GANs are equally capable of being used offensively — to synthesize evasive attack traffic designed specifically to fool a trained detector — a dual-use tension the review treats as a first-order concern for the field rather than a footnote [16]. Alauthman et al.'s more recent, broader survey of GAN-based intrusion detection systems extends this cataloguing effort across a wider span of GAN variants and application domains, and its stated emphasis on open research directions — few-shot and zero-day generalization, computational cost of adversarial training at line rate, and the evasion-robustness question Lim et al. raise directly — gives a reasonably current sense of where the sub-field considers itself unfinished [17].

It is worth naming plainly what these generative and graph-based architectures share with the recurrent and convolutional generation that preceded them: all of them are, at root, learning a model of normal multivariate correlation structure and flagging statistical deviation from it. The architectural sophistication has increased — explicit relational graphs, adversarial training objectives, normalizing flows for non-Gaussian latent structure — but the underlying detection logic, unsupervised or semi-supervised deviation-from-normal scoring trained on attack-free historical data, has been remarkably stable across a decade of otherwise rapid architectural churn. That stability is, in our reading, a genuine strength of the field's convergence rather than a lack of imagination: it reflects a hard constraint of the domain, namely that labeled attack examples are scarce and expensive (often only a few dozen per testbed run, as the SWaT and WADI numbers above illustrate), which structurally favors any method that can learn from abundant normal-operation data rather than requiring a large labeled attack corpus that, for most real plants, simply does not exist.

5. Testbeds, Benchmarks, and the Industrial Control Systems Case

Much of what is known empirically about intelligent CPS monitoring rests on a small number of physical or physically realistic testbeds, and the Secure Water Treatment (SWaT) testbed built at the Singapore University of Technology and Design has become the field's closest thing to a shared standard. Goh et al.'s dataset paper documents SWaT's design directly: a scaled-down but fully operational six-stage water

treatment process instrumented with programmable logic controllers, twenty-five sensors, and actuators representative of a real plant, run continuously to produce eleven days of operational data, seven days of attack-free normal operation followed by four days during which thirty-six distinct single- and multi-stage cyberattacks were launched against the live process [6]. That paper's contribution is not a detection algorithm at all but the dataset and its documented ground truth, and its influence is visible throughout this article: essentially every deep-learning detection paper discussed above (Inoue et al., Kravchik and Shabtai, and Deng and Hooi among them) either evaluates directly on SWaT or on its sibling WADI testbed, making Goh et al.'s infrastructure paper one of the more consequential pieces of the entire literature despite reporting no detection results of its own.

Table 1: Representative detection results reported on the SWaT and WADI water-treatment testbeds across the architectures discussed in this article, alongside the testbeds' own documented scale [6,9,8,14,13].

Technique	Source	Testbed	Reported Result
SWaT testbed design	Goh et al. [6]	SWaT	25 sensors/actuators; 7 days normal operation followed by 36 attacks over 4 days
1D CNN ensemble	Kravchik and Shabtai [9]	SWaT	F1 = 0.92 (precision 1.00, recall 0.85); 32 of 36 attacks detected
Unsupervised DNN autoencoder	Inoue et al. [8]	SWaT	F1 = 0.803 (precision 0.983, recall 0.678)
One-class SVM	Inoue et al. [8]	SWaT	F1 = 0.796 (precision 0.925, recall 0.699)
Graph Deviation Network (GDN)	Deng and Hooi [14]	SWaT / WADI	F1 = 0.81 / 0.57; 54% higher F-measure than next-best baseline on WADI
MAD-GAN	Li et al. [13], as evaluated in [14]	SWaT / WADI	F1 = 0.77 / 0.37

Adepu and Mathur's complementary work on the same testbed approaches the detection problem from a control-theoretic rather than a purely statistical-learning angle, distributing physical invariant checks — hand-derived relationships that must hold across the plant's water levels, flow rates, and valve states under correct physical operation — across each of SWaT's six process stages, so that a multi-point coordinated attack spanning several stages can be caught locally at each affected stage rather than requiring a single centralized detector to reason about the whole plant at once [7]. This invariant-based approach is a useful counterpoint to the learned-model architectures surveyed above: it

requires domain engineering effort up front (someone has to derive the physical invariants correctly) but does not depend on a learned statistical model at all, and is consequently far less vulnerable to the kind of adversarial evasion that concerns Lim et al. and Alauthman et al. for purely data-driven detectors, since there is no learned decision boundary to probe or fool [16,17].

Two survey papers help situate where the applied literature stands as of the past two years. Umer et al.'s survey of machine learning for industrial control systems intrusion detection reviews the space of applied techniques, datasets, and evaluation practices and is explicit about the field's persistent gaps: heavy reliance on a handful of public testbeds, limited attention to adversarial robustness, and a shortage of standardized, cross-testbed evaluation protocols that would let a reader compare, say, a graph-based detector evaluated on SWaT against a recurrent detector evaluated on a different ICS dataset with any real confidence [10]. Termanini et al.'s more recent survey of machine-learning-based network intrusion detection for industrial control systems covers substantially overlapping ground and reinforces the same conclusion from a slightly different angle, situating physical-process anomaly detection alongside network-traffic-based ICS intrusion detection and noting that the two approaches are more often studied in isolation than combined, despite the intuitive case — made forcefully by the Stuxnet case discussed earlier — that a sophisticated attacker will try to look normal at whichever layer is being watched, which argues for monitoring both simultaneously rather than choosing one [11].

6. Distributed, Federated, and Privacy-Aware Monitoring Architectures

Real CPS deployments rarely consist of a single testbed's worth of sensors feeding a single centralized model, and a growing strand of the literature addresses what changes when monitoring has to happen across many geographically distributed sites, each with its own sensitive operational data, that cannot or should not be pooled into one central training corpus. Federated learning, introduced by McMahan et al. as a way to train a shared model across many clients' local data without moving that raw data off-device, offers a natural fit: a federation of plants, substations, or building sites could jointly train an anomaly detector on their collective

operating experience while keeping each site's raw sensor logs — which may reveal proprietary process parameters or operational details an operator would rather not centralize — local [19]. Mothukuri et al.'s federated-learning-based anomaly detection framework for IoT security applies exactly this pattern to distributed device monitoring, training detection models across many participating devices via federated averaging rather than centralizing raw telemetry, and the framework's design case for CPS-adjacent IoT deployments — where sensor data can be commercially sensitive and bandwidth to a central server is often constrained — maps closely onto the industrial and infrastructure monitoring settings this article is concerned with [18].

Federated monitoring introduces its own security questions, and the most immediate is privacy: aggregating model updates across a population of physically distributed sites is not automatically private just because raw sensor data never left each site, since a sufficiently determined analysis of the shared model updates can, in principle, leak information about the underlying local data. Wei et al.'s analysis of differential privacy in federated averaging formalizes what happens when calibrated noise is injected into client updates before aggregation, characterizing the resulting privacy-utility tradeoff and showing that meaningful formal privacy guarantees are achievable, but at a measurable accuracy cost that grows more severe as the number of participating clients shrinks [20]. That tradeoff is considerably less forgiving in a CPS federation of, say, a few dozen plants than it is in the cross-device mobile-phone settings differential-privacy-augmented federated learning was originally evaluated against, since the aggregate population differential privacy relies on to dilute individual contributions is correspondingly smaller — a gap the CPS-specific literature has, to date, engaged with less thoroughly than the general federated learning literature has.

Digital twins represent a complementary rather than competing architectural direction: rather than distributing the learned model across many physical sites, build a high-fidelity simulated replica of a single physical process and run the anomaly detector, or a broader battery of security analyses, against the comparison between the live physical system and its simulated twin. Airehenbuwa et al.'s recent survey of

security-enhancing digital twins catalogues this direction across threat modeling, incident response, and live anomaly detection use cases, and frames a digital twin's principal advantage for monitoring purposes as the ability to simulate attack scenarios and edge-case operating conditions exhaustively, off the live physical process, in a way that generates far more labeled training and test data than any real plant's limited attack history realistically can [21]. That advantage directly addresses the scarce-labeled-attack-data constraint this article has returned to repeatedly, though the survey is equally candid that a digital twin's own fidelity to the real process it mirrors is itself an open validation problem: a monitoring model trained or tested primarily against a simulated twin inherits whatever gap exists between that twin and physical reality, and the literature has not yet converged on how to measure that gap rigorously.

A final distributed concern worth naming explicitly is the power grid, which sits somewhat apart from the industrial-control testbed literature discussed above but shares its underlying logic. False data injection attacks against power system state estimation manipulate the sensor measurements a grid operator uses to infer the true electrical state of the network, in a manner explicitly designed to evade the classical residual-based bad-data detection built into most state estimators, which makes grid FDIA detection a close cousin of the physical-invariant and learned-anomaly detection problems this article has focused on for water and industrial process control. Irfan et al.'s survey of FDIA detection and localization catalogues the resulting detection literature — statistical residual methods, machine-learning classifiers, and deep sequence models applied to the grid's own multivariate telemetry — and its emphasis on *localization*, identifying not just that an attack occurred but which specific measurements or grid locations were falsified, is a useful reminder that detection alone is an incomplete operational goal: an operator responding to an alarm needs to know where to look, not just that something, somewhere, is wrong [15].

7. Open Challenges

Several limitations recur across the literature synthesized here clearly enough to name directly rather than leave implicit. First, and most fundamentally, the evidence base is testbed-concentrated: SWaT, WADI, and a small number of sibling ICS testbeds account for a

disproportionate share of the reported detection numbers this article has cited, including the headline results from Inoue et al., Kravchik and Shabtai, Adepu and Mathur, and Deng and Hooi [8,9,7,14]. These are genuinely valuable shared benchmarks — without them, cross-paper comparison would be essentially impossible — but a detector's F1 score on SWaT's specific thirty-six attack scenarios is not automatically evidence of how it will perform against a genuinely novel attack strategy on a different plant with a different physical layout, different sensor noise characteristics, and a different attacker. Umer et al. and Termanini et al.'s surveys both flag this concentration explicitly as a field-level limitation rather than a criticism of any individual paper [10,11].

Second, the precision-recall tradeoff this article has traced across nearly every architecture surveyed — near-perfect precision purchased at the cost of missing roughly one attack in seven or worse, as in Kravchik and Shabtai's strongest reported configuration, or the sharp SWaT-to-WADI performance drop every method in Deng and Hooi's comparison exhibits — remains substantially unresolved rather than merely an artifact of any one architecture choice [9,14]. Attacks involving gradual drift or subtle actuator manipulation, specifically, are reported as harder to catch than abrupt, constant-value anomalies across multiple independent papers using different model families, which suggests the difficulty is a property of the detection problem itself rather than a solvable engineering gap in any particular method.

Third, adversarial robustness is comparatively under-explored relative to detection accuracy. Lim et al. and Alauthman et al.'s GAN surveys both raise, without fully resolving, the dual-use concern that any learned detector — deep sequence model, graph network, or GAN-based discriminator alike — is a potential target for adversarially crafted inputs designed specifically to evade it, and that the generative techniques increasingly used to *build* detectors are the same techniques an adversary could use to *defeat* them [16,17]. Invariant-based approaches like Adepu and Mathur's are comparatively robust to this specific failure mode precisely because they do not rely on a learned decision boundary at all, which argues for hybrid architectures — combining physics-derived invariants with learned statistical detection — rather than treating the two

traditions as competitors, though we are not aware of extensive published evaluation of such hybrids specifically.

Fourth, distributed and privacy-preserving monitoring architectures are earlier-stage for CPS specifically than the general federated learning literature might suggest. Mothukuri et al.'s framework and the broader federated learning machinery it draws on [19] are promising fits for multi-site industrial monitoring, but the privacy-utility tradeoff Wei et al. characterize becomes more severe exactly as the federation shrinks toward the handful-of-plants scale realistic for a single industrial operator, and we are not aware of CPS-specific empirical work that has characterized this tradeoff directly against real or realistic multi-plant industrial telemetry rather than the larger simulated client populations typical of federated learning benchmarks [20]. Digital twins offer one path around the scarce-labeled-data problem specifically, but Airehenbuwa et al.'s own survey is candid that twin-fidelity validation is itself an open methodological question rather than a solved prerequisite [21].

8. Conclusion

The literature surveyed here supports a specific and, we think, defensible claim: anomaly-aware intelligent monitoring — learning a statistical or graph-structured model of normal multivariate physical-process dynamics and scoring deviations from it, whether via recurrent autoencoders, convolutional ensembles, graph deviation networks, or adversarially trained generators — has become the dominant architectural pattern for cyber-physical system security research over the past decade, converging across independent research groups working on water treatment, power grids, and general industrial control [9,8,14,13,15]. It directly addresses the central lesson of the Stuxnet case: that a capable CPS attacker will try to falsify what the monitoring layer sees, which physics-grounded, multivariate anomaly detection can catch precisely because it does not trust any single reported sensor value in isolation [1]. It is also, on the evidence assembled in this article, less mature than the volume of published architectures might suggest: reported detection numbers concentrate heavily on a small set of shared testbeds, the precision-recall tradeoff for subtle and gradual attacks remains substantially open across model families, and adversarial robustness and

distributed privacy-preserving deployment are active rather than settled research questions [10,11,16,20].

None of these gaps are reasons to dismiss the pattern; if anything, the consistency with which independently developed architectures — statistical baselines, recurrent and convolutional networks, graph deviation networks, and physical-invariant checking — have converged on the same underlying detection logic across water treatment, power grid, and general industrial control domains is itself evidence that the pattern reflects something real about the structure of the problem, not a fashion effect in the research literature. Making anomaly-aware monitoring trustworthy for production deployment on critical infrastructure, rather than merely well-evidenced on shared testbeds, will depend on the field doing the less glamorous work this article's open-challenges section names directly: broader, harder, and more heterogeneous benchmarks; honest reporting of missed and gradual attacks alongside headline F1 scores; adversarial evaluation as a standard rather than optional component of any new architecture's evaluation; and empirical, not just architectural, treatment of what distributed and privacy-preserving monitoring actually costs in detection accuracy at realistic industrial scale.

9. References

- Langner, R. (2011) Stuxnet: Dissecting a Cyberwarfare Weapon. *IEEE Security & Privacy*, 9(3): 49-51. DOI: [10.1109/MSP.2011.67](https://doi.org/10.1109/MSP.2011.67)
- Mitchell, R., Chen, I.R. (2014) A Survey of Intrusion Detection Techniques for Cyber-Physical Systems. *ACM Computing Surveys*, 46(4): 1-29. DOI: [10.1145/2542049](https://doi.org/10.1145/2542049)
- Humayed, A., Lin, J., Li, F., Luo, B. (2017) Cyber-Physical Systems Security—A Survey. *IEEE Internet of Things Journal*, 4(6): 1802-1831. DOI: [10.1109/JIOT.2017.2703172](https://doi.org/10.1109/JIOT.2017.2703172)
- Luo, Y., Xiao, Y., Cheng, L., Peng, G., Yao, D. (2021) Deep Learning-Based Anomaly Detection in Cyber-Physical Systems: Progress and Opportunities. *ACM Computing Surveys*, 54(5): 1-36. DOI: [10.1145/3453155](https://doi.org/10.1145/3453155)
- Abshari, D., Sridhar, M. (2025) Cyber-Physical Systems Security: A Comprehensive Review of Anomaly Detection Techniques. *arXiv preprint*. DOI: [10.48550/arXiv.2502.13256](https://doi.org/10.48550/arXiv.2502.13256)
- Goh, J., Adepu, S., Junejo, K.N., Mathur, A. (2017) A Dataset to Support Research in the Design of Secure Water Treatment Systems. *Critical Information Infrastructures Security (CRITIS 2016)*, Lecture Notes in Computer Science 10242: 88-99. DOI: [10.1007/978-3-319-71368-7_8](https://doi.org/10.1007/978-3-319-71368-7_8)
- Adepu, S., Mathur, A. (2016) Distributed Detection of Single-Stage Multipoint Cyber Attacks in a Water Treatment Plant. *Proceedings of the 11th ACM Asia Conference on Computer and Communications Security (AsiaCCS 2016)*. DOI: [10.1145/2897845.2897855](https://doi.org/10.1145/2897845.2897855)
- Inoue, J., Yamagata, Y., Chen, Y., Poskitt, C.M., Sun, J. (2017) Anomaly Detection for a Water Treatment System Using Unsupervised Machine Learning. *arXiv preprint*. DOI: [10.48550/arXiv.1709.05342](https://doi.org/10.48550/arXiv.1709.05342)
- Kravchik, M., Shabtai, A. (2018) Detecting Cyberattacks in Industrial Control Systems Using Convolutional Neural Networks. *Proceedings of the 2018 Workshop on Cyber-Physical Systems Security and Privacy (CPS-SPC)*, 72-83. DOI: [10.1145/3264888.3264896](https://doi.org/10.1145/3264888.3264896)
- Umer, M.A., Junejo, K.N., Jilani, M.T., Mathur, A.P. (2022) Machine Learning for Intrusion Detection in Industrial Control Systems: Applications, Challenges, and Recommendations. *International Journal of Critical Infrastructure Protection*, 38: 100516. DOI: [10.1016/j.ijcip.2022.100516](https://doi.org/10.1016/j.ijcip.2022.100516)
- Termanini, A., Al-Abri, D., Bourdouce, H., Al Maashri, A. (2024) Using Machine Learning to Detect Network Intrusions in Industrial Control Systems: A Survey. *International Journal of Information Security*, 24(1): 20. DOI: [10.1007/s10207-024-00916-x](https://doi.org/10.1007/s10207-024-00916-x)
- Su, Y., Zhao, Y., Niu, C., Liu, R., Sun, W., Pei, D. (2019) Robust Anomaly Detection for Multivariate Time Series through Stochastic Recurrent Neural Network. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, 2828-2837. DOI: [10.1145/3292500.3330672](https://doi.org/10.1145/3292500.3330672)
- Li, D., Chen, D., Jin, B., Shi, L., Goh, J., Ng, S.K. (2019) MAD-GAN: Multivariate Anomaly Detection for Time Series Data with Generative Adversarial Networks. *Artificial Neural Networks and Machine Learning – ICANN 2019*, Lecture Notes in Computer Science 11730: 703-716. DOI: [10.1007/978-3-030-30490-4_56](https://doi.org/10.1007/978-3-030-30490-4_56)
- Deng, A., Hooi, B. (2021) Graph Neural Network-Based Anomaly Detection in Multivariate Time Series. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(5): 4027-4035. DOI: [10.1609/aaai.v35i5.16523](https://doi.org/10.1609/aaai.v35i5.16523)
- Irfan, M., Sadighian, A., Tanveer, A., Al-Naimi, S.J., Oligeri, G. (2024) A Survey on Detection and Localisation of False Data Injection Attacks in Smart Grids. *IET Cyber-Physical Systems: Theory & Applications*, 9(4): 313-333. DOI: [10.1049/cps2.12093](https://doi.org/10.1049/cps2.12093)
- Lim, W., Yong, K.S.C., Lau, B.T., Tan, C.C.L. (2024) Future of Generative Adversarial Networks (GAN) for Anomaly

Detection in Network Security: A Review. *Computers & Security*, 139: 103733. DOI: [10.1016/j.cose.2024.103733](https://doi.org/10.1016/j.cose.2024.103733)

17. Alauthman, M., Aslam, N., Al-Qerem, A., Aldweesh, A., Sureephong, P. (2026) Generative Adversarial Networks for Intrusion Detection Systems: A Comprehensive Survey of Applications, Challenges, and Research Directions. *Arabian Journal for Science and Engineering*, 51(1): 179-203. DOI: [10.1007/s13369-026-11103-6](https://doi.org/10.1007/s13369-026-11103-6)
18. Mothukuri, V., Khare, P., Parizi, R.M., Pouriyeh, S., Dehghantaha, A., Srivastava, G. (2022) Federated-Learning-Based Anomaly Detection for IoT Security Attacks. *IEEE Internet of Things Journal*, 9(4): 2545-2554. DOI: [10.1109/JIOT.2021.3077803](https://doi.org/10.1109/JIOT.2021.3077803)
19. McMahan, H.B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A. (2017) Communication-Efficient Learning of Deep Networks from Decentralized Data. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, PMLR 54: 1273-1282. DOI: [10.48550/arXiv.1602.05629](https://doi.org/10.48550/arXiv.1602.05629)
20. Wei, K., Li, J., Ding, M., Ma, C., Yang, H.H., Farokhi, F., Jin, S., Quek, T.Q.S., Poor, H.V. (2020) Federated Learning With Differential Privacy: Algorithms and Performance Analysis. *IEEE Transactions on Information Forensics and Security*, 15: 3454-3469. DOI: [10.1109/TIFS.2020.2988575](https://doi.org/10.1109/TIFS.2020.2988575)
21. Airehenbuwa, B., Hasan, T., Sarkar, S., Guin, U. (2025) Advancing Security with Digital Twins: A Comprehensive Survey. *arXiv preprint*. DOI: [10.48550/arXiv.2505.17310](https://doi.org/10.48550/arXiv.2505.17310)
22. Buczak, A.L., Guven, E. (2016) A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. *IEEE Communications Surveys & Tutorials*, 18(2): 1153-1176. DOI: [10.1109/COMST.2015.2494502](https://doi.org/10.1109/COMST.2015.2494502)
23. Tavallaee, M., Bagheri, E., Lu, W., Ghorbani, A.A. (2009) A Detailed Analysis of the KDD CUP 99 Data Set. *Proceedings of the 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA)*, 1-6. DOI: [10.1109/CISDA.2009.5356528](https://doi.org/10.1109/CISDA.2009.5356528)

10. Declarations

Funding

This literature synthesis received no specific grant or financial support from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of Interest

The authors declare that they have no known competing financial or personal interests relevant to this work.

Data Availability Statement

This article is a literature synthesis and did not generate new primary data; all datasets and results discussed are drawn from the cited publications, and readers should consult those original sources for their respective data and code availability statements.

Ethical Approval

Not applicable. This work involved no human participants, animal subjects, or collection of primary experimental data, and relies exclusively on previously published, publicly available research.