International Journal of AI, BigData, Computational and Management Studies
Noble Scholar Research Group | Volume 4, Issue 2, PP. 180-185, June - 2023
ISSN: 3050-9416 | https://doi.org/10.63282/3050-9416.IJAIBDCMS-V4I2P118

Artificial Intelligence & Machine Learning for Anomaly Detection in ICS Environments

Vilas Shewale
Independent Cybersecurity Researcher, USA.

Abstract: Since telemetry data generated in ICS environments can be both vast and difficult to interpret via traditional, rule-based intrusion detection methods, machine learning techniques have emerged to help close the gap in ICS-specific detection. Machine learning-including techniques like supervised classification (e. g, using Support Vector Machines or Decision Trees), unsupervised approaches (e. g. , for baseline normal activity or anomaly detection) and more recent deep models applied to time-series data-has recently been introduced and increasingly used in ICS security settings. This survey introduces machine learning techniques for ICS anomaly detection, discusses the implementation patterns used to create productive ML-based intrusion detection systems based on passive collection and feedback loops, highlights commonly seen operational challenges with these types of systems and names specific problems encountered in deployment, such as false positive economics, model drift, training data scarcity and an explainability gap that impacts analyst work. The target audience of this survey includes security engineers at facilities using industrial control systems. The survey should provide a sense of the value, limitations and pitfalls of modern AI-driven intrusion detection. It aims to equip the audience to more critically examine vendor claims for ML-based products.

Keywords: Machine Learning, Anomaly Detection, ICS, OT Security, Time-Series, Autoencoder, LSTM, Deep Learning.

1. Introduction

However, detection within industrial control system (ICS) environments poses different challenges than in enterprise IT networks. Traditional signature-based tools-effective against commodity malware- are ill-suited to ICS for multiple reasons: The protocols used are arcane and vendor-unique. Rare attack samples appropriate for antivirus engines are often classified. ICS are superimposed on physical processes whose normal behavior fluctuates with operating mode, season and demand. Relying only on known bad signatures will result in gaps.

Machine learning proposes an alternative: instead of characterizing what is malicious, learn what is normal and identify anomalies. The idea holds promise in ICS because process data is naturally recurrent: A pump operating steadily yields a tight band of vibration values. A pipeline at constant throughput registers a tight band of pressure differentials. A control room shift entails a narrow band of operator commands. A shift in the band indicates an alteration. These changes may be inconsequential-a setpoint update or a planned maintenance outage. Alternatively, they might signal the emerging traces of an attack.

This article discusses how this premise is realized. Section 2 reiterates why traditional detection methods perform poorly in ICS. Section 3 outlines the methods currently in use. Section 4 describes deployment configurations. Section 5 examines the ongoing operational challenges post-proof-of-concept. Section 6 concludes with a preview of future developments and the early effects of generative models on attack and defense strategies. The material is targeted at security architects and engineering managers than ML specialists. Mathematical specifics are omitted, references at the end guide the reader toward foundational literature.

2. Why Traditional Detection Underperforms In Ics

A good reason for new detections stems from the limitations inherent in previous detection strategies. There are at least three important such limits that ought to be explicitly identified.

Signature-based detection presumes the repetition of attack signatures, assuming the viability of collecting and sharing these signatures by defenders. Although a reasonably accurate assumption within commodity IT systems, this assumption holds less ground within OT. The TRITON malware, first discovered in 2017, was used in just one attempted assault on a safety system prior to its public unveiling. Similarly, the INCONTROLLER framework made its debut publicly in 2022 before any recorded in-the-wild activities [1]. In many cases, by the time a signature can be published, the threat campaign that generated it will be over or has evolved beyond it. OT operators relying on such feeds are necessarily working backward than forward, in time and detection.

Rule-based detection methods involve analysts writing specific "if this pattern occurs, trigger an alert" rules. This works as long as an analyst can clearly define what constitutes a harmful pattern. However, within OT, a high proportion of malicious

activity requires an understanding of its contextual situation, for instance, an attack command that is perfectly valid when issued during certain system states is suspicious if used during others. Likewise, a change to a setpoint can appear reasonable when made by a shift-working engineer on the premises, but entirely abnormal if made otherwise. Writing all the rules needed to capture these various contextual anomalies results in brittle systems where, after the original analyst has moved on, neither he nor anyone else can reasonably manage the growing rule set.

Network intrusion detection assumes that protocol standards are clearly and readily documented and that network traffic patterns tend to be relatively static. Unfortunately, OT protocols are often undocumented, private or poorly supported by the common open-source intrusion detection engines available for network use. Furthermore, even where OT protocols are well-understood, the underlying network traffic patterns are not generally stationary in the way enterprise traffic patterns sometimes are baseline network behaviors frequently change along with operational circumstances, such as during routine maintenance procedures. Network-based static rule sets therefore remain perennially either far too noisy or much too quiet depending on the moment in time.

No one of these limits should necessarily discourage the ongoing reliance on signatures, rules or network intrusion detection, they remain valuable as a component in an overall defense-in-depth approach to OT security. The argument to be made, however, is that their effectiveness has diminished over time and that this reduced effectiveness has precisely opened the gap where machine learning can be most beneficially applied.

3. Techniques Applied to ICS Detection

The landscape of machine learning techniques applied to ICS anomaly detection has expanded considerably over the past five years. Four families show up most often in the literature and in commercial products.

3.1. Supervised Classification

A supervised classification scheme can effectively create a direct transformation between an input and its associated category. For industrial control systems, attack kinds or whether traffic is normal versus harmful represent these categories. Machine learning approaches like decision tree forests, boosting-based techniques (such as boosted trees) and SVMs have been put to work and papers have often compared how well they fare on datasets like SWaT and a collection produced by Mississippi State [2] in terms of testbed performance for pipeline gas. Machine-learning classification algorithms have the potential to produce unbiased probabilities that clearly make sense if there is a sufficient quantity of labeled training examples. For systems such as industrial control systems, however, labeled training examples are seldom readily accessible. Attack sequences have seldom occurred on production systems and simulated attacks used in testbeds may only partly reflect the multitude of ways attackers might compromise operational systems.

3.2. Unsupervised Anomaly Detection

With unsupervised methods, data does not need labels. The machine first creates a baseline of normal behavior and flags instances where activity deviates from this behavior. Machine learning methods can range from simple statistical parametric methods (Gaussian mixture models, isolation forests) to non-parametric techniques (one-class SVMs, distance-based approaches). Unsupervised methods tend to dominate within ICS security applications, primarily because most available ICS operational data are labeled as normal operations, whereas ICS attack data are typically not labeled. A drawback is that almost any behavior that deviates from the data training used for establishing the model would be marked as an anomaly, even though it might be part of routine operation that the model was not trained to recognize (like regular maintenance activities, load changes for seasons or process deviations the plant experienced prior to the attack). All these will appear as alerts if the model has not already seen them or was intentionally created in a way to tolerate or ignore them.

3.3. Deep Learning for Time-Series

Because ICS telemetry is almost entirely time-series based, recurrent neural networks (especially LSTMs by Hochreiter & Schmidhuber 1997 [3]) can be used to model time dependencies in process variables and provide one-step-ahead prediction. We calculate anomaly scores as the difference between the predicted and true value. Anomaly detection techniques can also be built using the autoencoder framework. The inputs are first compressed into a latent space representation and then reconstructed. Differences between predicted and real values give us the anomaly signal. LSTM autoencoders and some more recent transformer autoencoders also fit within the same model schema. There are some recent publications using deep learning models with several baselines on some publicly available ICS datasets [4].

3.4. Graph-Based and Process-Aware Methods

The processes in a plant are connected in an organized manner: A pump supplies fluid into a storage tank, that tank feeds an appliance somewhere down the line and we can read what connects what right from looking at process and instrumentation diagrams (P&IDs). In graph-based detection approaches, that same network of connections is built into the process model, so that unusual anomalies are measured not merely against how individual sensors perform, but also in comparison to the combination of sensors implied by the way the equipment actually works. Domain experts use the identical idea by applying

their knowledge, as an instance, the pressure within equipment is too great to safely accommodate by way of hydrostatic limitations even if the model has not encountered this specific measurement reading before. In reality, joining intelligent, learned models with expert constraints on process dynamics yields lower percentages of unwarranted alarms than either approach taken independently.

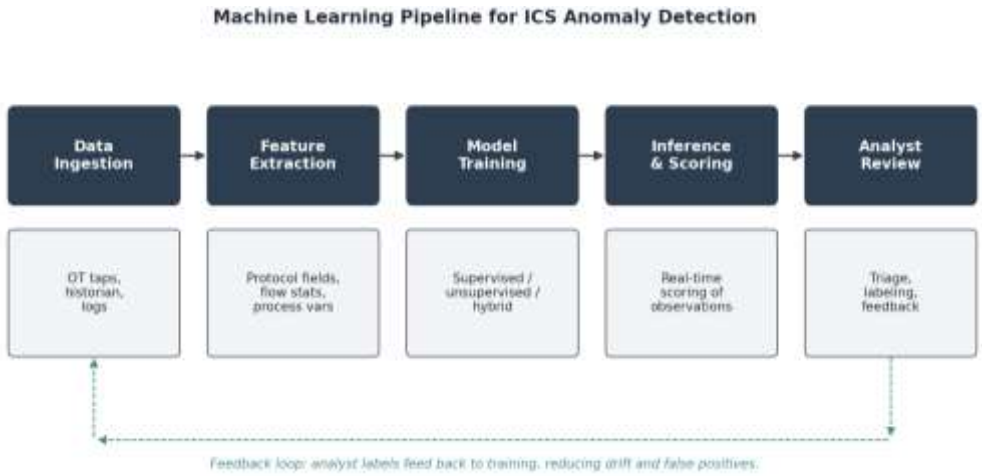


Figure 1: A Typical Machine Learning Pipeline for ICS Anomaly Detection. Analyst Feedback Closes the Loop and Helps the Model Adapt to Operational Change

4. Deployment Patterns

Choosing a technique is the easier half of the problem. Getting it deployed in a production OT environment without disrupting operations or burying analysts in alerts is the harder half.

4.1. Passive Collection

This predominantly takes on the setup of being out of band, network taps/SPANs in the switch ports send traffic to another host collecting this data, meanwhile, process information goes to the same collector from replica historians or read-only OPC-UA streams. Here the detection does not exist within the control loop or path and, therefore, the failure cannot affect operations. The plant standards usually require this to enforce no failures in the critical control path-which placing inline filters on this route would absolutely introduce and which most organizations will not tolerate [5].

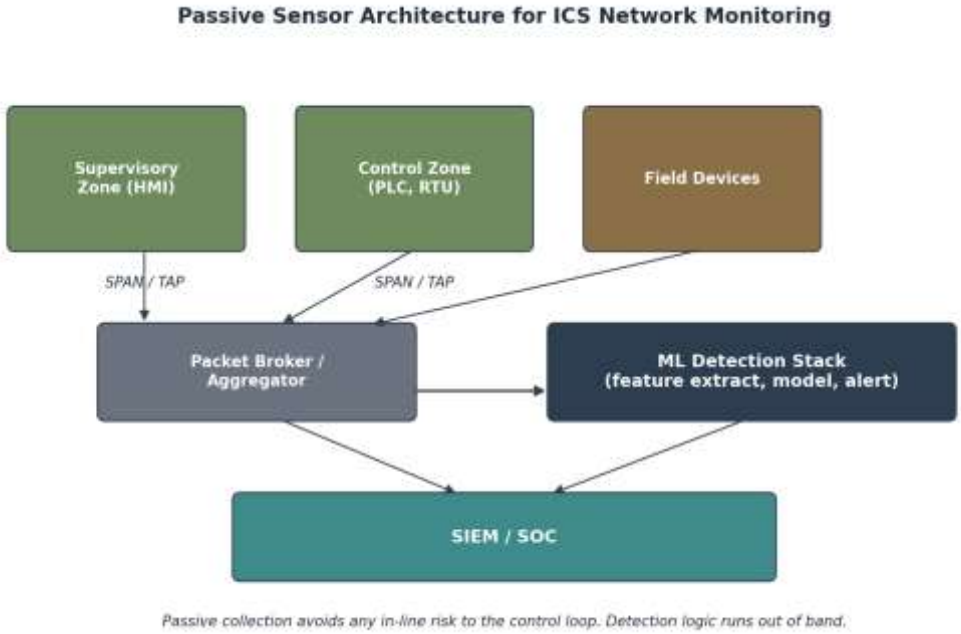


Figure 2: Passive Sensor Architecture. Detection Logic Runs Out Of Band; The Control Path is not Modified

4.2. Feature Engineering

Instead of using raw network packets and raw sensor readings, it is much better to use engineered features. Flow features, protocol field distributions, calculated rate of change statistics- these typically have a better out-of-sample accuracy and require simpler models. Here is the value that controls engineers bring to the process. A model that takes in just the generic packet capture will level out at accuracy much more quickly than a model which uses carefully chosen features that reflect the specifics of the control process.

4.3. Network-Aware vs. Process-Aware Detection

In truth, we ought to recognize there are two approaches to detection because they provide answers to different questions. I have contrasted those here. Network-aware looks at what is happening in terms of communication: what machine is talking to which, how often and using which protocols. We will generally find it effective in discovering rogue computers, hosts sniffing for vulnerable systems or protocol abuses. Process-aware detects what physical state the detected communication might suggest has been arrived at: the command that is making a system do what is described above, even though it is correct syntactically and otherwise looks benign. Most experienced detection deployers are running two types of detection, since they are not mutually exclusive and are excellent complements.

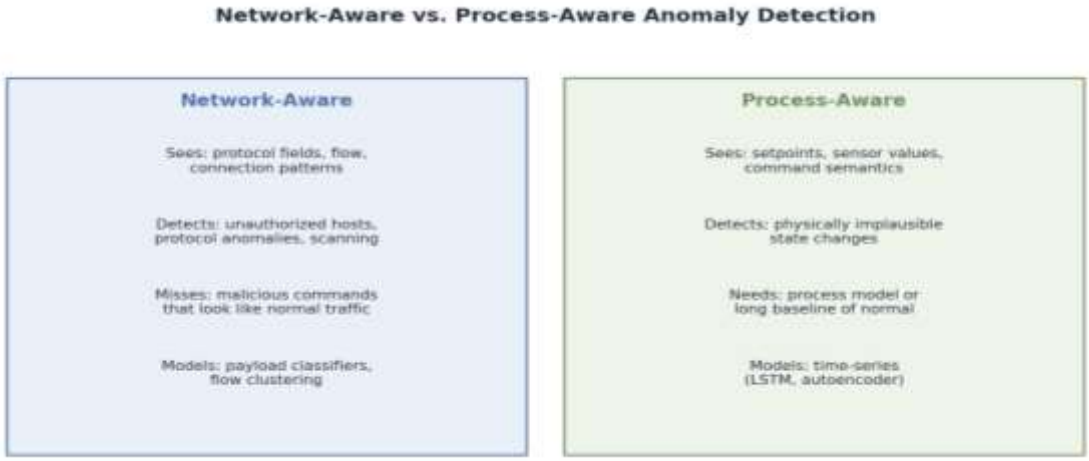


Figure 3: Network-Aware and Process-Aware Anomaly Detection Answer Different Questions. Mature Deployments Use Both

4.4. Federated and On-Premise Models

Plant operators are right to worry about sending OT data to vendors for model training and federated learning offers a potential workaround in that model updates are sent to a central server while the raw data stays within the plant walls[6]. However, FL is seldom seen in commercial practice partly due to the effort required to do it effectively and partly because the operators most paranoid about sensitive data getting stolen are often the least inclined to place trust in the person who controls the federated server. Most of the OT data security has thus remained on premise, even though it is less efficient for recognizing patterns across a large group of plants.

5. Mapping Detections to Adversary Behavior

However, anomaly scores do not really end a triage conversation, they simply start one. These conversations are made much easier when the output can be framed in a manner that everybody understands what threat actors are doing. Fortunately, in 2020, MITRE released ATT&CK for ICS, which, in addition to two annual updates in 2021 and 2022, has provided an appropriate vocabulary for threat actors within OT. [10] Each detection event that occurred should have, if possible, the most appropriate ATT&CK technique that it likely is annotated on the report output. A spike of increased engineering command volume originating from one of the OT servers might mean Modify Controller Tasking. If a non-authorized read attempt happened against your OT historian database, then Theft of Operational Information. Analysts get a hypotheses, but more importantly, it also allows detection to communicate its coverage in a method a security manager can easily relate to.

This process can also help shape the future model. One can monitor for missing coverage of the ATT&CK matrix and redirect focus onto other areas that can fill these gaps, instead of continuing to waste effort based on whatever latest benchmark came out from an arbitrary research paper. We do see several commercial OT detection products shipping with ATT&CK annotations baseline out-of-the-box, but we have moved into an era where the focus has to shift to operators properly integrating the annotations into their triage process instead of whether the annotations have been included in the system.

6. Operational Problems

A model that performs well on a benchmark dataset can still fail in production. The reasons cluster around a few themes that recur across deployments.

6.1. False Positive Economics

They live and die by their false positive rate, how this is calculated varies greatly. There is a real situation where 99% precision means one alert every hour and this is overwhelming. Focus not on stat quality, but analyst minutes. Those who do it right focus on getting the alert to a human and tuning the model down based on that human's capabilities. The rest fail and dump a massive amount of fancy demos before their analyst team revolt and walk off the job.

6.2. Model Drift

Equipment gets updated over the decades and with operational mode changes and a baseline model will naturally drift over time to the extent that a new model needs to be trained on current and relevant baseline behavior to function properly. However, you also cannot just retrain your system on incoming logs with malicious behavior baked in (that it did not flag) or you will eventually get stuck in an eternal loop of reinforcing the attack. The typical approach is to manually review any suspicious behavior, approve as a false alarm or threat, then retrain your models. Some operate on a faster-than-drift and slower-than-new-data refresh cycle.

6.3. Adversarial Examples

Well, ML models are not invincible. The work on adversarial examples started back in 2014 with Goodfellow and the boys, showing you could add minuscule amounts of noise to inputs to trick the classifier into making mistakes while being super sure of themselves[7]. For things like malware analysis, an adversary that knows the system's layout and models could create malicious commands or traffic streams that slip under the radar. The defenses are only partially effective, things like ensemble learning, where you have got multiple models checking each other and input validation can help keep things out. A major theme in network security is defense in depth-do not assume one line of defense can do everything by itself.

6.4. Training Data Scarcity

It is possible to find public datasets, but these are minimal, real data is confidential and generally non-public. Synthetic data creation may be appealing, but it appeals far more to the data-creator than the attacker creating the events. Organizations that wish to realistically measure and improve their models' performance need to build evaluation methods representative of their own operating circumstances, not merely measure themselves against vendor benchmarks, which can range in terms of meaningfulness. This can be expensive and the interest and resources devoted to it are useful predictors of effective versus merely pretty programs.

6.5. Explainability

An anomaly score itself is unhelpful for analysts. What are the key features contributing to the anomaly? How do historical trends affect interpretation of anomalies today? Which physical machines exhibit this anomaly? Techniques for explaining anomalies include feature attributions, counterfactual examples or attention maps-some are research topics while others are commercialized. While the explainability problem is not solved, the situation improved over three years. Systems should evaluate explication depth, making it a factor than an add-on.

7. Where the Field Is Heading

There are two trends going on in OT security around this area in 2023. First, LLMs are starting to have real implications across cybersecurity broadly. As ChatGPT and its ilk show astonishing capabilities to the public, a range of SOC activities, like alert prioritization, log summarization and assistance to analysts, have caught the attention of security teams. But, like any advancement, the attack side is quick to adopt LLMs as well for things like more convincing phishing campaigns, automating reconnaissance and mapping attacks and coding tasks that historically required more focused human input[8]. We have early examples in OT-specific usage patterns [8], but the trend looks pretty firm.

Second, we are now seeing maturation around specific AI frameworks, focusing on risk in OT environments. In Jan 2023, NIST published the AI Risk Management Framework [9] and government entities such as CISA are actively incorporating an AI perspective into OT guidance. If you are a consumer of an AI-driven OT security tool, expect that the tool vendor will be forced to justify how it fits within these frameworks, for regulatory purposes and for good measure.

8. Conclusion

Machine learning, although now one of the many tools available to industrial control system (ICS) intrusion detection specialists, has not displaced traditional detection approaches, they remain relevant and functional as they did for the past 20 years. ML addresses a real gap, but only up to a point. Industries that are most effective when deploying ML for detection use it in concert with multiple other detection methods, invest in time-consuming processes like feature creation and analyst

training and keep their expectations grounded despite vendor promises that currently dominate the market. AI-based detection is effective, however, realistic expectations about its limitations are also critical.

References

1. U.S. Cybersecurity and Infrastructure Security Agency, Department of Energy, Federal Bureau of Investigation, and National Security Agency, "APT Cyber Tools Targeting ICS/SCADA Devices," Joint Advisory AA22-103A, April 13, 2022.
2. J. Goh, S. Adepu, K. N. Junejo, and A. Mathur, "A Dataset to Support Research in the Design of Secure Water Treatment Systems," Proc. CRITIS, 2016.
3. S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
4. P. Malhotra, A. Ramakrishnan, G. Anand, L. Vig, P. Agarwal, and G. Shroff, "LSTM-Based Encoder-Decoder for Multi-Sensor Anomaly Detection," Proc. ICML Workshop on Anomaly Detection, 2016.
5. K. Stouffer, V. Pillitteri, S. Lightman, M. Abrams, and A. Hahn, "Guide to Industrial Control Systems (ICS) Security," NIST Special Publication 800-82 Revision 2, May 2015; Initial Public Draft of Revision 3, April 2022.
6. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," Proc. AISTATS, 2017.
7. I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," arXiv:1412.6572, 2014.
8. European Union Agency for Cybersecurity (ENISA), "Artificial Intelligence Cybersecurity Challenges," December 2020; ENISA "Threat Landscape 2022," November 2022.
9. National Institute of Standards and Technology, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," January 26, 2023.
10. Dragos, Inc., "Year in Review 2022: ICS/OT Cybersecurity," February 2023.
11. MITRE Corporation, "ATT&CK for ICS," framework documentation, ongoing.
12. National Institute of Standards and Technology, "Framework for Improving Critical Infrastructure Cybersecurity," Version 1.1, April 2018.
13. D. Yang, A. Usynin, and J. W. Hines, "Anomaly-based Intrusion Detection for SCADA Systems," Proc. NPIC-HMIT, 2006 (foundational reference).
14. Verizon, "2022 Data Breach Investigations Report," May 2022.
15. World Economic Forum, "Global Cybersecurity Outlook 2023," January 2023.