International Journal of
Technology and Emerging Research
ISSN: 3068-109X (online) | ISSN: 3068-1995 (print) | https://www.ioro.org/ijter

Cyber Attack Identification in Industrial Control Systems: A Review of Dynamic Watermarking and Machine Learning Applications

Maidhili Mohan K.[1], **Aysha M. K.** [2]

[1] *Department of Computer Science, St. Thomas College (Autonomous), Thrissur, India*

[2] *Department of Computer Science, St. Thomas College (Autonomous), Thrissur, India*

* **Corresponding author:** maidhilimohan@gmail.com • ORCID: 0009-0003-9202-5546

Abstract

Industrial control systems (ICS) are fundamental to critical infrastructures such as electrical grids, water processing plants, and manufacturing plants, where Programmable Logic Controllers (PLCs) and Supervisory Control and Data Acquisition (SCADA) systems regulate physical processes that cannot stop operating at ease. These formerly isolated systems acquire a greater variety of challenges as they become interconnected with IT networks, including command-injection, replay, and false-data injection attacks that may actually result in serious physical harm in addition to some data loss. Two defense methods that have been evolved are discussed in this review. By integrating a private random signal on the control command, dynamic watermarking (DW) takes on an active, physics-based approach resulting in any anomaly in the sensor-actuator feedback loop becoming statistically visible. In contrast, machine learning (ML) approaches try to learn what malicious behavior appears from data. We keep track of the innovations in both directions to distinguish their supplementing powers and blind spots, as well as look at the simple but growing set of work that attempts to bring them together. Also, we employed the case studies from PLC-controlled water-tank testbeds, networked control systems, and power-system automatic generation control. The paper ends by discussing the barriers that prevent these research findings from being put into practice, including the ability to scale, adversarial robustness, and real-time deployment on legacy PLC hardware.

Keywords: Industrial Control Systems; SCADA, Programmable Logic Controller; Cyber-Physical Systems; Machine Learning; Intrusion Detection; Dynamic Watermarking; Anomaly Detection

1. Introduction

The technologies like Programmable Logic Controllers (PLCs), Remote Terminal Units (RTUs), sensors, actuators, and SCADA software are all associated with Industrial Control Systems (ICS) to maintain physical processes in fields like discrete production, oil and gas, water and wastewater treatment, and electric power. For a long period of time, these systems relied on security-by-obscurity, isolation, and proprietary protocols. Commercial-off-the-shelf hardware, IP-based networking, and remote access have minimized the cost of running and tracking ICS, but they also made a PLC-dependent infrastructure prone to the same threats that focus on regular IT networks [1]. The most frequently reported incidents are Stuxnet, Industroyer/CrashOverride, and Triton, all of them have shown that a successful ICS intrusion is not restricted to a screen. It could be an interruption to any service that people rely on, or a safety risk.

Two philosophies dominate the response to this problem. One is data-driven: machine learning models are trained on network traffic, protocol fields, or process telemetry to recognize the statistical fingerprint of an attack [2]. The other works from physics rather than statistics — dynamic watermarking injects a private, statistically independent random signal into the control command and checks, through hypothesis testing on the returned sensor data, whether the plant's response is consistent with what its known dynamics predict [3]. The appeal of DW is that it needs no attack history and probes the physical layer directly, which lets it catch stealthy replay and false-data-injection attacks that would otherwise slip past a purely statistical detector; ML, in turn, is better positioned to catch anomalies at the network and protocol layer that DW was never designed to see. The ICS threat landscape is covered first in this study, subsequently addressing ML-based detection, DW, developing hybrid work, and finally remaining concerns.

2. ICS Architecture and Threat Landscape

A PLC implementing ladder-logic or structured-text program codes that access the sensor to compute a control action, and activate actuators often makes up an ICS control loop. Meanwhile, a SCADA layer incorporates Ethernet/IP, Modbus, Profinet, or DNP3 for facilitating set-point adjustment and supervisory monitoring. To cover up an ongoing attack, any user with network access may potentially change the sensor data that the controller sees, alter the commands given to actuators, rewrite PLC logic completely, or just repeat already recorded valid data. In order to arrange their anomaly detector, Choi et al. organized this set of methods onto the MITRE ATT&CK for ICS framework: primary access, execution, persistence, evasion, discovery, lateral motion, a collection, command-and-control, inhibit-responsive-function, impair-process-control, and impact [2]. Huang et al. make a similar point from a physical point of view: sensor-data manipulation alone pushes PLC-dependent water and chemical tank systems into overflow or uncontrolled operation without setting off any conventional alert system. This is completely the case for a defense that directly analyzes the physical signal rather than just the network [1]. Detector designs in the following sections are finally looked at against a number of simulation and PLC-in-the-loop studies that have gone deeper by thoroughly evaluating which control algorithms and attack timings lead to the greatest damage.

3. Machine-Learning-Based Detection

ICS ML-based intrusion detection systems typically belong to one of three groups. The first group, known as supervised classifiers, consist of random forests, support vector machines, and gradient-boosted trees, that are trained on labeled normal/attack traffic. The second group is known as unsupervised or semi-supervised detectors, such as autoencoders, isolation forests, and one-class SVMs, which create a framework of usual behavior and indicate anything that varies from it, rather than employ attack labels at all. A third group uses deep sequential models, including GRU networks and LSTM, to recognize temporal dependencies that basic models fail to recognize in the process or network time series.

The studies of Choi et al. give a fine example of the path that field of study is taking. Based on the finding, the ICS traffic is minimal and fluctuating rather than always continuous, they devised a Zero-Inflated Poisson (ZIP)-based GRU model that challenges the Gaussian-noise assumptions that are found in traditional models [2]. The detector had more than 95 percent detection accuracy when tested against simulated Stuxnet and Industroyer cases. Significantly, the anomalies were traced back onto MITRE ATT&CK for ICS techniques, transforming a raw alarm into something resembling an actionable diagnosis. For it being obvious that an operator may act on a specific method of identification in a way that they cannot act on an anomaly flag, the

integration of a statistical or deep-learning score with a structured threat-intelligence framework is becoming a common practice.

The ability of machine learning (ML) to capture complex and high-dimensional patterns directly from data removes the necessity of manual construction of a physical model of the plant. It deals with nearly the exact reverse of that strength: training data must be unbiased, attack scenarios are unusual, and real process data is generally considered too sensitive to share; models are prone to adversarial defense evasion and concept drift as normal operating circumstances fluctuate over time; and their results are not always clear enough to interpret, and this is important when an operator needs to defend a shutdown or an incident report that has to hold up to scrutiny.

4. Dynamic Watermarking

Satchidanandan and Kumar's Dynamic Watermarking relies on a different approach. Before the initial control input reaches the plant, the controller introduces a watermark—a private, zero-mean, statistically independent random signal [3]. Because an attacker does not know this signal, any sensor or actuator manipulation that is inconsistent with how the real plant dynamics would have carried that watermark forward shows up in one of two statistical tests run on the returned sensor data: a correlation test between the watermark and the measurement residual, and a variance test on the residual itself. What sets this apart from a purely statistical anomaly detector is that, under fairly mild conditions on the system model, it comes with an actual proof — an attacker cannot stay undetected while also driving an arbitrarily large deviation from the true state. That is exactly the property that makes DW effective against sophisticated replay and stealthy false-data-injection attacks, which are built precisely to look statistically unremarkable.

The idea has since moved off the whiteboard. Huang et al. implemented DW on a PLC-dependent coupled water-tank testbed built around an Allen-Bradley Micro820 controller, embedding the watermark in the signal driving the tank's motor and pump, and validated detection across several attack scenarios, including sensor-measurement compromise aimed at inducing overflow [1]. Their results are a useful existence proof: DW runs on commodity industrial hardware, in real time, at modest computational cost. Continuously transmitting and testing a watermark is not free, however, and Du et al. addressed that overhead with an event-triggered variant for networked control systems, activating the watermark and its statistical tests only when a triggering condition is met — lowering bandwidth use without giving up the detection guarantees [4]. Huang, Satchidanandan, Kumar, and Xie pushed the idea into the power-systems domain, extending DW to Automatic Generation Control (AGC) loops and generalizing the hypothesis tests to AGC's multi-area, multi-timescale structure, then validating the approach against attacks that would otherwise trigger unsafe frequency deviations [5]. Taken together, these results trace DW's path from a single laboratory testbed to networked and large-scale cyber-physical infrastructure.

DW's strength is that it is grounded in the physics of the plant rather than in a history of past attacks, which is also what makes it resistant to stealthy manipulation that looks normal at the network layer. Its weaknesses come from the same source: it needs an accurate — or accurately identified — plant model, the watermark itself adds a small amount of actuation noise, and integrating the watermark generator and its statistical tests into a legacy PLC control loop takes real engineering effort, more so for nonlinear or switching systems.

5. Toward Hybrid ML–Dynamic-Watermarking Architectures

ML and DW operate at different layers and lean on different assumptions, and a growing body of work treats that as an opportunity rather than a conflict. ML is a natural fit for network- and protocol-layer anomalies, high-dimensional feature spaces, and classifying an attack once something has already been flagged; DW is a natural fit for verifying the physical integrity of the sensor-actuator loop with an actual statistical guarantee, even when the network traffic looks entirely legitimate. The hybrid designs that have emerged so far tend to go one of two directions. Some feed DW's correlation and variance-test residuals into an ML classifier as engineered features, letting the model fuse physical-layer integrity evidence with network-layer indicators. Others run it the other way: ML anomaly scores at the network layer adaptively tighten DW's detection thresholds, or trigger event-based watermarking in the spirit of [4] only once suspicious network activity has already appeared, cutting down on continuous overhead. A further pattern uses ML, once a DW alarm has fired, to classify what kind of attack it is — replay, bias-injection, denial-of-service — and map that onto a MITRE ATT&CK for ICS technique as in [2], pairing DW's detection rigor with ML's diagnostic and response-oriented strengths.

Table I shows the comparative analysis across the two paradigms and their hybrid combination.

Table 1. Comparison of Detection Paradigms.

	Machine Learning	**Dynamic Watermarking**	Hybrid ML–DW
Detection layer	Network / protocol / process data	Physical control loop	Both layers jointly
Attack-label dependence	Needed for supervised variants	None — needs a plant model instead	Reduced reliance on either alone
Robustness to stealthy/replay attacks	Moderate to low	High, with formal guarantees	High, plus attack classification
Computational overhead	Low to moderate	Low to moderate, at the controller	Reduced via event-triggering [4]
Physical plant model required	No	Yes (identified or known)	Partially offset by ML features

6. Open Challenges and Future Directions

There are a few limitations preventing this study from being widely used. Lightweight and event-triggered architecture designs are crucial since real-time constraints on traditional PLC hardware restrict the computational resource available for watermark generation and validation, even with any on-device ML inference [4]. The ML component also has its own exposure: an attacker aware of the detector could craft traffic specifically to evade it, and DW's statistical guarantees are one of the few dependable backstops against that. Scaling beyond single-loop laboratory testbeds to large, multi-PLC, multi-protocol ICS networks is still largely unexplored territory, and it is not obvious how cleanly DW generalizes to nonlinear, switching, or multi-area systems like AGC [5]. Progress on all of this is also bottlenecked by data: the field needs shared, realistic ICS testbeds that pair network traffic with physical process data and ground-truth attack labels, since

without that kind of shared benchmark it is hard to compare ML, DW, and hybrid detectors on equal footing. Finally, mapping detector output onto a structured framework like MITRE ATT&CK for ICS — as [2] does for ML — is a promising way to turn a raw detection signal into something an operator can act on; doing the same for physics-based DW alarms remains, as far as we can tell, an open problem.

7. Conclusion

Machine learning and dynamic watermarking approach the same problem from different directions: one learns the statistical signature of malicious behavior from network or process data, the other actively and provably checks the physical integrity of the sensor-actuator feedback loop. Both have now been demonstrated on realistic PLC-controlled testbeds and on networked or power-system infrastructure, and the early hybrid work suggests that combining physical-layer watermarking evidence with data-driven anomaly detection and threat-framework mapping produces a more robust — and more actionable — defense than either approach used alone. What remains is largely engineering and evaluation: lightweight real-time implementations, adversarial robustness, and shared testbeds that let these results be compared fairly and, eventually, deployed to protect infrastructure that cannot afford to fail.

Acknowledgements

We acknowledge the productive contributions of the scientific community whose work in the field of cyber-attack identification in Industrial Control Systems has laid the foundation for this review work. Special thanks to the developers and researchers behind the pioneering tools such as PLC-dependent infrastructure [1], Zero-Inflated Poisson (ZIP)-based GRU model [2], and Dynamic watermarking [3], whose innovations continue to drive the field forward. This review would not have been possible without the collective efforts of enormous researchers over the past five decades. We acknowledge the support and resources provided by St. Thomas College (Autonomous), Thrissur which has facilitated the successful completion of this research.

Funding

This research received no external funding.

Conflict of Interest

The authors declare no conflict of interest.

AI Usage Disclosure

No generative AI tools were used.

Author Contributions

Conceptualization, Maidhili Mohan K. and Aysha M. K.; methodology, Maidhili Mohan K.; analysis, Aysha M. K.; writing—original draft, Maidhili Mohan K.; writing—review and editing, all authors. All authors have read and agreed to the published version of the manuscript.

References

[1] P.-H. Huang, P. R. Kumar, J. Rajendran, and P. Enjeti, “Enhancing Cybersecurity for Industrial Control Systems: Innovations in Protecting PLC-Dependent Industrial Infrastructures,” IEEE Internet of Things Journal, vol. 11, no. 22, pp. 36486–36493, Nov. 2024, doi: 10.1109/JIOT.2024.3408098.

[2] W. Choi, S. Pandey, and J. Kim, “Detecting Cybersecurity Threats for Industrial Control Systems Using Machine Learning,” IEEE Access, vol. 12, pp. 153550–153563, 2024, doi: 10.1109/ACCESS.2024.3478830.

[3] B. Satchidanandan and P. R. Kumar, “Dynamic Watermarking: Active Defense of Networked Cyber–Physical Systems,” Proceedings of the IEEE, vol. 105, no. 2, pp. 219–240, Feb. 2017, doi: 10.1109/JPROC.2016.2575064.

[4] D. Du, C. Zhang, X. Li, M. Fei, and H. Zhou, “Attack Detection for Networked Control Systems Using Event-Triggered Dynamic Watermarking,” IEEE Transactions on Industrial Informatics, vol. 19, no. 1, pp. 351–361, Jan. 2023, doi: 10.1109/TII.2022.3168868.

[5] T. Huang, B. Satchidanandan, P. R. Kumar, and L. Xie, “An Online Detection Framework for Cyber Attacks on Automatic Generation Control,” IEEE Transactions on Power Systems, vol. 33, no. 6, pp. 6816–6827, Nov. 2018, doi: 10.1109/TPWRS.2018.2829743.