VOL. 08 NO. 9,
MAY, 2025
E-ISSN 3026-958X
P-ISSN 3027-1169

African Scholar Publications & Research International
www.africanscholarpub.com

Journal of Science Innovation & Technology Research (JSITR)

Adversarial Threats in Industrial Control Systems: A Machine Learning Approach to Securing the U.S. Energy Grid

Fatima Rilwan Ododo[1]; Ridwan Rahmat Sadiq[2]; & Nicholas Addotey[3]

[1]Computer Science, Montana State University, Bozeman, MT 59717, USA. [2]Computer Science, Nasarawa State University, Keffi, Nigeria. [3]Cyber Security, Montana State University, Bozeman, MT 9717, USA

Correspondence author: fatimaododo@montana.edu

DOI: https://doi.org/10.70382/ajsitr.v8i9.040

Abstract

As the U.S. energy sector undergoes rapid digital transformation, its reliance on industrial control systems (ICS) such as SCADA has introduced new cybersecurity vulnerabilities. While machine learning (ML) models have become essential tools for intrusion detection and predictive maintenance in these systems, they are increasingly being targeted by adversarial attacks designed to deceive or disable them. This paper presents a structured literature review that explores the intersection of adversarial machine learning and critical infrastructure security, focusing on how malicious actors can exploit ML models embedded in power grid operations. We categorize common attack types—such as evasion, poisoning, and model inversion—and demonstrate their potential impact on ICS environments. Drawing on recent developments in adversarial defense, we propose strategies to harden ML-based security systems through adversarial training, robust feature selection, and anomaly-aware architectures. The review also identifies existing research gaps and discusses policy implications. This study is limited in scope to the U.S. power sector and does not include experimental validation.

Keywords: Adversarial machine learning, industrial control systems, SCADA security, critical infrastructure protection, machine learning-based intrusion detection, evasion and poisoning attacks, resilient AI models, U.S. energy sector

Introduction

The increasing digitization of the U.S. energy sector has transformed the management of power generation, transmission, and distribution, enabling enhanced efficiency and real-time system control (Ododo and Sadiq, 2025b). Industrial Control Systems (ICS)—including Supervisory Control and Data Acquisition (SCADA) and Distributed Control Systems (DCS)—form the backbone of this infrastructure, facilitating the continuous operation of critical processes across utilities (Stouffer et al., 2011; Kalonde et al., 2024b). However, as these systems become increasingly interconnected and software-driven, they inherit significant cybersecurity vulnerabilities (Ododo and Sadiq, 2025c). Cyberattacks targeting ICS can lead to widespread blackouts, physical equipment damage, and disruptions to essential services, thereby threatening national security and societal resilience (Li, 2022).

To enhance the cyber defense capabilities of ICS environments, machine learning (ML) has been widely adopted for intrusion detection, anomaly recognition, and behavioral modeling (Ododo and Sadiq, 2025a; Garba et al., 2019). ML-based systems are adept at recognizing subtle deviations from normal operations, making them effective tools for detecting zero-day threats and advanced persistent attacks (Hodo et al., 2016; Buczak and Guven, 2015). However, the growing reliance on ML introduces a new class of vulnerabilities. Specifically, adversarial machine learning (AML)—which involves crafting deceptive inputs to mislead or manipulate ML models—has emerged as a critical threat vector (ODODO and ADDOTEY, 2025a; Kalonde et al., 2024a). These attacks can bypass intrusion detection systems, compromise training data integrity, or extract sensitive operational parameters, thereby undermining the very models intended to enhance ICS security (Biggio and Roli, 2018; Papernot et al., 2016a).

This paper aims to synthesize literature on adversarial ML threats to ICS in the U.S. energy sector, identify critical gaps in current defense mechanisms, and propose context aware policy and technical recommendations to improve infrastructure resilience. The scope of this work encompasses the architecture of ICS, the role of ML in ICS cybersecurity, the taxonomy of adversarial threats, and the evaluation of proposed defense mechanisms in practical settings.

To ensure comprehensive coverage and methodological rigor, a structured literature review was conducted. Relevant sources

were retrieved from academic databases including IEEE Xplore, arXiv, SpringerLink, and Google Scholar. Searches were guided by keywords such as "adversarial machine learning in ICS," "SCADA cybersecurity," "machine learning based intrusion detection," and "resilient AI in critical infrastructure." The review focused on publications from 2010 to 2025, capturing both foundational concepts and recent advancements. Inclusion criteria required that studies explicitly addressed adversarial threats or defense strategies in the context of industrial or cyber-physical systems—particularly those related to energy infrastructure. Articles focused solely on generic ML methods without ICS relevance were excluded. In total, approximately 75 peer-reviewed papers and technical reports were selected for detailed review based on their relevance, methodological clarity, and contribution to the cybersecurity discourse in critical infrastructure.

By highlighting dominant research trends, evaluating the applicability of defense solutions, and identifying actionable gaps, this paper contributes to the development of robust, explainable, and adversarial resilient ML systems tailored for the operational realities of ICS. The findings offer valuable insights for policymakers, utility operators, and security practitioners tasked with defending the U.S. power grid against increasingly sophisticated cyber threats.

Overview of Industrial Control Systems and Their Vulnerabilities

Industrial Control Systems (ICS) are specialized digital systems designed to monitor, control, and automate physical processes in critical infrastructure sectors such as energy, water, and manufacturing. Within the U.S. power sector, ICS components—including Supervisory Control and Data Acquisition (SCADA) systems, Programmable Logic Controllers (PLCs), and Remote Terminal Units (RTUs)—enable real-time supervision of substations, transmission lines, and generation plants (Stouffer et al., 2011). These systems operate under strict timing and reliability constraints and are often integrated with legacy technologies not originally designed with cybersecurity in mind.

Historically, ICS environments were physically isolated and relied on proprietary protocols, which provided a layer of" security through obscurity." However, the shift toward open standards (e.g., TCP/IP), remote access, and internet connectivity has increased their exposure to cyber threats. Today's ICS are often connected to corporate IT networks and cloud-based systems for monitoring, data analytics, and predictive maintenance—creating a broader attack surface for adversaries (Knowles et al., 2015).

Vulnerabilities in ICS typically arise from both technical and organizational factors. Technically, many ICS lack basic security features such as encryption, authentication, and real-time patching mechanisms. Because uptime and availability are paramount, many systems operate for years without software updates, making them susceptible to known exploits (Zhu et al., 2011). Organizationally, gaps in cybersecurity awareness, inadequate

incident response procedures, and lack of coordination between IT and OT (Operational Technology) teams further compound the risk (Conway et al., 2016).

The consequences of exploiting ICS vulnerabilities can be severe. The 2015 Ukraine power grid cyberattack, for example, involved coordinated exploitation of ICS components to cause intentional blackouts affecting hundreds of thousands of citizens (Case, 2016). More recently, the 2021 Colonial Pipeline ransomware incident, while primarily targeting corporate IT systems, resulted in the shutdown of operational infrastructure due to precautionary measures—highlighting how tightly coupled ICS and enterprise systems have become (Case, 2016).

In light of these risks, there is growing interest in leveraging machine learning for anomaly detection, early warning, and autonomous response within ICS. However, the deployment of ML models must account for the unique constraints of ICS environments—such as real-time performance, explainability, and resistance to manipulation (ODODO and ADDOTEY, 2025b). These concerns are amplified when considering the emerging class of adversarial attacks, which exploit the trust placed in ML models and can weaponize prediction errors to disrupt physical processes.

Adversarial Threat Landscape in Industrial Control Systems

As machine learning becomes more embedded in the security and operations of industrial control systems (ICS), new vulnerabilities are emerging—specifically from adversarial machine learning (AML). Adversarial threats are deliberate attempts to manipulate the inputs or training processes of machine learning (ML) models to cause misclassification, evasion of detection, or model exploitation Addotey and Ododo (2025). These threats pose unique risks to ICS environments, where ML is increasingly used for intrusion detection, predictive maintenance, and anomaly recognition in real-time control systems.

One of the most common forms of attack is the evasion attack, in which an adversary subtly modifies input data so that a malicious behavior appears normal to a trained model. In the context of an ML-based intrusion detection system (IDS) within a SCADA network, such an attack could allow unauthorized commands or traffic patterns to pass undetected (Papernot et al., 2016a). For example, modifying just a few control variables or timing signals can bypass classifiers trained to identify anomalies in power distribution metrics.

Poisoning attacks, on the other hand, target the model during the training phase. Attackers inject malicious but plausible data into the training dataset, thereby corrupting the model's decision boundaries (Biggio et al., 2012). In ICS, where models may be retrained periodically on streaming sensor data, poisoning attacks could permanently distort the ML model's perception of ''normal'' operational behavior—leading to delayed responses or false positives/negatives in critical scenarios.

Another emerging threat is model inversion and extraction attacks, where adversaries query an ML model (e.g., via APIs or through network access) to reconstruct sensitive

features of the training data or replicate the model itself (Tram`er et al., 2016). In ICS settings, this could result in leakage of proprietary operational parameters, control logic, or infrastructure blueprints—raising both industrial espionage and sabotage concerns.

The attack surface in ICS is further complicated by the operational constraints of control systems: low latency requirements, lack of computational redundancy, and the inability to rapidly update software or firmware make real-time defense challenging. Moreover, adversarial threats in this domain are particularly dangerous due to their stealth and physical impact. Unlike traditional malware, which may trigger system alerts, adversarial inputs can manipulate ML models silently, leading to subtle but cumulative damage (Huang et al., 2011; ODODO and SADIQ, 2025).

Recent studies underscore the urgency of addressing these threats. For instance, researchers have demonstrated that adversarial perturbations to sensor readings in water treatment systems can bypass anomaly detectors while maintaining operational realism (Carlini and Wagner, 2017). Similarly, in power grid simulations, even small adversarial changes to voltage or load measurements have been shown to mislead fault detection mechanisms (Bhamare et al., 2020).

Given the growing use of AI in ICS and the expanding threat landscape, it is crucial to develop defense strategies tailored to adversarial scenarios. The following section explores the range of countermeasures being investigated to build robust, resilient, and trustworthy ML models for critical infrastructure security.

Defense Strategies Against Adversarial Machine Learning in ICS

As adversarial machine learning (AML) threats to industrial control systems (ICS) continue to grow, so too does the need for targeted and effective countermeasures. These strategies must not only detect and mitigate adversarial inputs but also account for the unique operational constraints of ICS—such as strict latency requirements, limited computational redundancy, and the need for high interpretability in safety-critical settings. Accordingly, defense mechanisms can be broadly categorized into preventive, detective, and reactive approaches, with varying degrees of feasibility in real-time ICS environments.

Preventive Measures: One of the most extensively studied preventive techniques is adversarial training, which involves augmenting the training data with adversarial examples to improve model robustness (Madry et al., 2017). While this method has shown promise in traditional ML applications, its utility in ICS remains limited due to the high computational overhead and the potential reduction in accuracy on non-adversarial inputs. In latency-sensitive systems like SCADA, such trade-offs may be unacceptable unless adversarial training is implemented during offline retraining cycles rather than continuous deployment. Similarly, architectural defenses such as model distillation and ensemble

learning aim to enhance robustness by smoothing decision boundaries or aggregating classifier outputs (Strauss et al., 2017; Papernot et al., 2016b). These methods offer stronger protection against perturbations but may introduce additional complexity and reduce real-time explainability, which is a critical requirement in operator-in-the-loop environments.

Detective Measures: Detection-based defenses typically operate alongside ML models to identify adversarial activity at runtime. Techniques such as input sanitization, statistical noise filtering, and feature squeezing have been applied to flag anomalous inputs based on confidence scores, gradient sensitivity, or reconstruction error (Xu et al., 2017; Fu et al., 2023). In ICS contexts, such methods are feasible if they are computationally lightweight and capable of preserving legitimate anomalies in system telemetry. For example, feature squeezing has been piloted in SCADA lab testbeds to detect minor adversarial perturbations without disrupting normal operations. These approaches are more practical for embedded systems, especially when paired with rule-based thresholds informed by domain-specific telemetry.

Reactive Measures: Reactive defenses involve post-hoc model governance and systemic hardening to minimize the long-term impact of adversarial threats. Secure system design practices—such as access control, audit logging, and periodic retraining using validated datasets—are foundational in ICS deployments. Additionally, Explainable AI (XAI) methods have gained traction in supporting human operators to interpret model decisions, reducing the risk of blind reliance on opaque predictions (Doshi-Velez and Kim, 2017; Ododo and Addotey, 2025b). While these techniques do not directly block adversarial inputs, they enhance operational trust and support timely human intervention in response to suspicious model behavior.

Despite growing research interest, most existing defenses remain general-purpose and reactive, offering limited guidance for industrial environments. There is a pressing need for domain-specific countermeasures tailored to ICS—including physics-informed models, cyber-physical redundancy, and attack-resilient sensor fusion. Few solutions have advanced beyond simulation to deployment, though initial experiments in power grid emulators and water treatment SCADA systems demonstrate the potential of integrating adversarial detection layers without compromising real-time control performance.

Ultimately, defending ICS against AML threats requires not only robust technical mechanisms but also careful calibration of model behavior, operational feasibility, and safety assurance. A layered defense posture that blends preventive, detective, and reactive approaches—aligned with ICS constraints—is essential for ensuring long-term system resilience.

Policy and Practical Recommendations

While technical defenses play a crucial role in mitigating adversarial threats within industrial control systems (ICS), broader strategies involving policy, operational planning, and public private collaboration are essential for securing the U.S. energy grid. Given the national importance of utilities and power infrastructure, a coordinated approach that combines technological innovation with governance and regulation is necessary to address the evolving threat landscape.

First, federal agencies such as the Department of Energy (DOE), Cybersecurity and Infrastructure Security Agency (CISA), and National Institute of Standards and Technology (NIST) should prioritize the development of sector-specific guidelines for integrating adversarially robust machine learning in ICS environments. While general cybersecurity standards exist for critical infrastructure (e.g., NIST SP 800-82), these frameworks must be expanded to incorporate adversarial machine learning concerns—including threat modeling, robustness testing, and secure ML lifecycle management.

Second, utility companies and system operators should establish internal governance mechanisms for the safe deployment of AI and ML technologies. This includes model validation procedures, access controls, audit trails, and policies for continuous retraining using trusted data sources. A focus on explainability and operator trust is equally important. Machine learning systems should be designed to provide interpretable alerts and justifications to human operators, especially in safety-critical settings where human oversight remains vital (Doshi-Velez and Kim, 2017).

In addition, there is a need for cross-sector threat intelligence sharing between energy providers, cybersecurity researchers, and government agencies. Public-private partnerships, such as the Electricity Information Sharing and Analysis Center (E-ISAC), can be leveraged to disseminate information about adversarial techniques and ML-targeted attack patterns. These collaborations can accelerate the creation of benchmarks, datasets, and red-teaming exercises specific to AI systems used in ICS.

Educational and workforce development programs must also evolve to include AI security training for engineers, analysts, and infrastructure administrators. As adversarial attacks become more subtle and model-specific, ensuring that the workforce can recognize and respond to these threats is just as important as deploying technical defenses.

Lastly, investment in research and innovation should be encouraged through competitive grants and national AI security initiatives. Supporting academic and industry research into robust and physics-aware ML models tailored for ICS will strengthen the nation's long-term cyber defense posture.

The intersection of adversarial machine learning and critical infrastructure protection is not just a technical challenge—it is a strategic imperative. Policymakers, engineers, and researchers must act now to ensure that AI enhances, rather than compromises, the resilience of the nation's utilities and power grid.

Conclusion

As industrial control systems (ICS) evolve to include intelligent, data-driven capabilities, machine learning has become a valuable tool for enhancing the security and efficiency of critical infrastructure. However, the rise of adversarial machine learning has revealed a dangerous new frontier in cybersecurity—where attackers can subtly manipulate inputs to deceive ML models and compromise the safety of national utilities. In the context of the U.S. energy grid, the consequences of such attacks could be catastrophic, ranging from widespread blackouts to physical damage and public safety hazards.

This paper has provided a structured review of the adversarial threat landscape affecting ML-integrated ICS. We examined the nature of common attacks, including evasion and poisoning, and explored the unique vulnerabilities posed by the convergence of cyber and physical systems. Furthermore, we assessed current defense strategies—from adversarial training and model hardening to detection and explainability—highlighting both their promise and limitations in real-time operational environments.

Our findings underscore the urgent need for a holistic defense posture—one that includes not only robust machine learning models but also sound policy frameworks, secure system design, and workforce readiness. As the U.S. power sector continues to modernize, stakeholders must prioritize the development and deployment of adversarially resilient AI systems. Failure to do so risks undermining the very infrastructure on which modern society depends.

Securing critical infrastructure in the era of intelligent automation will require continuous innovation, cross-sector collaboration, and forward-looking regulatory leadership. By addressing the challenges of adversarial machine learning today, we can lay the foundation for a safer, more resilient, and more intelligent energy future.

References

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Aliyu Garba, Sandip Rakshit, and Fatima Rilwan. Detection and sentiment analysis of hate speech on twitter in Nigerian politics. In *Proceedings of: 2nd International Conference of the IEEE Nigeria*, page 285, 2019.

Anna L Buczak and Erhan Guven. A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications surveys & tutorials*, 18(2): 1153–1176, 2015.

Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pages 2154–2156, 2018.

Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. *arXiv preprint arXiv:1206.6389*, 2012.

Bonnie Zhu, Anthony Joseph, and Shankar Sastry. A taxonomy of cyber attacks on scada systems. In *2011 International conference on internet of things and 4th international conference on cyber, physical and social computing*, pages 380–388. IEEE, 2011.

Christine Li. Securing us critical infrastructure against cyber attacks. In *Harvard model congress*, 2022.

Defense Use Case. Analysis of the cyber attack on the ukrainian power grid. *Electricity information sharing and analysis center (E-ISAC)*, 388(1-29):3, 2016.

Deval Bhamare, Maede Zolanvari, Aiman Erbad, Raj Jain, Khaled Khan, and Nader Meskin. Cybersecurity for industrial control systems: A survey. *computers & security*, 89: 101677, 2020.

Elike Hodo, Xavier Bellekens, Andrew Hamilton, Pierre-Louis Dubouilh, Ephraim Iorkyase, Christos Tachtatzis, and Robert Atkinson. Threat analysis of iot networks using artificial neural network intrusion detection system. In *2016 International Symposium on Networks, Computers and Communications (ISNCC)*, pages 1–6. IEEE, 2016.

FATIMA ODODO and NICHOLAS ADDOTEY. Advancements and challenges in deep learning for cyber threat detection. *International Journal of Science Research and Technology*, 2025a.

FATIMA ODODO and NICHOLAS ADDOTEY. Understanding the influence of outliers on machine learning model interpretability. *International Journal of African Sustainable Development Research*, 2025b.

FATIMA ODODO and RIDWAN SADIQ. The role of power sector cybersecurity in safeguarding food systems and national stability. *International Journal of Science Research and Technology*, 07 2025. doi: 10.70382/tijsrat.v08i9.052.

Fatima Rilwan Ododo and Ridwan Rahmat Sadiq. A review of outlier detection techniques in cybersecurity: A machine learning perspective. *Journal of Science Innovation and Technology Research*, 2025c.

Fatima Rilwan Ododo and Ridwan Rahmat Sadiq. Adversarial attacks in cybersecurity: A machine learning perspective. *Journal of Science Innovation and Technology Research*, 2025a.

Fatima Rilwan Ododo and Ridwan Rahmat Sadiq. Machine learning for cybersecurity in the utilities and power sector: A practical approach to infrastructure protection. *Journal of Science Innovation and Technology Research*, 2025b.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Florian Tram`er, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction {APIs}. In *25th USENIX security symposium (USENIX Security 16)*, pages 601–618, 2016.

Gilbert Kalonde, Samuel Boateng, Lateefat Sanni, Silas Chotwe, and Fatima Ododo. Artificial intelligence and special education: The use and the integration. In *Society for Information Technology & Teacher Education International Conference*, pages 1926–1932. Association for the Advancement of Computing in Education (AACE), 2024b.

Gilbert Kalonde, Samuel Boateng, Silas Chotwe, Lateefat Sanni, and Fatima Ododo. Integration of computer science in rural math classroom: A case study. In *Society for Information Technology & Teacher Education International Conference*, pages 1757–1761. Association for the Advancement of Computing in Education (AACE), 2024a.

Keith Stouffer, Joe Falco, Karen Scarfone, et al. Guide to industrial control systems (ics) security. *NIST special publication*, 800(82):16–16, 2011.

Ling Huang, Anthony D Joseph, Blaine Nelson, Benjamin IP Rubinstein, and J Doug Tygar. Adversarial machine learning. In *Proceedings of the 4th ACM workshop on Security and artificial intelligence*, pages 43–58, 2011.

Nicholas Addotey and Fatima Rilwan Ododo. Bridging static and dynamic frontiers: A systematic review of malware detection techniques. *2025 Intermountain Engineering, Technology and Computing (IETC)*, pages 1–6, 2025.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017.

Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, pages 372–387. IEEE, 2016a.

Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE symposium on security and privacy (SP)*, pages 582–597. IEEE, 2016b.

T Conway, R Lee, and M Assante. Analysis of the cyber attack on the ukrainian power grid. *Electricity Information Sharing and Analysis Center*, 2016.

Thilo Strauss, Markus Hanselmann, Andrej Junginger, and Holger Ulmer. Ensemble methods as a defense to adversarial perturbations against deep neural networks. *arXiv preprint arXiv:1709.03423*, 2017.

Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*, 2017.

William Knowles, Daniel Prince, David Hutchison, Jules Ferdinand Pagna Disso, and Kevin Jones. A survey of cyber security management in industrial control systems. *International journal of critical infrastructure protection*, 9:52–80, 2015.

Yuqian Fu, Yu Xie, Yanwei Fu, and Yu-Gang Jiang. Styleadv: Meta style adversarial training for cross-domain few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24575–24584, 2023.