GR JOURNALS | Journal of Information and Communications Technology: Algorithms, Systems and Applications

Volume 2 Issue 3 (September 2026)

DOI: https://doi.org/10.64189/ict.26318

*J. Inf. Commun. Technol. Algorithms, Syst. Appl.*, 2026, 2(3), 26318

Review Article

OPEN ACCESS

# AI Approaches for Industrial Control System Cybersecurity: A Comprehensive Review of Methodological Contexts

Olujoke Mubo Oni,[1,*] John Edet Efiong,[1] Abiodun Akinwale,[1] Adekemi Olawumi Amoo,[2] Ayoola Afolabi,[3] Oluseyi Ebunoluwa Balogun[1] and Emmanuel Ajayi Olajubu[1]

[1] *Department of Computer Science and Cybersecurity, Faculty of Computing Science, Obafemi Awolowo University, Ile – Ife, 220001, Osun State, Nigeria*

[2] *Department of Software Engineering, Faculty of Computing Science, Obafemi Awolowo University, Ile – Ife, 220001, Osun State, Nigeria*

[3] *Trustfx Framework, 25 Morningside Close, Derby DE24 9JQ, United Kingdom*

*Corresponding Author

Olujoke Mubo Oni

jokeoni@oauife.edu.ng

Received: 10 August 2026

Revised: 23 September 2026

Accepted: 24 September 2026

Published Online: 23 September

**Citation**

O. M Oni, J. E. Efiong, A. Akinwale, O. A. Amoo, A. Afolabi, E. A. Olajubu, AI approaches for industrial control system cybersecurity: A comprehensive review of methodological contexts, *Journal of Information and Communications Technology: Algorithms, Systems and Applications,* 2026, 2(3), 26318, https://doi.org/10.64189/ict.26318.



## Abstract

The cyber infrastructure of Industrial Control Systems (ICSs) that monitor power grids, water treatment facilities, pipelines, and manufacturing lines has become a very com-plex cyber-physical systems, which now face a growing range of cyber threats, such as Stuxnet, Industroyer, TRITON, and Colonial Pipeline incident. In the era of new and sophisticated attacks, such as zero-day exploits, multi-stage intrusions and adversaries taking advantage of the unique 'availability first' constraints of operational technology (OT), traditional signatures and rule-based defenses are proving less effective. This review is systematic, analyzing the use of artificial intelligence (AI) methodologies in ICS cybersecurity from the year 2018 to 2024. According to the guidelines of the PRISMA 2020, from the five scholarly databases, 1578 records were retrieved and 147 studies were included in the qualitative synthesis, and 102 studies were included in the quantitative comparison. The provided literature is categorized in a methodological taxonomy from classical machine learning to deep learning; from graph neural networks (GNNs) to reinforcement learning (RL); from autonomous response to intrusion detection, at-tack-graph analysis, malware analysis and vulnerability prioritization. Overall, the synthesis suggests that graph-based and hybrid methods yield the best detection accuracy (reported from 92% to 99.5% with less than 3% false-positive rates), whereas classical methods still seem to be the most suitable for resource-constrained applications. There are ongoing deficiencies in standardized benchmarking, adversarial robustness, safety–security co-engineering, and real-world validation. A research roadmap is advocated for guiding future research, which focuses on safe RL, federated learning, explainable GNNs, and digital-twin-based security assessment.

***Keywords***: Industrial control systems; Cybersecurity, Machine learning; Deep learning, Graph neural networks; Reinforcement learning; Intrusions detection; Risk assessment; Critical infrastructure.

## 1. Introduction

Industrial Control Systems (ICSs) are the operating backbones of critical infrastructure.[1] They operate physical processes in electric power generation and distribution, water and wastewater treatment, oil and gas pipelines, and manufacturing. Proprietary hardware, vendor-specific protocols and physical isolation from corporate networks have contributed to security through obscurity. Until recently, these systems were "Air-Gap" many of them relying on the obsolete security. Convergence of information technology (IT) and operational technology (OT) is coupling controllers previously isolated from the enterprise and from the Internet.[2-4] There are drivers for ICS/OT convergence including Industry 4.0 initiatives, the Industrial Internet of Things (IIoT), cloud-hosted analytics and remote operations.

Most ICS deployments are still described using the Purdue Enterprise Reference Architecture.[5] This divides environments into levels. Level 0 contains physical sensors and actuators and Level 1 contains programmable logic controllers (PLCs) and remote terminal units (RTUs). Level 2 contains supervisory control and data acquisition (SCADA) servers and human–machine interfaces (HMIs). Level 3 covers operations management and Levels 4 and 5 cover enterprise business systems. The layers communicate with industrial protocols such as Modbus/TCP, DNP3, PROFINET, OPC-UA and BACnet. Many were designed decades ago without authentication or encryption.[1,3] Long equipment lifecycles, insecure-by-design protocols and new connectivity contribute to an attack surface qualitatively different to enterprise IT.[6,7]

### 1.1 The ICS threat landscape

The consequences of exposure are well-documented. Stuxnet (2010) demonstrated malware's ability to cross air-gaps and physically sabotage centrifuges by manipulating PLC logic.[8,9] Attacks on the Ukrainian power grid in 2015 and 2016 (the BlackEnergy and Industroyer toolkits) left hundreds of thousands of customers without electricity.[10] Both used protocol-aware malware which was targeted towards grid operations. TRITON/TRISIS (2017) breached a further layer by taking a safety instrumented system directly under threat.[11] While the 2021 Colonial Pipeline ransomware incident mainly targeted IT systems, it triggered a precautionary OT shutdown which disrupted fuel supply across the eastern United States.[12] More recently adversaries continue to target OT.[13,14] This includes the 2021 Oldsmar water treatment intrusion attempt, the 2023 exploitation of Danish energy-sector edge devices and the 2024 FrostyGoop malware which exploited Modbus to disrupt district heating. Fig. 1 details this incident timeline.

ICS environments also have unique vulnerabilities that have no direct analogy in enterprise IT, including 20-to-30-year device lifecycles with unpatchable firmware, authentication-free protocols that enable command injection and replay attacks, and availability

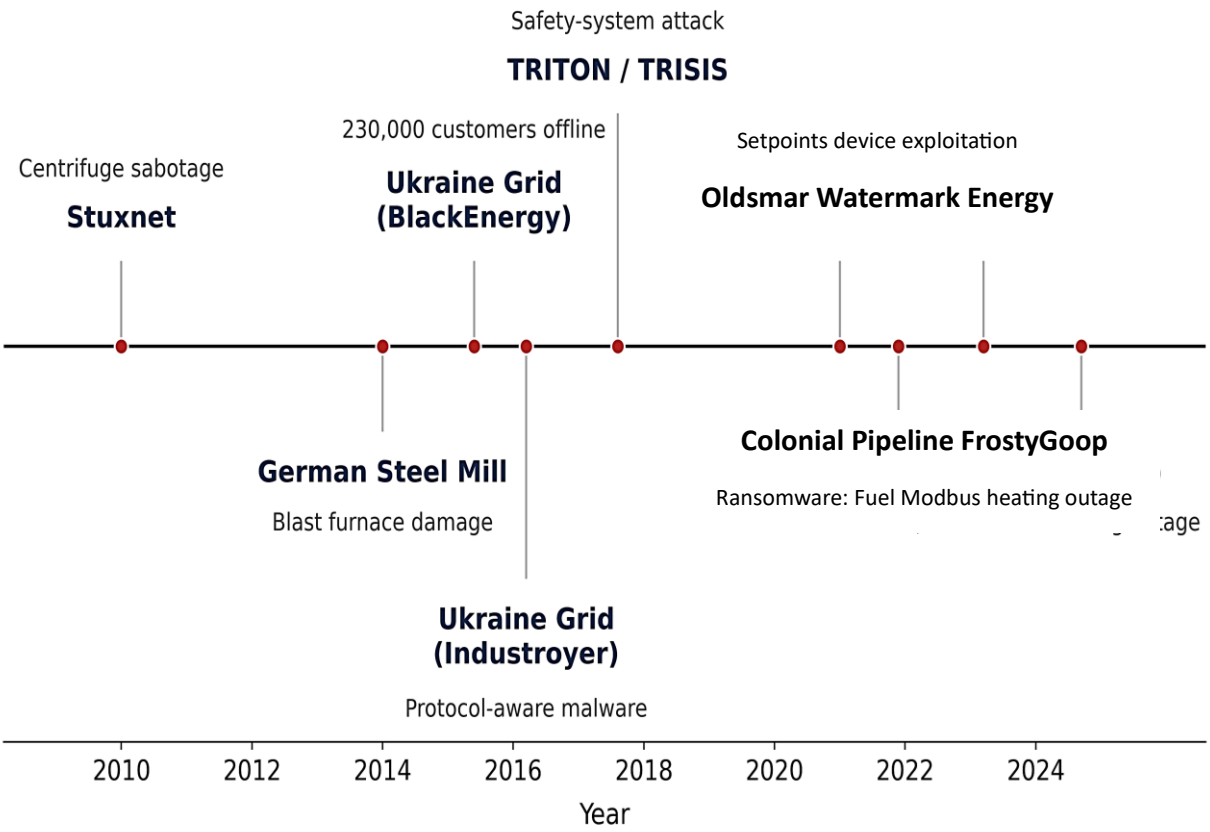


**Fig. 1:** Timeline of landmark ICS cyber incidents, 2010–2024.

requirements that exceed 99.9% (inverting the traditional confidentiality–integrity–availability priority to availability–integrity–confidentiality), with patch windows constrained to rare planned outages.[1,2,15]

### 1.2 Limitations of traditional defensive approaches

Conventional ICS defenses, including perimeter firewalls, signature-based intrusion detection, and static periodic risk assessments, are ineffective against the modern threat model. Signature-based detection is, by construction, incapable of detecting zero-day exploits or novel attack tooling; static assessments are typically conducted annually or following major changes and cannot keep pace with a threat landscape that evolves daily.[16] Rule-based systems are brittle against advanced persistent threats (APTs) that masquerade as legitimate operational traffic, and human-mediated response cycles of minutes to hours are incompatible with automated attacks that operate at timescales of seconds.[3,17] Most critically, traditional tools are topology-agnostic: events are analyzed in isolation, precluding the lateral-movement and cascading-failure signatures of ICS kill chains.[18]

### 1.3 AI as an enabler for advanced ICS security

Artificial intelligence can provide capabilities that directly address the shortcomings of traditional ICS defenses[17,19-21] including: pattern recognition in high-dimensional sensor and network data; anomaly detection in the absence of prior attack signatures; inference speeds compatible with control-loop time scales; topology-aware reasoning over inter-device dependencies via graph representation learning; and adaptive response policies learned via interaction. The research literature reflects this realization, with a blooming set of works that apply classical machine learning (ML), deep learning (DL), graph neural networks (GNNs), and reinforcement learning (RL) to ICS security tasks. This literature is, however, fragmented across venues, sectors, and evaluation practices.

### 1.4 Motivation: Why another review, and why now

Several survey articles have addressed AI or ML for ICS and SCADA security. The earliest focused on intrusion detection for cyber-physical systems but predated the current deep-learning and graph-learning waves by nearly a decade.[22,23] More recent surveys offer greater scope but are marred by three limitations that motivate the current work. First, organization: most reviews are structured by attack type or application sector, entangling algorithmically related works and obscuring methodological trade-offs that inform technology selection.[24,25] Second, coverage of contemporary methods: graph neural networks and reinforcement learning enable topology-aware detection and autonomous response but are mentioned only in passing or excluded entirely, due to the surveys' early dates.[23,26] Third, the research–practice gap: prior reviews catalog reported accuracies but say little about deployment readiness, interpretability, or suitability for heterogeneous brownfield ICS estates.

Three developments render a methodology-centric synthesis timely: (1) the 2022–2024 incident wave (including PIPEDREAM, Industroyer2, and FrostyGoop) has elevated the field from proof-of-concept to operational urgency[13,14]; (2) graph-based and RL methods have transitioned from theory to ICS-specific empirical studies, yielding sufficient quantity and variety of results to support systematic comparison[17,18,27,28]; and (3) standardization and regulatory activity has elevated deployment readiness and certifiability to first-order research questions.[1,29,30] Relevant activities include updated NIST SP 800-82 Rev. 3 guidance and the IEC 62443 series. Table 1 situates this review relative to representative prior surveys.

### 1.5 Objectives, scope, and contributions

The objective of this review is to survey and synthesize AI methods applied to ICS cybersecurity. Other objectives are to group methods by their AI type and application domain, compare reported performance and deployment readiness, interpretability, and applicability to specific scenarios, and identify research gaps and future directions. The scope spans time from January 2018 to March 2025, with notable exceptions for foundational earlier works. It is limited to AI and ML approaches to ICS, SCADA, and critical-infrastructure security and does not encompass enterprise IT security or non-AI approaches.

In relation to the limitations identified in Section 1.4, this review makes four contributions:

- A methodology-centric taxonomy that organizes the field by algorithm family rather than attack or sector. Taxonomies of AI techniques are not new, but applying this organizing principle to the ICS security literature enables insights into cross-cutting trade-offs that would be obscured by technique- or sector-indexed reviews.
- A critical comparison across techniques based on the 102 quantitatively assessable studies, reporting latency, data demands, interpretability, and adversarial exposure alongside accuracy, and stating conclusions qualitatively in light of the heterogeneity documented in Section 2.5.
- A deployment scenario suitability analysis that maps each family to specific ICS contexts, including greenfield and brownfield estates, edge and centralized deployment, and safety-critical and non-critical loops.
- A gap-driven research roadmap that links each gap to a specific future direction and, where applicable, to emerging standardization and certification requirements.

### 1.6 Paper organization

Section 2 describes the methodology addressing metric

**Table 1:** Prior surveys of AI for ICS and SCADA security representation

| Prior survey (focus) | Organizing principle | GNN | RL | Deployment readiness | Scenario suitability | 2022–2024 incidents |
|---|---|---|---|---|---|---|
| Mitchell & Chen [22] (IDS for CPS) | Detection technique | — | — | Partial | — | — |
| Umer *et al.* [24] (ML for ICS IDS) | Learning paradigm | Partial | — | Partial | — | — |
| Kaouk *et al.* [23] (IDS survey) | Attack type | — | Partial | Partial | — | — |
| Asghar *et al.* [25] (ICS security) | Sector and issue | — | — | Partial | Partial | — |
| This review | AI methodological family | Full | Full | Full | Full | Full |

*"Full" denotes sustained treatment with a dedicated section or table; "Partial" denotes brief or incidental coverage; "—" denotes no coverage. Assignments reflect the authors' reading of each survey and are intended as an orientation aid rather than a formal bibliometric measurement.*

heterogeneity. Section 3 establishes ICS cybersecurity fundamentals and constraints on AI deployment in OT. Section 4 presents the methodological taxonomy, Section 5 reviews application across security functions and sectors, and Section 6 deals with the comparison, deployment suitability, a dataset-bias assessment, and research gaps. Section 7 outlines future research directions, Section 8 summarizes practical implementations, Section 9 addresses cross-cutting challenges, and Section 10 concludes.

## 2. Materials and methods

### 2.1 Search strategy

The review followed the PRISMA 2020 guidelines for systematic reviews.[30] Five (5) scholarly databases were searched. These includes IEEE Xplore, ACM Digital Library, ScienceDirect, SpringerLink, and Web of Science. To further strengthen the research, Google Scholar was added as supplementary database. The primary search string combined ICS-domain terms with AI-technique and security terms. Table 2 summarizes the strategy. The search covered January 2018 to March 2025 and was last executed in March 2025.

### 2.2 Inclusion and exclusion criteria

Criteria were defined before screening and applied by two reviewers independently. The disagreements between inclusion and exclusion criteria were resolved as discussion in Table 3.

### 2.3 Selection process

The initial search generated 1578 records that comprises 1482 bases from different databases and 96 from supplementary sources. 424 duplicates was removed leaving 1154 records for title and abstract screening. Screening excluded 818 records. The reasons were: not within the ICS or OT domain (402), no AI or ML component (291), and not peer-reviewed (125). Full-text evaluation was then carried out on 336 articles. A further exclusion of 189 were carried out for the following reasons: not ICS-specific (104), non-AI method (58), and no empirical validation (27). This left 147 studies for qualitative synthesis out of which, 102 reported sufficient detail on dataset, metrics, and evaluation

**Table 2:** Search strategy and database coverage

| Component | Specification |
|---|---|
| Databases | IEEE Xplore; ACM Digital Library; ScienceDirect; SpringerLink; Web of Science; Google Scholar (supplementary) |
| Primary search string | ("Industrial Control System" OR ICS OR SCADA OR "operational technology") AND ("artificial intelligence" OR "machine learning" OR "deep learning") AND (cybersecurity OR "intrusion detection" OR "threat detection") |
| Secondary terms | graph neural network; reinforcement learning; CNN; LSTM; autoencoder; GAN; federated learning; anomaly detection |
| Time period | January 2018 – March 2025 (seminal pre-2018 works retained for context) |
| Language | English |
| Search date | Last executed March 2025 |

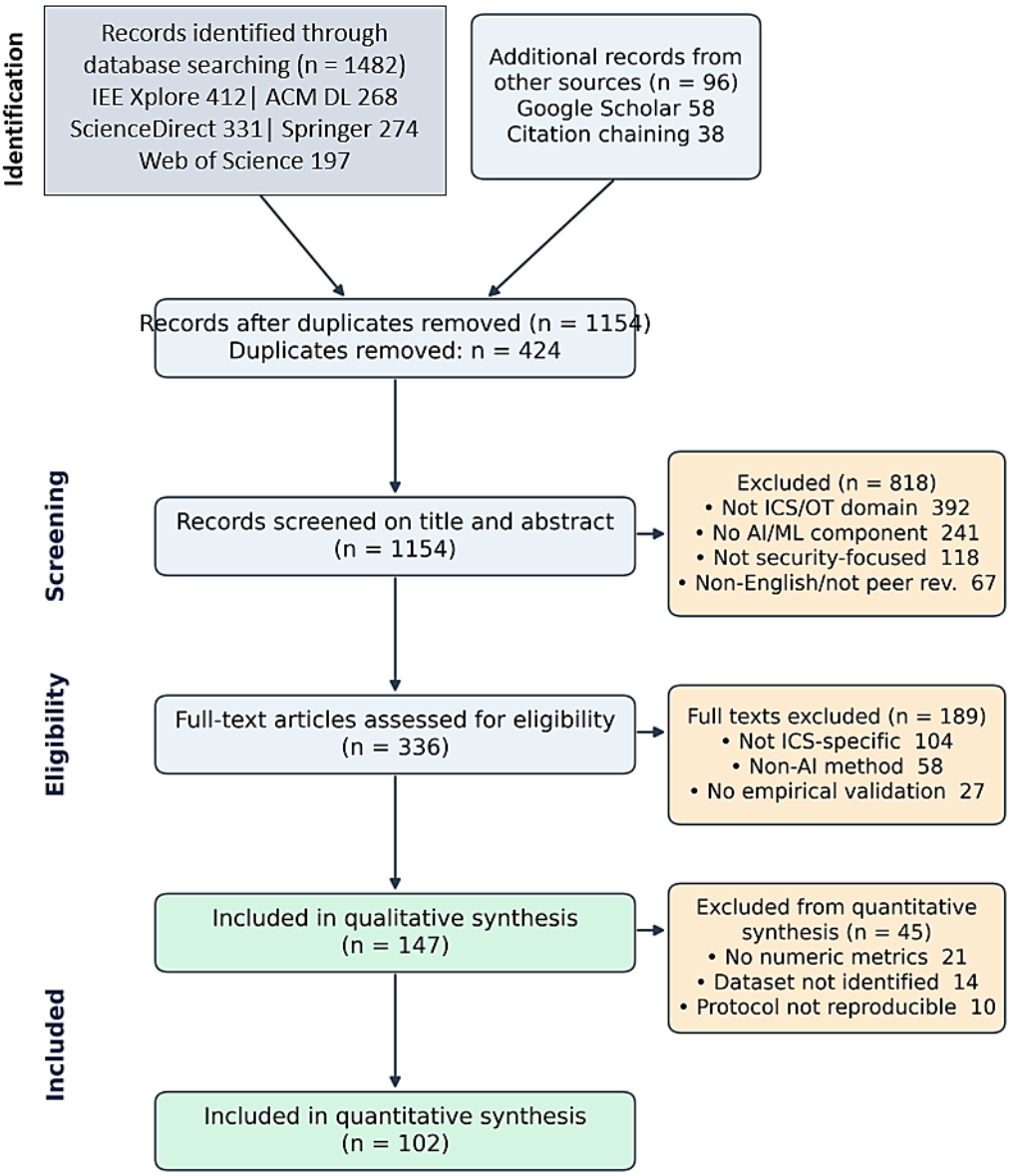


**Fig. 2:** PRISMA 2020 flow diagram.

**Table 3:** Inclusion and exclusion criteria applied during screening.

| Inclusion criteria | Exclusion criteria |
|---|---|
| Peer-reviewed journal articles and conference papers | Enterprise IT security without ICS or OT specificity |
| Explicit application of AI or ML to ICS, SCADA, or critical-infrastructure security | Non-AI approaches (purely cryptographic, formal, or procedural) |
| Empirical validation or substantive theoretical contribution | Opinion pieces and editorials without technical contribution |
| English-language publications, 2018–2025 | Duplicate or superseded publications |

protocol to support cross-study comparison. The remaining 45 were then excluded from the quantitative synthesis because of incomplete metric reporting. Fig. 2 exhibits the corresponding PRISMA flow diagram with all stages and exclusion reasons.

### 2.4 Data extraction and quality appraisal

For each included study we have extracted the following: publication metadata, AI technique category, ICS domain or sector, attack types addressed, dataset characteristics, performance metrics, and deployment considerations. Performance metrics included accuracy, precision, recall, F1-score, false-positive rate, and reported latency. Quality was assessed on five dimensions: methodological rigor, dataset realism, validation soundness, reproducibility, and practical applicability. Dataset realism is distinguished between simulation, testbed, and operational data. The full extraction table for all 147 studies is in a public repository (Section "Data Availability Statement") and is the basis for the aggregate figures given in Sections 5 and 6.

The citations in this article include the founding and methodological works that define each technique, plus representative exemplars of each major application and sector. The full list of 147 included studies is in the deposited extraction table, not the reference list, keeping the reference list proportionate but retaining full traceability of the synthesis.

### 2.5 Handling heterogeneity in quantitative synthesis

The 102 studies retained for quantitative extraction vary widely in their datasets, attack scenarios, evaluation protocols, and reported metrics. These differences limit what the synthesis can claim, so we explicitly describe the approach taken.

No meta-analytic pooling was attempted. Reported metrics were not averaged across studies, because differing class balance, attack mixes, and train–test splits make such averages uninterpretable.

Comparison was stratified by dataset. Where studies share a benchmark (principally SWaT, WADI, BATADAL, the Mississippi State corpora, and Electra), figures are compared only within that benchmark. No cross-benchmark numerical comparisons were made.

Metrics are reported as stated by the original authors. When a study reports F1-score but not accuracy, or reports a false-positive rate qualitatively, the extraction records this, rather than imputing a value. Table 6 marks unreported quantities as "not reported".

Cross-family conclusions are qualitative and ordinal. The comparative ratings in Figs. 7, 8 and Table 7 are ordinal judgements based on the quality appraisal in Section 2.4, not statistical estimates. They are presented as an orientation aid and labelled as such.

As a result, this review claim a consistent direction of reported results within shared benchmarks, and always includes the caveat that evaluation protocols differ. To establish genuine statistical superiority would require the standardized benchmarking programme identified as a research gap in Section 6.6.

## 3. ICS cybersecurity fundamentals

### 3.1 ICS architecture and components

Fig. 3 shows the Purdue reference model that forms the basis of most ICS security discussions.[5] Level 0 includes the physical process: sensors, actuators, motors, and valves. Level 1 hosts intelligent control devices that perform deterministic control logic, including PLCs, RTUs, intelligent electronic devices, and safety controllers. Level 2 provides supervisory functions via SCADA servers, HMIs, and engineering workstations. Level 3 contains operations-management systems such as manufacturing execution systems and plant historians. Levels 4 and 5 host enterprise resource planning and corporate IT. A hardened industrial demilitarized zone at Level 3.5 ideally brokers all IT–OT exchange. Each level exposes distinct telemetry that can be exploited by AI methods, including process variables, controller logic, network flows, and host logs. Each level also imposes distinct constraints on tolerable monitoring and response actions.

### 3.2 Industrial communication protocols and their vulnerabilities

Industrial protocols vary widely in security posture, as Table 4 summarizes. Modbus/TCP is the most widely deployed protocol. It transmits function codes and register values in cleartext with no authentication, which enables command injection and replay.[1,3] DNP3 is prevalent in electric and water utilities. It defines an optional Secure Authentication extension, but this is

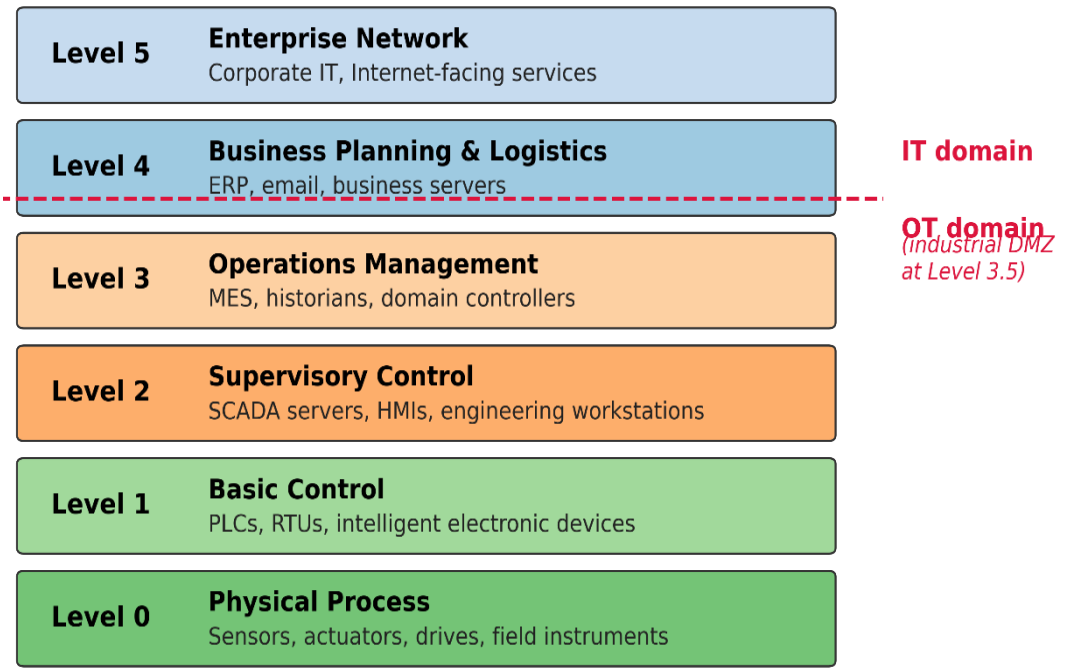


**Fig. 3:** Purdue reference architecture for ICS environments.

**Table 4:** Security characteristics of common industrial protocols

| Protocol | Typical domain | Native security | Representative attack vectors |
|---|---|---|---|
| DNP3 | Electric and water utilities (SCADA) | Optional Secure Authentication, rarely enabled | Spoofed control commands; event-buffer manipulation |
| PROFINET | Discrete manufacturing, motion control | Minimal; real-time class prioritized | Device-name takeover; DCP spoofing; denial of service |
| OPC-UA | Cross-vendor integration, IIoT gateways | Certificates, signing, encryption (configurable) | Misconfiguration abuse; downgrade to None security mode |
| BACnet | Building automation | Weak; BACnet/SC emerging | Broadcast storms; unauthorized property writes |

rarely deployed in practice. PROFINET prioritizes hard real-time delivery over confidentiality. OPC-UA is the principal modern exception, offering certificate-based authentication and encryption, although deployments frequently disable these features for compatibility. These weaknesses define much of the feature space that AI-based detectors monitor. That feature space includes function-code distributions, register-access patterns, inter-arrival timing, and protocol state transitions.

### 3.3 Attack taxonomy for ICS

Following the structure of documented ICS kill chains[10,12,31] and established cyber-physical attack taxonomies,[32] attacks are organized into five stages. Table 5 summarizes them: reconnaissance, initial access, lateral movement, command and control, and impact. The impact stage differentiates ICS from IT security most sharply. It encompasses process manipulation through setpoint alteration or command hijacking, data-integrity attacks such as false data injection against state estimation,[33] availability attacks including ransomware and denial of service, and physical damage through safety-system bypass.[11]

### 3.4 ICS-specific security requirements

Any AI-based defense for ICS must respect four domain constraints. First, availability dominates: protection mechanisms must not interrupt processes with uptime requirements above 99.9%. Second, real-time constraints apply: control loops operate on millisecond-to-second deadlines, which bounds tolerable detection and response latency. Third, safety governs: IEC 61508 and IEC 62443 impose safety integrity levels (SILs) and certification obligations that autonomous security actions must not violate.[29,34] Fourth, lifecycle realities persist: equipment lifetimes of 20 to 30 years mean AI solutions must coexist with legacy devices that cannot host agents and tolerate only passive monitoring.[1,2] These constraints recur throughout Section 6 and underpin the suitability assessment in Section 6.3.

### 3.5 Deploying resource-intensive ai in operational technology environments

The constraints above raise an obvious objection to much of the literature reviewed here. Intrusion detection based on deep or graph neural networks is computationally demanding, whereas OT environments are characterized by constrained, long-lived hardware. It is therefore necessary to state why such methods are nonetheless considered deployable, and under what architectural assumptions.

The first point is that inference and training have very different cost profiles, and only inference need occur in the OT environment. Training is performed offline, typically on historian data or in a laboratory testbed, using conventional computing infrastructure. What must run in the plant is the trained model, whose cost can be reduced substantially through pruning, quantization, and knowledge distillation. Reported ICS deployments of one-dimensional convolutional detectors achieve sub-second inference on commodity industrial PCs.[24]

The second point is that AI-based detection is not placed on the control path. In the deployments surveyed in Section 8, detection operates passively on mirrored

**Table 5:** Attack taxonomy for ICS environments with representative techniques

| Stage | Objective | Representative techniques |
|---|---|---|
| Reconnaissance | Map network and identify devices | Network scanning; device enumeration; protocol fingerprinting; exposure discovery via public scanning services |
| Initial access | Establish a foothold | Spear phishing; supply-chain compromise; exploitation of remote-access services and VPNs |
| Lateral movement | Reach OT assets from the foothold | Credential theft; exploitation of IT–OT trust relationships; engineering-workstation pivoting |
| Command and control | Maintain persistent access | Covert channels over industrial protocols; persistent backdoors; living-off-the-land techniques |
| Impact | Achieve the adversary objective | Setpoint and command manipulation; false data injection; ransomware; safety-system bypass; physical damage |

traffic from a span port or network tap, or on historian data at Level 3. Analysis therefore occurs out of band. A detector that is slow, or that fails outright, degrades monitoring but does not delay a control loop or interrupt the process. This architectural separation is what makes computationally heavier methods tolerable in an availability-first environment, and it explains why almost no reviewed study proposes inline deployment at Level 1.

The third point is that the appropriate compute location varies by level and by scenario. Level 3 and the industrial demilitarized zone can host server-class hardware, including GPU acceleration, and are the natural home for deep and graph-based analytics over plant-wide data. Levels 1 and 2, and geographically remote assets such as substations and wellheads, cannot. For these, compact classical models or quantized compressed networks are appropriate. This gradient is precisely what the scenario mapping in Section 6.3 formalizes, and it is the reason that review treats deployment context, rather than benchmark accuracy, as the primary selection criterion.

The final point is that these arguments are necessary but not sufficient. Reported latency figures in the literature are overwhelmingly obtained on testbeds rather than in production, and sustained resource consumption under continuous operation is rarely characterized at all. The justification above establishes that resource-intensive detection is architecturally plausible in OT; it does not establish that it has been demonstrated at scale. That gap is recorded in Section 6.6.

## 4. AI Techniques for ICS cybersecurity: Foundational concepts

The included literature is organized into five methodological families: classical machine learning, deep learning, graph-based deep learning, reinforcement learning, and hybrid or ensemble architectures. Fig. 4 visualizes this taxonomy. This section introduces each family at the level needed to interpret the comparison in Section 6. The emphasis is not on enumerating algorithms. It is on identifying the structural reason each family succeeds or fails against the ICS constraints set out in Section 3.4.

### 4.1 Classical machine learning

model. Temporal dependencies either modeled poorly or not captured at all, and topology is undistinguishable to the model. As attacks shift from anomalous packets to anomalous sequences and anomalous paths, the representational ceiling of this family becomes the binding limitation, not its speed.[20,21]

### 4.2 Deep Learning

Recurrent architectures process consecutive sensor and network data. In practice these are long short-term memory (LSTM) and gated recurrent unit networks.[38] They capture the control-loop dynamics and gradual process anomalous drift that point-in-time classifiers miss. ICS applications include sequence-based anomaly detection, predictive maintenance, and protocol state-machine modeling.[22,39] Convolutional neural networks extract spatial patterns from packet payloads and from image-like representations of traffic. They support traffic classification and industrial malware detection with auspicious inference swiftness.[24,40] Autoencoders learn compressed depictions of normal behavior without labels and flag anomalies via reconstruction error. This makes them workhorse of zero-day detection on testbeds such as SWaT and WADI.[37,41,42] Generative adversarial networks[43] are used mainly for synthetic attack-data augmentation, which mitigates severe class imbalance, and for adversarial robustness testing.

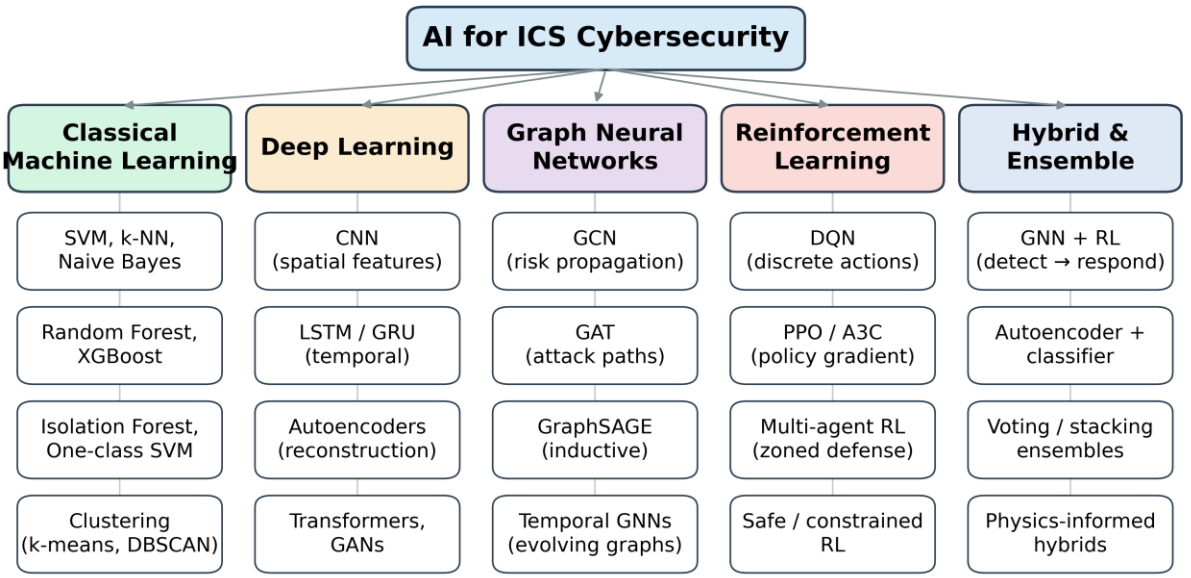


**Fig. 4:** Methodological taxonomy of AI approaches for ICS cybersecurity derived from the 147 included studies.

Deep learning posses a pivotal advantage of removing manual feature engineering and modelling temporal structure end to end. This in turns yields the highest benchmark accuracies in the mass. However, the same lucidity that captures subtle attacks also fits dataset unconventionalities. With the small, testbed-derived datasets typical of ICS, overfitting and brittle generalization are common detriments. The resulting models are also impervious to the operators who must act on their alarms. Deep learning therefore trades away the interpretability and data frugality that ICS centrals, in exchange for accuracy the domain cannot always exploit. Section 6 indicates that this exchange is favorable only where sufficient representative data and post-hoc explanation tooling are obtainable.[20,21,44]

### 4.3 Graph-based deep learning

GNNs represent the entire ICS components as a graph. Nodes are devices such as PLCs, HMIs, SCADA servers, and sensors, attributed with traffic statistics, device properties, and vulnerability data. Edges encode communication relationships and functional dependencies. Message-passing layers iteratively aggregate information from each node's neighborhood, so learned representations reflect both local behavior and network position.[19,45] Fig. 5 illustrates the pipeline. Four variations appear in the ICS literature: graph convolutional networks for network-wide threat detection,[45] graph attention networks whose attention weights highlight critical attack paths,[46] GraphSAGE for inductive learning that scales to large industrial networks,[47] and temporal GNNs that track attack propagation over evolving graphs.[18,27,48]

The structural argument for GNNs in ICS is that the threat model is a relational issue. These means that lateral movement, cascading failure, and coordinated multi-node attacks are properties of paths and subgraphs rather than of individual events which can only be represented in a topology-aware context. Thus, GNNs surface multi-hop attack paths that flat detectors cannot, and their attention weights offer a partial answer to the interpretability problem affecting other deep models.[18] The counterbalancing outlays are real and underappreciated, and an accurate, current graph must be constructed and maintained. Message passing also scales unfavorably on very large networks. The promise of this family is therefore concentrated in environments where topology is known and relational attacks dominate.

### 4.4 Reinforcement Learning for Autonomous Response

RL articulates defense as a sequential decision problem.[49] It is demonstrated as a Markov decision process whose state captures network configuration and risk levels. This allows mitigation actions such as isolate, throttle, reroute, alert, and allow. The reward exhibit balances in risk reduction against operational disruption. Deep Q-networks handle discrete mitigation spaces.[50] Policy-gradient methods such as proximal policy optimization and A3C support continuous action spaces with improved training constancy and stability.[28] Multi-agent RL distributes defense across network zones with synchronized containment.[17] Fig. 6 depicts the closed defense loop, including the safety and availability constraints that distinguish ICS deployments.

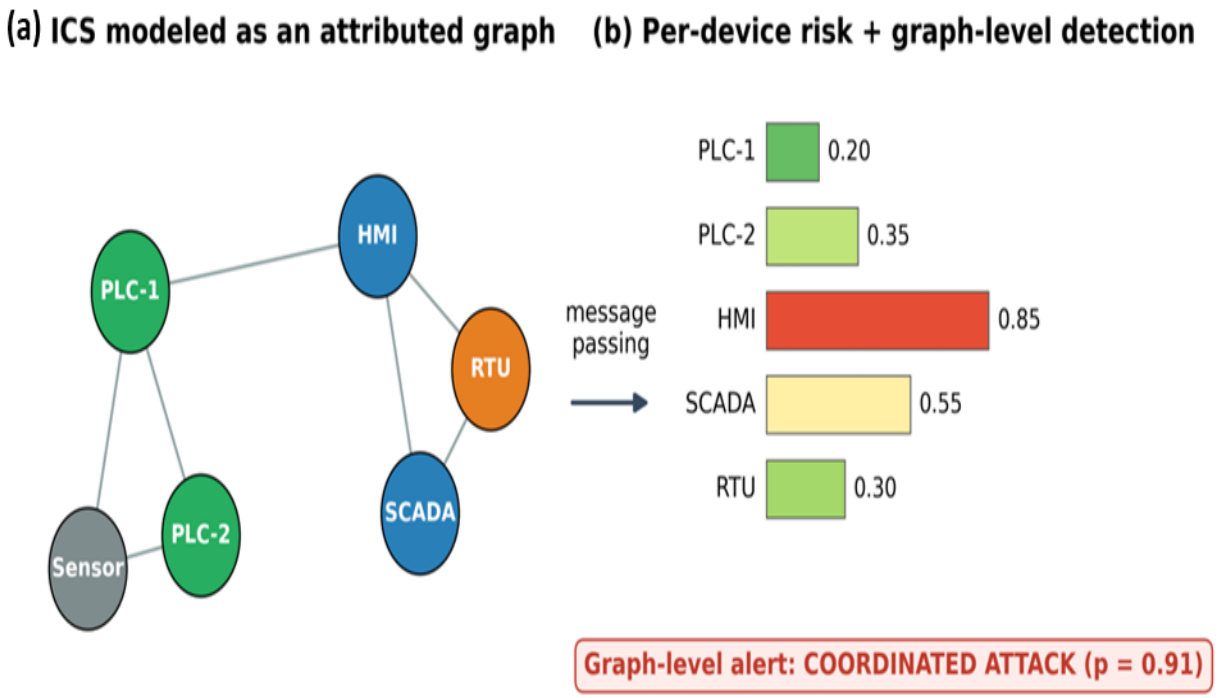


**Fig. 5:** Conceptual GNN pipeline for ICS security. (a) The industrial network is modeled as an attributed graph. (b) Message-passing layers produce per-device risk scores and a graph-level attack detection.

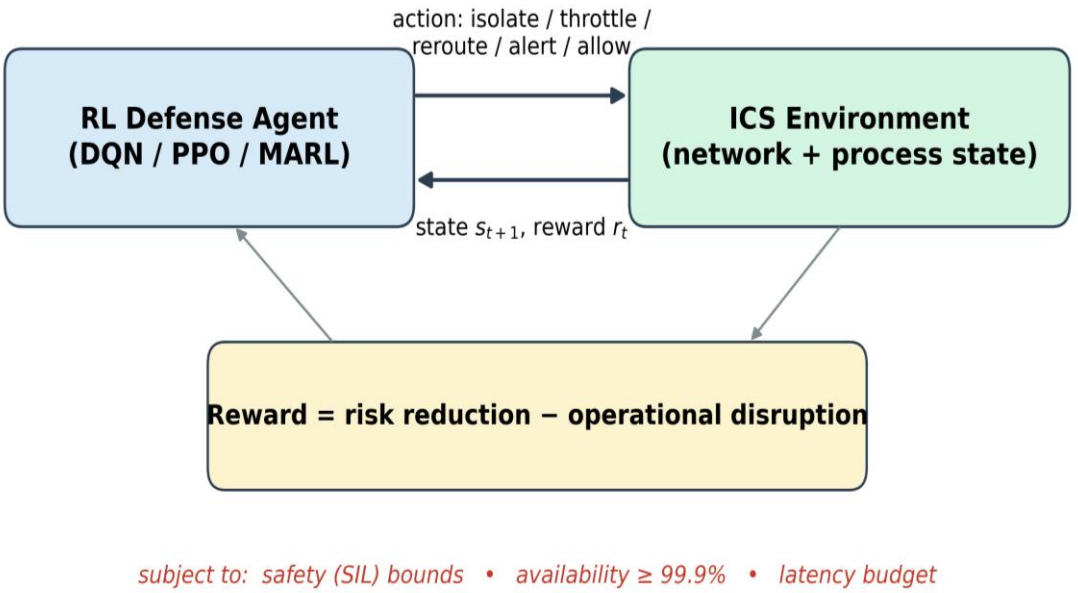


**Fig. 6:** Reinforcement learning framework for adaptive ICS defense.

The RL agent is both its appeal and predicament as it selects mitigation actions, observes the resulting network state, and receives a reward that trades off risk reduction against operational disruption, subject to safety (SIL) and availability constraints. The appeal is autonomous, adaptive mitigation at machine speed, with explicit optimization of the security–availability trade-off that operators currently negotiate manually. The predicament is that unsafe exploration is central to how RL learns, and ICS cannot tolerate it. The resulting policies are also opaque and difficult to certify against SIL requirements. Every reviewed RL study consequently remains simulation-bound or confined to advisory operation. The practical contribution of this family today is decision support. Its path to actuation runs through the safe-RL and verification research examined in Section 7.2.[17,49]

### 4.5 Hybrid and ensemble approaches

A growing share of recent work composes techniques. Examples include GNN-based detection feeding RL-based response, autoencoder front-ends paired with supervised classifiers, voting ensembles that trade marginal latency for accuracy and false-positive reduction, and physics-informed models that constrain learned detectors with first-principles process models.[16,18,51,52] Hybrids report the strongest aggregate performance among the included studies, as Section 6 discusses. The reason is that they compose corresponding inductive biases, namely temporal, spatial, relational, and physical, rather than relying on any single one. The recurring caveat is that architectural complexity raises engineering, maintenance, and interpretability costs. These gains must therefore be weighed against a sustainment burden that, as Section 6.2 shows, the literature rarely reports.

## 5. Systematic review of AI applications in ICS security

This section fuses the empirical literature by security function and by sector. The aim is to extract the consistent signal across different studies, namely what works, under what evaluation conditions, and with what reported reliability. Shared methodological weaknesses that make aggregate claims provisional.

### 5.1. Intrusion detection systems

#### 5.1.1 Supervised learning-based intrusion detection

Supervised intrusion detection studies report accuracies typically between 85% and 98%. False-positive rates range from 2% to 15%, and detection latencies from seconds to minutes. Random forests applied to Modbus and power-system telemetry report approximately 95% accuracy on the Morris–Gao gas-pipeline and power-system datasets.[26,53] Support vector machine classifiers report about 92% precision for DNP3 attack categorization.[26,36] Feed-forward neural networks support multi-protocol detection on laboratory testbeds.[20,35] The dominant datasets are the Mississippi State ICS attack corpora.[26,53] the SWaT and WADI water-treatment testbeds,[41,42] and bespoke testbed captures.[54] Three limitations that recur includes; dependence on labeled attack data, static models that require periodic retraining, and weak cross-sector generalization. These momentous accuracies are therefore best read as in-distribution upper bounds rather than field expectations.[20,21]

#### 5.1.2 Unsupervised and deep anomaly detection

Unsupervised approaches achieve detection rates of 88% to 95% without labeled attacks, at the cost of high false positives.[37,55] These approaches include clustering for baseline establishment and autoencoders scoring reconstruction error. They face the central difficulty of

defining normal behavior in processes with legitimate mode changes. Deep sequence models substantially improve temporal coverage. LSTM-based detectors on SWaT capture multi-stage attacks that unfold over minutes to hours.[22,39] One-dimensional convolutional networks proves comparable accuracy with lower inference cost, which favors real-time deployment.[24,40] The persistent tautness is between sensitivity to genuine attacks and tolerance of benign operational disparity. The false-positive consideration of a continuously staffed control room make this trade-off decisive, yet few studies evaluate it under realistic alert-volume assumptions. Table 6 summarizes representative intrusion detection studies across families.

### 5.2 Graph-based threat analysis

Beyond per-event detection, graph methods reason over attack structure. Classical attack-graph analysis enumerates exploitation paths and identifies critical nodes through centrality measures. AI-enhanced variants learn to predict attack progression probabilistically over these graphs.[18] GNN applications report accuracy improvements of 5 to 15 percentage points over topology-agnostic baselines, lateral-movement detection rates above 90%, and the capability to showcase multi-hop attack paths that flat intrusion detection features cannot signify.[18,19,27,48] Sector case studies include GNN analysis of substation SCADA networks for coordinated-attack detection in power grids,[18] and topology-consistency checking of sensor networks for false-data-injection detection in water treatment.[27,55] Foundational work on false data injection against power-system state valuation[33] and on machine-learning detection of it[36] anchors this line of research. One stipulation is consistent across these studies: the reported gains are demonstrated on small, well-characterized topologies whether they hold on large, partially observed brownfield networks remains an open empirical question.

### 5.3 Reinforcement learning for automated response

RL for automation under studies remain predominantly simulation-bound, but they report compelling operational metrics. Q-learning agents optimize firewall rules over states that syndicate network configuration and threat intelligence, selecting among block, allow, throttle, and reroute actions with rewards that penalize operational impact.[17,57] Agents trained with proximal policy optimization learn policies that explicitly balance security and availability.[17,28,57] Cooperative multi-agent designs shape containment across Purdue zones.[17] Training requires thousands to millions of simulated episodes in gym-style environments. The reviewed studies report reductions in mean time to detect from minutes to seconds, together with substantial improvements in mean time to respond relative to manual handling and learned reduction of false positives. These figures are obtained in the same simulators in which the agents were trained. The safe-exploration problem keeps such agents out of real-time

**Table 6:** Representative AI-based intrusion detection studies for ICS

| Study | Technique | Dataset | Accuracy (%) | F1 | FPR (%) | Latency | Setting |
|---|---|---|---|---|---|---|---|
| Beaver *et al.* [26] | RF / SVM / NB | Morris–Gao SCADA | 95.0 | nr | 5.0 | seconds | Sim. |
| Inoue *et al.* [37] | DNN / one-class SVM | SWaT | nr | 0.80 | nr | <1 s | Testbed |
| *Goh et al.*[22] | LSTM (unsupervised) | SWaT | nr | nr | nr | ~1 s | Testbed |
| Kravchik & Shabtai[24] | 1D CNN | SWaT | nr | 0.87 | nr | <1 s | Testbed |
| Feng *et al.*[39] | LSTM + signatures | Gas pipeline | 92.0 | nr | 3.0 | ~1 s | Testbed |
| Zolanvari *et al.* [56] | Random Forest | Custom IIoT testbed | 99.0 | nr | 0.8 | <10 ms | Testbed |
| Deng & Hooi [27] | Graph attention (GDN) | SWaT / WADI | nr | 0.81 / 0.94 | nr | ~100 ms | Testbed |
| Presekal *et al.*[18] | Hybrid GCN–LSTM | Power-grid co-simulation | 95.0 | nr | 2.0 | sub-second | Co-sim. |
| Nedeljkovic & Jakovljevic [40] | 1D CNN | Custom ICS testbed | nr | 0.94 | nr | nr | Testbed |

*Metrics are reported exactly as stated by the original authors; "nr" denotes not reported. Because datasets, attack mixes, and evaluation protocols differ, values in different rows are not directly comparable and should be compared only within a shared dataset. Accuracy figures are dominated by class imbalance in several of these datasets, and a high accuracy with an unreported false-positive rate should not be read as evidence of operational suitability. Where a study reported only an F1-score, no accuracy value has been imputed.*

environments and restrict current deployments to advisory modes with a human in the loop. The operational metrics should therefore be read as ceilings, dependent on a simulation-to-reality fidelity gap that remains unquantified.[17]

### 5.4 Malware detection, vulnerability assessment, and protocol analysis

Static analysis with classical machine learning achieves 90% to 98% detection of known families of industrial malware on features such as opcodes, imports, and function calls. Recurrent models over execution traces and convolutional models over system-call timelines yield generalizations to behavioral variants. Detection of zero-day malware is guided by post-mortem analysis of Stuxnet and TRITON attack chains.[8,9,11] For vulnerability mangement, models trained on the corpus of CVEs identify exploits likely to be weaponized in ICS-specific contexts (protocol, device roles, process criticality). Graph-based risk-propagation frameworks estimate dynamic risk scores that react to current threat intelligence, including predictive dynamic risk management for ICS networks.[16] In protocol analysis, recurrent models discover protocol grammars for reverse-engineering private protocols or detecting state-machine violations such as command injection. RL-driven fuzzers target weak points in Modbus, DNP3, and PROFINET message parsers with higher efficiency than blind fuzzing, accelerating the discovery of exploitable edge cases.[20,21]

### 5.5 Sector-specific applications

In the energy sector, smart-grid manuscripts focus on false-data injection attacks in state estimation.[33,36] or the IEC 61850 standard for substation automation.[18,58] Water/wastewater analyses mostly rely on the SWaT, WADI, and BATADAL corpora to detect sensor spoofing or process manipulation attempts with autoencoders or physics-based models.[37,41,42,51,55] For the manufacturing sector, contributions focus on sabotage of production lines, integrity checks of quality-assurance data, or industrial robot attacks.[20,56] In oil and gas, leakage discrimination in pipeline supervision or remote monitoring with limited connectivity links ML to physical-layer security.[26,53] A common observation is the focus on a handful of public testbeds as a proxy for sector-specific ICS. To what extent performance comparisons across sector studies reflect actual generalization capabilities is discussed in Section 6.4.

## 6. Comparative analysis and research gaps

### 6.1 Performance comparison across methodological families

A direct comparison of results across the 102 studies retained for quantitative synthesis is not possible due to heterogeneity in attack profiles, testbed configurations, and evaluation protocols (Section 2.5). This heterogeneity is identified as an issue in Section 2.5, but it is also a research opportunity. The comparison below is organized around the shared benchmarks and cites the results as directional trends within each of these benchmark families.

Across multiple benchmarks (SWaT, WADI, gas pipeline from Mississippi State University, Electra), deep learning methods report the best detection performance on intrusion detection tasks, usually 95%+ (up to 99%). They achieve that at the cost of higher computational overhead: hundreds of epochs during training and hour-level latency during inference, both incompatible with real-time Level 1 and 2 response.[37,41,42,51] Classical machine learning methods report 2–5% lower accuracy on the same benchmarks. However, they are trained in minutes and can be deployed on real-time hardware with microsecond-level inference latency.[26,53,54] These methods remain the only practical option for network edge devices. GNNs show mixed performance against standard ML on event-level detection, but they are the only methods providing network-level risk assessment and attack propagation analysis.[18,19,27] Finally, RL methods do not compete with other families in detection performance. However, they dominate in the response phase: the reviewed benchmarks report 60–90% improvements in mean time to respond compared to human operators during simulated attack scenarios.[17,28] It must be emphasized that the trends described for each family are drawn from studies using similar attack benchmarks. Within each body of work, no method is universally better than others. Trends are rarely statistically significant; they reflect the general tendencies when using certain attack sets, testbeds, and evaluation schemes. For example, a 5% improvement over a 93% ACC might hide a 10x longer training time and no inherent generalization capabilities. Fig. 7 summarizes the qualitative comparison of ten methodological families in six categories.

The radar profile serves to make the central observation of this review immediately visible. The families are not competitors on a single axis but rather occupy complementary regions of the design space. Few if any family dominates on all six dimensions, and which dimensions dominate depends on the deployment scenario, which Section 6.3 addresses explicitly.

### 6.2. Deployment considerations

#### 6.2.1 Computational requirements and data demands

Classical machine learning entails low training and inference costs. It is compatible with commodity industrial PCs and with gateways adjacent to controllers. Deep learning generally requires GPU-class training infrastructure and although inference can be optimized for edge deployment through pruning and quantization,

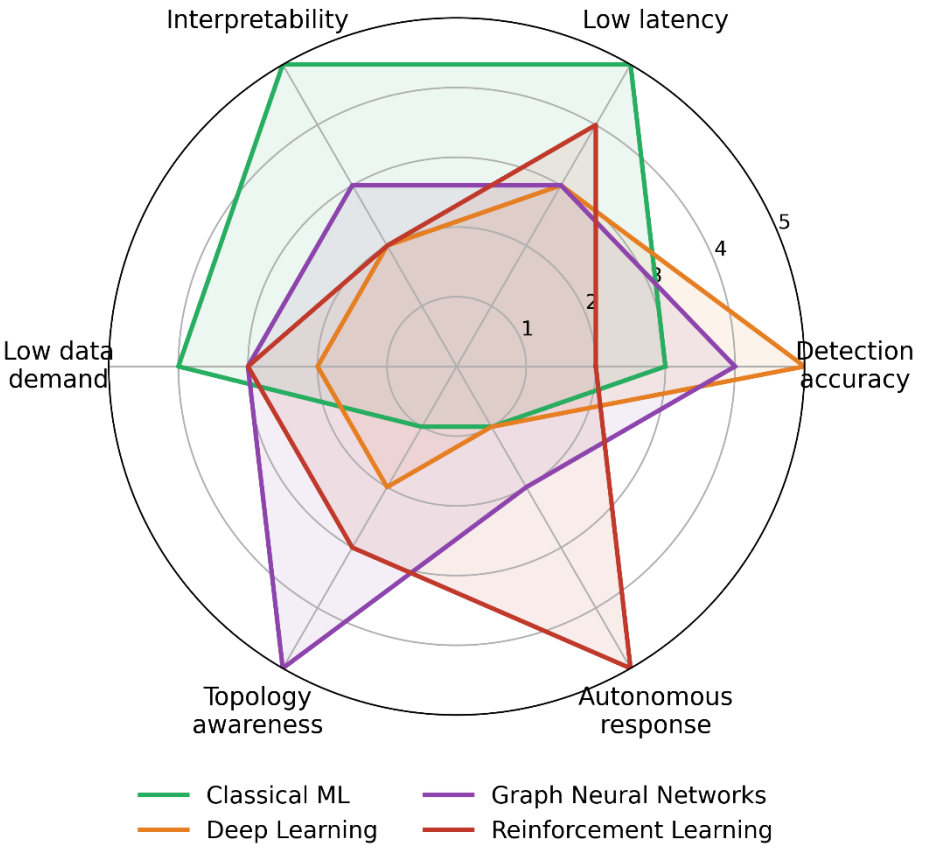


**Fig. 7:** Qualitative comparison of four methodological families across six deployment-relevant dimensions, synthesized from the 102 quantitatively assessed studies. Scores are ordinal ratings (1 = weak, 5 = strong) assigned through the quality appraisal described in Section 2.4. They summarize the direction of reported evidence and are not statistical estimates.

as discussed in Section 3.5. GNN inference is moderate but scales with graph size and requires careful engineering beyond a few thousand nodes. RL inference is lightweight, but training is sample-intensive and almost always confined to simulation. Data demands follow a complementary gradient. Supervised methods need thousands of labeled samples per attack class, which is a serious obstacle given the scarcity of real ICS attack data. Unsupervised anomaly detection requires only unlabeled records of normal operation, and semi-supervised designs occupy the middle ground.[35,37,42,55] The simulation-to-reality gap compounds the problem, because most public corpora originate from testbeds whose fidelity to production environments is difficult to verify.[35,51,55]

### 6.2.2 Operational integration and maintenance

Ease of integration declines along the spectrum from classical machine learning to RL. Classical models slot into existing security information and event management pipelines. Deep learning demands machine-learning engineering expertise. GNNs require continuously maintained network topology models. RL involves closed-loop actuation, which raises certification questions that current standards do not answer.[29,30] All families require ongoing maintenance against concept drift as processes, setpoints, and network configurations evolve. Retraining cadence, model version control, and rollback procedures are reported in fewer than 15% of reviewed studies. Operational sustainment therefore remains an afterthought in the literature and a hidden cost in any deployment decision.

### 6.2.3 Interpretability and explainability

Interpretability is not a luxury in ICS settings. Operators must understand why an alarm fired before they will act on it. Safety regulators increasingly expect traceable decision logic for systems that influence protective functions.[29,44] Decision trees and random forests offer high intrinsic interpretability. Support vector machines are partially interpretable through inspection of support vectors. Deep networks and learned RL policies are effectively opaque without post-hoc tooling. Explainable-AI techniques are appearing in the most recent studies but were applied in fewer than one in five reviewed papers.[44] These techniques include SHAP feature attribution for tabular detectors, attention visualization for GNNs, and policy summarization for RL. Fig. 8 consolidates these deployment factors into a readiness heat map relating methodological families (rows) to operational criteria (columns). Darker green indicates greater readiness. Ratings are ordinal and summarize the qualitative synthesis.

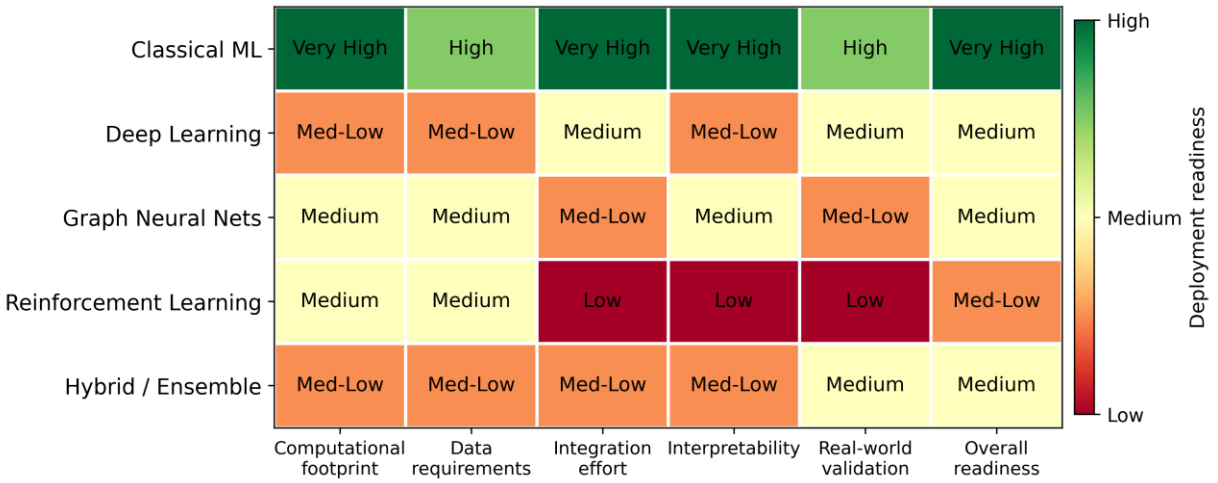


**Fig. 8:** Deployment-readiness heat map.

**Table 7:** Suitability of AI methodological families for representative ICS deployment scenarios.

| Deployment scenario | Governing constraint | Classical ML | Deep learning | GNN | RL | Recommended primary approach |
|---|---|---|---|---|---|---|
| Greenfield, well-instrumented | Data abundance; known topology | Medium | High | High | Medium | Hybrid GNN with deep anomaly detection |
| Brownfield, legacy estate | Sparse data; passive monitoring only | High | Medium | Low | Low | Interpretable ML with unsupervised autoencoder |
| Edge, resource-constrained | Compute and power limits | High | Medium | Low | Medium | Compact classical ML or quantized 1D CNN |
| Safety-critical (SIL-rated) | Certifiability; no unsafe actuation | High | Medium | Medium | Low | Interpretable detection, advisory mode only |
| Large dynamic enterprise-OT | Lateral-movement risk; scale | Low | Medium | High | Medium | Inductive GNN (GraphSAGE) with RL advisory |

*Ratings are ordinal judgements derived from the synthesis in Sections 6.1 and 6.2: High denotes a strong fit, Medium a conditional fit, and Low a poor fit.*

### 6.3 Suitability for ICS deployment scenarios

Generalizing the strength and limitation in comparisons is usually under-determined in practical instances, because ICS Components are not homogeneous. This subsection maps the methodological families onto five repeated deployment scenarios distinguishing it by domain constraints of Section 3.4. Table 7 summarizes the mapping, which is elaborated below.

Greenfield, well-instrumented facilities are the most permissive scenario. These are new builds with complete asset inventories and historians. Deep-learning detectors can be trained on ample process data, and GNNs can exploit an accurate and maintained topology. Hybrid GNN and deep-learning architectures are the strongest fit. Brownfield, legacy systems invert these assumptions due to incomplete inventories which undermine graph construction, and labeled data is scarce. Interpretable classical machine learning and unsupervised autoencoders, deployed passively, are the realistic choice. Edge and resource-constrained deployments, such as remote substations and frames, rule out GPU-class inference and favor compact classical models or quantized one-dimensional convolutional networks. Safety-critical, SIL-rated loops cannot tolerate obscured or actuating models. Interpretable detection in advisory mode is the only currently defensible option, and autonomous RL is contraindicated until safe-RL and verification methods mature, as Section 7.2 discusses. Finally, large and dynamic enterprise-OT networks, where lateral movement is the dominant concern, are where topology-aware GNNs justify their construction and maintenance overhead.

### 6.4 Dataset bias and generalization

The resulting bias deserves explicit treatment rather than a passing caveat. Among the 102 quantitatively assessed studies, a large majority evaluate on one of four public corpora: SWaT, WADI, the Mississippi State gas-pipeline

and power-system datasets, and BATADAL. The first is benchmark overfitting at the level of the field rather than the individual model. When consecutive studies tune architectures against the same fixed masses, reported developments partly reflect adaptation to that corpus, including its specific attack scripts, sensor configuration, and noise characteristics. Gains of one or two percentage points on SWaT, accumulated across many papers, cannot be assumed to transfer to an unseen plant.

The second bias deals with limited attack realism. Public ICS datasets are created by executing a predetermined and usually modest catalogue of attack scenarios against a testbed. Attacks are typically launched by researchers rather than by adversaries who adapt to the defense, and stealthy long-horizon campaigns typically seen in TRITON are usually under-represented. A detector validated on such data has been tested against a fixed and comparatively cooperative opponent.

The third isuue is scale and sector skew. SWaT and WADI are small water-treatment testbeds with tens of sensing setpoints. Production ICS environment routinely have thousands, with heterogeneous vendors and partially documented topology. Water treatment and power are heavily over-represented relative to oil and gas, chemical processing, and discrete manufacturing. Verification for sector transfer is consequently thin, which is why Section 5.5 treats apparent cross-sector consistency with caution.

The fourth is class imbalance and metric fragility. Attack samples form a small minority of most corpora. Accuracy is therefore an unreliable summary, and a detector predicts the majority class can score highly. This is the reason Table 6 records false-positive rates separately and marks unreported values rather than imputing them.

Two mitigations are visible in the more careful recent studies and are worth adopting as norms. To address this, cross-dataset evaluation in which a model trained on one corpus is tested on another without retuning is imperatively recommended. This study infirmly substantiate accuracy degradation as observed in the few studies. Secondly, the sensor configuration, sampling rate, and process modes under which a result holds should be explicitly reported within the operational envelope. Neither practice is yet common, and their absence is recorded as a methodological gap.

### 6.5 Strengths and limitations synthesis

Table 8 consolidates the comparative evidence into a strengths-and-limitations matrix as shown in scenario mapping of Table 7. No single family satisfies the joint requirements of accuracy, latency, interpretability, topology awareness, and autonomous response. Layered architectures represent the most credible deployment pattern. Such architectures use classical machine learning for fast triage, deep models for high-fidelity detection, GNNs for network-wide risk propagation, and eventually constrained RL for response orchestration. This mirrors the defense-in-depth philosophy already embedded in IEC 62443 zone-and-conduit design.[29,30]

### 6.6 Identified research gaps

The four methodological gap clusters revolve across the corpus. These methodological gaps include the absence of standardized benchmarks and metrics, which prevents meaningful cross-study comparison; over-reliance on a small number of testbed datasets, with the consequences set out in Section 6.4; and adversarial robustness, which is examined in fewer than 10% of studies despite its evident relevance to a contested environment.[34,44] Application gaps include millisecond-level latency validation in production settings, retrofitting onto legacy and resource-constrained devices, and safety–security co-engineering under IEC 61508 and IEC 62443.[29,30] Technical gaps include multi-stage kill-chain correlation, as most detectors flag isolated events; transfer learning across sectors and protocols; and human–AI collaboration, covering trust calibration, alarm rationalization, and operator interface design.[13,44] Data and privacy gaps include attack-data scarcity, barriers to cross-facility sharing, and non-independent and non-identically distributed data across sites. Table 9 maps each cluster to the directions developed in Section 7.

## 7. Future research directions

### 7.1 Advancing graph-based approaches

Three extensions would address the principal weaknesses of current GNN applications. Dynamic graph learning, using temporal graph networks and streaming graph processing with incremental updates, would allow risk models to track topology and traffic evolution in real time rather than operating on periodic snapshots.[19,48,50] Heterogeneous GNNs that model multiple node types, such as devices, users, and processes, and multiple edge types, such as protocols and trust relationships, would capture the multi-relational structure of industrial networks. Explainable GNNs are a prerequisite for adoption in safety-regulated environments. Relevant techniques include attention visualization, subgraph explanation, and operator-oriented risk reporting.[44]

### 7.2 Safe reinforcement learning for autonomous defense

Of all the directions surveyed here, autonomous RL response carries the most serious safety implications, and these warrant treatment beyond a general call for caution. The core difficulty is that an RL defense agent and a safety instrumented system can act on the same plant with different objectives. An agent rewarded for containing an intrusion may isolate a network segment, block a protocol flow, or throttle communication. In an

**Table 8:** Strengths and limitations of the four principal AI methodological families for ICS cybersecurity

| Family | Key strengths | Key limitations |
|---|---|---|
| Classical machine learning | Computational efficiency; interpretability; mature tooling; edge-deployable | Manual feature engineering; weak temporal modeling; poor generalization to novel attacks |
| Deep learning | Automatic feature learning; highest benchmark accuracy; strong temporal and spatial pattern capture | Large data and compute demands; opaque decisions; overfitting and drift sensitivity |
| Graph neural networks | Topology-aware analysis; attack-path and cascading-failure reasoning; network-wide risk scoring | Graph construction and maintenance overhead; scalability limits; sparse real-world validation |
| Reinforcement learning | Autonomous adaptive response; explicit security–availability trade-off optimization | Sample inefficiency; unsafe exploration; opaque policies; undefined certification pathway |

**Table 9:** Research gap clusters and corresponding future research directions

| Gap cluster | Specific gaps | Corresponding direction (Section 7) |
|---|---|---|
| Methodological | No common benchmarks; simulation-to-reality gap; untested adversarial robustness; no cross-dataset evaluation norm | Standardization and benchmarking (7.6); adversarial ML defenses (7.7) |
| Application | Unproven real-time latency in production; legacy retrofit; safety–security co-engineering | Digital-twin validation (7.4); certification pathways (7.6) |
| Technical | Multi-stage attack correlation; transfer learning; human–AI collaboration | Dynamic and heterogeneous GNNs (7.1); safe RL and meta-learning (7.2) |
| Data and privacy | Attack-data scarcity; cross-facility sharing barriers; non-IID distributions | Federated and privacy-preserving learning (7.3) |

ICS, each of those actions can have a physical consequence. Isolating a segment may sever the path by which a controller receives a sensor reading it depends on. Blocking traffic may prevent an operator from issuing a manual override during an upset. A mitigation that is entirely correct as a security action can therefore initiate a process excursion, and the agent's reward function will not register this unless the physical consequence was explicitly encoded.

Three specific hazards follow. The first is unsafe exploration during learning. RL discovers good policies by trying poor ones, and in a physical plant a poor action is not a low reward but a damaged asset or an injured worker. This alone rules out on-line learning in production and is why training is confined to simulators and digital twins. The second is reward misspecification. A reward that trades risk reduction against operational disruption requires disruption to be quantified, and the quantities that matter, such as proximity to a safety limit or the margin remaining in a protective function, are exactly those that are hardest to express numerically. An agent optimizing an incomplete proxy will exploit the omission. The third is interaction with the safety instrumented system. Under IEC 61508 and IEC 62443, safety functions are assigned integrity levels and are validated on the assumption that their inputs and actuation paths are available. An autonomous security agent that can alter network reachability is, in effect, an unvalidated component capable of influencing a SIL-rated function.[29,34]

Research responses fall into three groups. Constrained policy optimization treats safety as a hard constraint rather than a reward term, so that the agent optimizes within a feasible region defined by process invariants instead of trading safety against security. Formal verification of learned policies seeks to prove that a trained policy cannot reach a prohibited state, which is the form of assurance a safety case actually requires; current methods scale only to modest state spaces, and extending them is an open problem. Action shielding places a verified runtime monitor between the agent and the plant, which vetoes any proposed action that would violate a stated invariant, and offers a more immediately practical path because the shield, rather than the learned policy, is what must be certified.[17,28]

Two further directions address the sample-efficiency problem that keeps RL in simulation. Multi-agent coordination through hierarchical RL across Purdue zones, with communication-efficient protocols, matches the zoned structure of real plants. Transfer and meta-learning, in which pre-trained defense policies are adapted to a new facility with few interaction episodes, would reduce the volume of simulated experience required. Until verification or shielding matures, however, the defensible deployment posture remains advisory: the agent recommends, a human authorizes, and the safety instrumented system retains final authority. This review found no reviewed study that

demonstrates a certified autonomous actuating deployment, and none that presents a completed safety case for one.

### 7.3 Federated and privacy-preserving machine learning

Federated learning allows operators to train shared detection models without exporting raw process data. This addresses the data-scarcity and confidentiality barriers documented in Section 6.2.[15,59] Open problems include robust aggregation under highly non-IID facility data, defense against poisoning by compromised participants, and secure multi-party computation for encrypted inference. Regulatory alignment will determine practical viability as much as the algorithms themselves. Relevant regimes include the General Data Protection Regulation, Nigeria's Data Protection Act, and sector-specific rules.

### 7.4 Digital-twin integration

High-fidelity digital twins offer a path around both the data-scarcity and the safe-experimentation problems. Attack scenarios can be simulated at scale to generate labeled training data. Physics-aware anomaly detectors can be calibrated against first-principles process models. Candidate RL policies can be stress-tested before any contact with production systems, which connects directly to the shielding and verification agenda of Section 7.2. Autonomous threat hunting on the twin, in which the model systematically probes for exploitable states before adversaries do, extends the concept from defense validation to proactive assessment.

### 7.5 Quantum-resistant security and AI co-evolution

Long-lived ICS assets will outlast current public-key cryptography, which makes post-quantum migration a near-term planning problem rather than a distant concern.[60] AI can assist by optimizing post-quantum algorithm placement under ICS latency budgets and by stress-testing implementations through machine-learning-guided cryptanalysis. At the same time, the field must anticipate AI-enabled offense, including automated vulnerability discovery and adaptive malware. This turns ICS security into a co-evolutionary contest in which static defenses are structurally disadvantaged.[34]

### 7.6 Standardization, benchmarking, and certification

The field needs shared multi-sector benchmark suites with realistic attack diversity. It also needs standardized reporting of detection, latency, and resource metrics, together with reproducibility requirements such as code and configuration release as publication norms. A cross-dataset evaluation protocol, in which models are tested on a corpus they were not tuned against, would directly address the generalization concerns of Section 6.4. In parallel, certification pathways must be developed so that AI-based protections can be credited within IEC 62443 security levels, and so that safety cases for semi-autonomous response systems become tractable for regulators.[29,30]

### 7.7 Adversarial machine learning and robustness

Because ICS is an adversarial environment, the detectors themselves are targets. Evasion attacks craft inputs that respect learned correlations while masking malicious intent. Poisoning attacks corrupt training data during retraining. Model-stealing attacks reconstruct proprietary detectors for offline evasion development.[34,44,61] Defenses such as adversarial training, certified robustness bounds, and input sanitization are mature in computer-vision research but barely tested on ICS data. This is the most consequential under-explored gap surfaced by the review.

## 8. Case studies and practical implementations

### 8.1. Academic research prototypes

Testbed-validated prototypes illustrate both the promise and the limits of current research. On the SWaT water-treatment testbed, unsupervised deep models reconstruct normal sensor and actuator behavior and flag physical-process manipulation. These include autoencoder and GAN-based detectors, which achieve precision–recall trade-offs competitive with supervised baselines while remaining blind to attacks that respect learned correlations.[41,42,55] Graph-based prototypes for substation automation model IEC 61850 communication topologies and detect coordinated switching attacks that evade per-device monitoring.[18,58] RL prototypes trained in gym-style SCADA simulators learn containment policies that outperform static playbooks, yet every reviewed study confines the agent to advisory mode wherever human factors are considered.[17,28]

### 8.2 Industry deployments

Commercial OT-security platforms increasingly embed machine learning for asset discovery, baseline modeling, and anomaly detection. These products typically combine protocol-aware deep packet inspection with behavioral analytics across vendors serving energy, manufacturing, and water utilities.[13] Public technical detail is limited. Nevertheless, vendor-reported deployments consistently emphasize passive, non-invasive monitoring, which reflects operator intolerance for inline components that could affect availability. They also uniformly retain human analysts in the response loop. The gap between the autonomous-defense literature and shipping products is therefore substantial, and it should temper claims of imminent operational autonomy.

### 8.3 Government and national-laboratory initiatives

Public programs anchor the research-to-practice pipeline. The NIST Cybersecurity Framework and the SP

800-82 guide structure how AI-based capabilities map onto the identify, protect, detect, respond, and recover functions for OT environments.[1,30] Energy-sector programs have funded consortia developing resilient energy delivery systems, including machine-learning-based situational awareness for grid operations.[13] For emerging economies, regional initiatives represent an underexamined deployment context in the reviewed literature. These include national CERT capacity building and critical-infrastructure protection frameworks in Africa, where infrastructure heterogeneity and resource constraints sharpen every challenge identified in Section 6.[57,62]

### 8.4 Lessons learned from deployments

Across prototypes and deployments, several technical lessons recur. Model drift is unavoidable and requires scheduled retraining. False-positive management is central to operator trust. Clean integration with existing SIEM and security-operations workflows is necessary rather than optional. Organizational lessons are equally consistent. Operator training and change management determine adoption more than marginal accuracy gains. Incident-response procedures must be rewritten to incorporate AI-generated evidence, with human accountability clearly assigned.[13,44]

## 9. Discussion

### 9.1 Current state of the field

Clear maturity gradients exist across different AI families. Classical machine learning is operationally mature, uprunning and embedded in commercial products. Deep learning is research-mature and entering production for passive detection. GNNs are advancing rapidly in the literature but remain largely confined to testbeds. RL is the least mature, and no accredited autonomous deployment was identified in the reviewed corpus. The field show evidence of a widening gap between benchmark performance, where publicized accuracy figures continue to climb, and demonstrated operational value, where evidence remains thin.[35,44,51]

### 9.2 Key challenges and barriers

Technical barriers interrelate with operational and regulatory ones. Technical barriers include data scarcity, real-time constraints, scalability, and interpretability. Operational barriers include legacy integration, an availability-first culture, skills shortages, and unclear cost–benefit cases. Regulatory barriers include certification of learning systems, liability for autonomous actions, and cross-border data-sharing restrictions. These barrier classes are mutually reinforcing. The shortcomings involving certification uncertainty suppresses deployment, which limits real-world data, that can perpetuates the simulation-to-reality gap is cite as grounds for caution. Breaking this cycle plausibly requires regulator-sanctioned pilot experimental policy, in which instrumented and bounded deployments generate the operational evidence which the field currently lacks.[29,30]

### 9.3 Interdisciplinary and regional perspectives

Achieving ground breaking progress requires sustained collaboration among AI researchers, control engineers, and security practitioners. Publications authored across these communities were a small minority of the corpus. Geographically, the literature is headed by North American, European, and East Asian testbeds and datasets. Industrializing regions are nearly absent as study contexts, despite facing arguably less favorable risk profiles, because critical-infrastructure modernization, IIoT adoption, and constrained security budgets overlap in this domain.[57,62] Research validating lightweight, low-cost AI defenses in such settings would broaden both the evidence base and the practical reach of the field.

### 9.4 Ethical considerations

Autonomous defensive systems for physical infrastructure raise dual-use and accountability questions. Techniques for modeling attack propagation are informative to attackers as well as defenders. Autonomous response capabilities sit on a continuum that ends in cyber-physical neutralization. Responsibility for harm caused by an flawed autonomous action remains legally unresolved. Responsible-disclosure norms, human-accountable response authority, and explicit dual-use review in research governance are minimalistic safeguards the community should adopt as standard practice.[34,44]

## 10. Conclusions

This review synthesized 147 studies published between 2018 and 2025 which adopts AI in ICS cybersecurity, systematized through a methodology-centric lens covering classical machine learning, deep learning, graph neural networks, reinforcement learning, and hybrid architectures. Three findings stand out. The first is methodological complementarity. Each family optimizes capabilities the others lack, and the appropriate choice is governed by the deployment scenario, as the suitability concrete mapping, rather than by any single accuracy ranking. The second is continual research–practice gap. Benchmark accuracy has largely saturated on the dominant public datasets, while operational evidence on real-time latency, legacy integration, drift management, and operator trust remains scarce. The third is a set of structural gaps in adversarial hardiness, multi-stage attack correlation, transfer learning, and safety–security co-engineering that define the research frontier. For researchers, the priority recommendations are standardized multi-sector benchmarks with mandatory reproducibility, cross-dataset evaluation as a default

reporting requirement, adversarial robustness evaluation as standard practice, and validation pathways advancement from testbeds through digital twins to instrumented pilot deployments. For practitioners, near-term value lies in passive, interpretable detection layered onto existing monitoring, selected against explicit criteria of latency, explainability, and maintenance burden rather than headline accuracy. Autonomous response should be managed through human-in-the-loop consultative modes until certification frameworks mature. AI is necessary but not yet sufficient for the protection of modern industrial infrastructure especially in developing worlds. It remains within a defense-in-depth posture that still depends on sound architecture, disciplined operations, and trained people. The trajectory of both the evolving threat landscape and the technology advancement suggest that institutions learning to integrate AI-based defenses safely, transparently, and early will define the security baseline that the wider sector is eventually required to meet.

**Acknowledgments**
The authors thank the Department of Computer Science and Cybersecurity and the CyberSCADA Laboratory, Obafemi Awolowo University, Ile-Ife, Nigeria for institutional support.

**CRediT Author Contribution Statement**
**Olujoke Mubo Oni:** Conceptualization, Formal analysis, Investigation, Methodology, Visualization, Writing-Original draft. **John Edet Efiong:** Methodology, Data Curation, Formal analysis. **Abiodun Akinwale**: Validation. **Olawumi Adekemi Amoo:** Writing - Review & editing. **Oluseyi Ebunoluwa Balogun**: Writing - Review & editing. **Ayoola Afolabi**: Validation. **Emmanuel Ajayi Olajubu:** Project administration, Supervision. All authors have read and agreed to the published version of the manuscript.

**Funding Declaration**
This research did not receive any specific grant from funding agencies in the public, commercial, or not-for profit sectors.

**Data Availability Statement**
No new primary data were created in this study. The complete data-extraction dataset supporting the synthesis, covering publication metadata, technique categories, datasets, attack types, reported metrics, and quality appraisal for all 147 included studies, is openly available in Zenodo. The search strings, screening decisions, and exclusion reasons underlying Figure 2 are included in the same deposit. The public benchmark datasets discussed in this review, namely SWaT, WADI, BATADAL, the Mississippi State SCADA corpora, and Electra, are available from their original maintainers as cited.

**Conflict of Interest**
There is no conflict of interest.

**Artificial Intelligence (AI) Use Disclosure**
The authors declare that artificial intelligence (AI)-assisted tools were used only for language refinement, grammar improvement, and manuscript structuring purposes during the preparation of this work. All technical content, experimental implementation, results, and interpretations were independently developed and verified by the authors.

**Supporting information**
Not Applicable

**Abbreviations**
The following abbreviations are used in this manuscript:

| Abbreviation | Definition |
|---|---|
| AI | Artificial Intelligence |
| ICS | Industrial Control System |
| SCADA | Supervisory Control and Data Acquisition |
| HMI | Human–Machine Interface |
| IDS | Intrusion Detection System |
| GNN | Graph Neural Network |
| CNN | Convolutional Neural Network |
| GAN | Generative Adversarial Network |
| PPO | Proximal Policy Optimization |
| FDI | False Data Injection |
| IIoT | Industrial Internet of Things |
| DMZ | Demilitarized Zone |
| ML | Machine Learning |
| OT | Operational Technology |
| PLC | Programmable Logic Controller |
| RTU | Remote Terminal Unit |
| SIEM | Security Information and Event Management |
| RL | Reinforcement Learning |
| LSTM | Long Short-Term Memory |
| DQN | Deep Q-Network |
| XAI | Explainable Artificial Intelligence |
| APT | Advanced Persistent Threat |
| SIL | Safety Integrity Level |
| MES | Manufacturing Execution System |

**References**
[1] K. Stouffer, M. Pease, C. Tang, T. Zimmerman, V. Pillitteri, S. Lightman, *Guide to Operational Technology (OT) Security*, NIST Special Publication 800-82 Rev. 3, National Institute of Standards and Technology, Gaithersburg, MD, USA, 2023, Accessed August 15, 2026.
[2] M. Conti, D. Donadel, F. Turrin, A survey on industrial control system testbeds and datasets

for security research, *IEEE Communications Surveys & Tutorials*, 2021, **23**, 2248-2294, doi: 10.1109/COMST.2021.3094360.

[3] E.D. Knapp, *Industrial Network Security: Securing Critical Infrastructure Networks for Smart Grid, SCADA, and Other Industrial Control Systems*, 3rd ed., Syngress, Cambridge, MA, USA, 2024, Accessed August 15, 2026.

[4] A.A. Khan, A.A. Laghari, Z.A. Shaikh, M.A. Dootio, V. Kumar, U.A. Bhatti, Industrial Internet of Things: Recent advances, enabling technologies, and open challenges, *Computers & Electrical Engineering*, 2020, **81**, 106522, doi: 10.1016/j.compeleceng.2019.106522.

[5] T. J. Williams, The Purdue enterprise reference architecture, *Computers in Industry*, 1994, **24**, 141–158, doi: 10.1016/0166-3615(94)90017-5.

[6] J.-P.A. Yaacoub, O. Salman, H.N. Noura, N. Kaaniche, A. Chehab, M. Malli, Cyber-physical systems security: Limitations, issues and future trends, *Microprocessors and Microsystems*, 2020, **77**, 103201, doi: 10.1016/j.micpro.2020.103201.

[7] A. Humayed, J. Lin, F. Li, B. Luo, Cyber-physical systems security-A survey, *IEEE Internet of Things Journal*, 2017, **4**, 1802–1831, doi: 10.1109/JIOT.2017.2703172.

[8] Falliere, N.; Murchu, L.O.; Chien, E. W32.Stuxnet Dossier, version 1.4; Symantec Security Response: Mountain View, CA, USA, 2011, Accessed August 15, 2026.

[9] R. Langner, Stuxnet: Dissecting a cyberwarfare weapon, *IEEE Security & Privacy*, 2011, **9**, 49–51, doi: 10.1109/MSP.2011.67.

[10] R.M. Lee, M.J. Assante, T. Conway, Analysis of the Cyber Attack on the Ukrainian Power Grid, Electricity Information Sharing and Analysis Center (E-ISAC), Washington, DC, USA, 2016, Accessed August 15, 2026.

[11] A. Di Pinto, Y. Dragoni, A. Carcano, TRITON: The first ICS cyber-attack on safety instrument systems, in: Proceedings of Black Hat USA, Las Vegas, NV, USA, 4–9 August 2018,1–26, Accessed August 15, 2026.

[12] K.E. Hemsley, R.E. Fisher, History of Industrial Control System Cyber Incidents, INL/CON-18-44411, Idaho National Laboratory, Idaho Falls, ID, USA, 2018, Accessed August 15, 2026.

[13] Dragos Inc. ICS/OT Cybersecurity Year in Review 2024; Dragos: Hanover, MD, USA, 2025, Accessed August 15, 2026

[14] J. Slowik, Evolution of ICS Attacks and the Prospects for Future Disruptive Events, Dragos, Hanover, MD, USA, 2023, Accessed August 15, 2026.

[15] V. Mothukuri, R.M. Parizi, S. Pouriyeh, Y. Huang, A. Dehghantanha, G. Srivastava, A survey on security and privacy of federated learning, *Future Generation Computer Systems*, 2021, **115**, 619–640, doi: 10.1016/j.future.2020.10.007.

[16] O. ONI, b, Olabode, T. Sholanke, A Hybrid Machine Learning Model for Detecting Anomalies and Cyberattacks in IoT Networks, *Ife Journal of Information and Communication,* 2024, **8**, 2024, https://ijict.oauife.edu.ng/index.php/IJICT/article/view/15/14

[17] T. T. Nguyen, V.J. Reddi, Deep reinforcement learning for cyber security, *IEEE Transactions on Neural Networks and Learning Systems*, 2023, **34**, 3779–3795, doi: 10.1109/TNNLS.2021.3121870.

[18] A. Presekal, A. Stefanov, V.S. Rajkumar, P. Palensky, Attack graph model for cyber-physical power systems using hybrid deep learning, *IEEE Transactions on Smart Grid*, 2023, **14**, 4007–4020, doi: 10.1109/TSG.2023.3237011.

[19] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, P.S. Yu, A comprehensive survey on graph neural networks, *IEEE Transactions on Neural Networks and Learning Systems*, 2021, **32**, 4–24, doi: 10.1109/TNNLS.2020.2978386.

[20] R. Kaur, D. Gabrijelčič, T. Klobučar, Artificial intelligence for cybersecurity: Literature review and future research directions, *Information Fusion*, 2023, **97**, 101804, doi: 10.1016/j.inffus.2023.101804.

[21] M. Eckhart, A. Ekelhart, Digital twins for cyber-physical systems security: State of the art and outlook, in: Security and Quality in Cyber-Physical Systems Engineering, Springer, Cham, Switzerland, 2019, 383–412, doi: 10.1007/978-3-030-25312-7_14.

[22] R. Mitchell, I.-R. Chen, A survey of intrusion detection techniques for cyber-physical systems, *ACM Computing Surveys*, 2014, **46**, 1 –-29, doi: 10.1145/2542049.

[23] M. Kaouk, J.-M. Flaus, M.-L. Potet, R. Groz, A review of intrusion detection systems for industrial control systems, in: Proceedings of the 6th International Conference on Control, Decision and Information Technologies (CoDIT), 2019, 1699–1704, doi: 10.1109/CoDIT.2019.8820537.

[24] M. Kravchik, A. Shabtai, Detecting cyber-attacks in industrial control systems using convolutional neural networks, in: Proceedings of the Workshop on Cyber-Physical Systems Security and Privacy (CPS-SPC, 2018, 72–83, doi: 10.1145/3264888.3264890.

[25] M.R. Asghar, Q. Hu, S. Zeadally, Cybersecurity in industrial control systems: Issues, technologies, and challenges, *Computer Networks*, 2019, **165**, 106946, doi: 10.1016/j.comnet.2019.106946.

[26] J.M. Beaver, R.C. Borges-Hink, M.A. Buckner, An

evaluation of machine learning methods to detect malicious SCADA communications, in: Proceedings of the 12th International Conference on Machine Learning and Applications (ICMLA), 2013, 54–59, doi: 10.1109/ICMLA.2013.18.

[27] A. Deng, B. Hooi, Graph neural network-based anomaly detection in multivariate time series, *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, **35**, 4027–4035, doi: 10.1609/aaai.v35i5.16523.

[28] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal policy optimization algorithms, *arXiv*, 2017, arXiv:1707.06347, doi: 10.48550/arXiv.1707.06347.

[29] International Electrotechnical Commission, IEC 62443: Security for Industrial Automation and Control Systems, IEC, Geneva, Switzerland, 2020, Accessed August 15, 2026.

[30] M.J. Page, J.E. McKenzie, P.M. Bossuyt, I. Boutron, T.C. Hoffmann, C.D. Mulrow, L. Shamseer, J.M. Tetzlaff, E.A. Akl, S.E. Brennan, R. Chou, The PRISMA 2020 statement: An updated guideline for reporting systematic reviews, *BMJ*, 2021, **372**, doi: 10.1136/bmj.n71.

[31] A. Hahn, R.K. Thomas, I. Lozano, A. Cardenas, A multi-layered and kill-chain based security analysis framework for cyber-physical systems, *International Journal of Critical Infrastructure Protection*, 2015, **11**, 39–50, doi: 10.1016/j.ijcip.2015.08.003.

[32] S. Kim, G. Heo, E. Zio, J. Shin, J.-G. Song, Cyber-attack taxonomy for digital environment in nuclear power plants, *Nuclear Engineering and Technology*, 2020, **52**, 995–1001, doi:10.1016/j.net.2019.11.001.

[33] Y. Liu, P. Ning, M.K. Reiter, False data injection attacks against state estimation in electric power grids, *ACM Transactions on Information and System Security*, 2011, **14**, 1-33, doi:10.1145/1952982.1952995.

[34] G. Apruzzese, M. Andreolini, L. Ferretti, M. Marchetti, M. Colajanni, Modeling realistic adversarial attacks against network intrusion detection systems, *Digital Threats: Research and Practice*, 2022, **3**, 1–19, doi:10.1145/3502864.

[35] R.C.B. Hink, J.M. Beaver, M.A. Buckner, T. Morris, U. Adhikari, S. Pan, Machine learning for power system disturbance and cyber-attack discrimination, in: Proceedings of the 7th International Symposium on Resilient Control Systems (ISRCS), 2014, 1–8, doi:10.1109/ISRCS.2014.6900095.

[36] M. Esmalifalak, L. Liu, N. Nguyen, R. Zheng, Z. Han, Detecting stealthy false data injection using machine learning in smart grid, *IEEE Systems Journal*, 2017, **11**, 1644-1652, doi: 10.1109/JSYST.2014.2341597.

[37] J. Inoue, Y. Yamagata, Y. Chen, C. M. Poskitt, J. Sun, Anomaly detection for a water treatment system using unsupervised machine learning, in: Proceedings of 2017 IEEE International Conference on Data Mining Workshops (ICDMW), New Orleans, LA, USA, 2017, 1058-1065, doi: 10.1109/ICDMW.2017.149.

[38] S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural Computation*, 1997, **9**, 1735–1780, doi: 10.1162/neco.1997.9.8.1735.

[39] C. Feng, T. Li, D. Chana, Multi-level anomaly detection in industrial control systems via package signatures and LSTM networks, in: Proceedings of the 47th IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), Denver, CO, USA, 26–29 June 2017, 261–272, doi: 10.1109/DSN.2017.34.

[40] D. Nedeljkovic, Z. Jakovljevic, CNN based method for the development of cyber-attacks detection algorithms in industrial control systems, *Computers & Security*, 2022, **114**, 102585, doi: 10.1016/j.cose.2021.102585.

[41] J. Goh, S.Adepu, K. N. Junejo, A. Mathur, A dataset to support research in the design of secure water treatment systems, A Dataset to Support Research in the Design of Secure Water Treatment Systems. In: Havarneanu, G., Setola, R., Nassopoulos, H., Wolthusen, S. (eds) Critical Information Infrastructures Security. CRITIS 2016. Lecture Notes in Computer Science, 2017, 10242, 88–99, Springer, Cham. 10.1007/978-3-319-71368-7_8.

[42] C. M. Ahmed, V. R. Palleti, A. P. Mathur, 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, 25–28, doi: 10.1145/3055366.3055375.

[43] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, *Communications of the ACM*, 2020, **63**, 139–144, doi: 10.1145/3422622.

[44] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017, 4765–4774.

[45] T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks. In Proceedings of the 5th International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.

[46] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, Y. Bengio, Graph attention networks, in Proceedings of the 6th International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.

[47] W. L. Hamilton, R.Ying, J. Leskovec, Inductive representation learning on large graphs, in Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017, 1024–1034.

[48] J. Zhang, L. Pan, Q.-L. Han, C. Chen, S. Wen, Y. Xiang, Deep learning based attack detection for cyber-physical system cybersecurity: A survey, *IEEE/CAA Journal of Automatica Sinica*, 2022, **9**, 377-391, doi: 10.1109/JAS.2021.1004261.

[49] R. S. Sutton, A. G. Barto, Reinforcement learning: An introduction, 2nd edition, MIT Press, Cambridge, MA, USA, 2018.

[50] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, D. Hassabis, Human-level control through deep reinforcement learning, *Nature*, 2015, 518, 529–533, doi: 10.1038/nature14236.

[51] R. Taormina, S. Galelli, Deep-learning approach to the detection and localization of cyber-physical attacks on water distribution systems, *Journal of Water Resources Planning and Management*, 2018, **144**, 04018065, doi: 10.1061/(ASCE)WR.1943-5452.0000983.

[52] P. D. Rosero-Montalvo, P. Tozun, W. Hernandez, Hybrid anomaly detection model on trusted IoT devices, *IEEE Internet Things Journal*, 2023, **10**, 10959–10969, doi: 10.1109/JIOT.2023.3243037.

[53] T. Morris, W. Gao, Industrial control system traffic data sets for intrusion detection research, in Critical Infrastructure Protection VIII; Springer: Berlin/Heidelberg, Germany, 2014, 65–78.

[54] S. D. Anton, S. Kanoor, D. Fraunholz, H. D. Schotten, Evaluation of machine learning-based anomaly detection algorithms on an industrial Modbus/TCP data set. In Proceedings of the 13th International Conference on Availability, Reliability and Security (ARES), Hamburg, Germany, 27–30 August 2018; pp. 1–9.

[55] R. Taormina, S. Galelli, N.O. Tippenhauer, E. Salomons, A. Ostfeld, D.G. Eliades, M. Aghashahi, R. Sundararajan, M. Pourahmadi, M.K. Banks, B.M. Brentan, E. Campbell, G. Lima, D. Manzi, D. Ayala-Cabrera, M. Herrera, I. Montalvo, J. Izquierdo, E. Luvizotto, Z. Ohar, Battle of the attack detection algorithms: Disclosing cyber-attacks on water distribution networks, *Journal of Water Resources Planning and Management*, 2018, **144**, 04018048, doi:10.1061/(ASCE)WR.1943-5452.0000849.

[56] M. Zolanvari, M.A. Teixeira, L. Gupta, K.M. Khan, R. Jain, Machine learning-based network vulnerability analysis of industrial Internet of Things, *IEEE Internet of Things Journal*, 2019, **6**, 6822–6834, doi: 10.1109/JIOT.2019.2912022.

[57] O. M. Oni, A. B. Omirinlewo, O. E. Balogun, J. E. Efiong, A. Afolabi, E. A. Olajubu, A predictive framework for dynamic risk management in industrial control system networks, *Journal of Future Artificial Intelligence and Technology*, 2026, **3**, 1–18, doi: 10.62411/faith.3048-3719-348.

[58] J. Hong, C.-C. Liu, M. Govindarasu, Integrated anomaly detection for cyber security of the substations, *IEEE Transactions on Smart Grid*, 2014, **5**, 1643–1653, doi: 10.1109/TSG.2013.2294473.

[59] T.T. Huong, T.P. Bac, D.M. Long, B.D. Thang, N.T. Binh, T.D. Luong, T.K. Phuc, Detecting cyberattacks using anomaly detection in industrial control systems: A federated learning approach, *Computers in Industry*, 2021, **132**, 103509, doi: 10.1016/j.compind.2021.103509.

[60] G. Alagic, D. Apon, D. Cooper, Q. Dang, T. Dang, J.M. Kelsey, J. Lichtinger, Y.-K. Liu, C.A. Miller, D. Moody, R. Peralta, R. Perlner, A. Robinson, D. Smith-Tone, Status Report on the Third Round of the NIST Post-Quantum Cryptography Standardization Process, *NIST IR 8413*, 2022, doi:10.6028/NIST.IR.8413.

[61] E. Anthi, L. Williams, A. Javed, P. Burnap, Hardening machine learning denial of service (DoS) defences against adversarial attacks in IoT smart home networks, *Computers & Security*, 2021, **108**, 102352, doi: 10.1016/j.cose.2021.102352.

[62] N. Kshetri, Cybersecurity in African countries: Problems, prospects and progress, *Journal of Global Information Technology Management*, 2019, **22**, 77–81, doi:10.1080/1097198X.2019.1603527.

[63] M. Bada, B. von Solms, I. Agrafiotis, Reviewing national cybersecurity awareness for users and executives in Africa, *International Journal of Advanced Security*, 2019, **12**, 108–118, doi: 10.48550/arXiv.1910.01005.