A Literature Survey on the Cyber-Physical Security of Water Treatment Systems: From Simulation to Predictive Defense

High-Fidelity Emulation of Water Treatment Processes: Simulators, Digital Twins, and Testbeds

The foundational layer for any rigorous investigation into the cybersecurity of Industrial Control Systems (ICS) is the creation of a realistic, high-fidelity environment for experimentation. Academic literature demonstrates a clear consensus that such environments are essential for developing, testing, and validating security measures against threats that target the nexus of cyber and physical operations. This section surveys the evolution of ICS testbeds, from early virtualized platforms to modern hybrid systems, and highlights the emergence of the Digital Twin (DT) as the state-of-the-art paradigm for cyber-physical security research.

The Spectrum of ICS Testbeds: From Virtualization to Physical Fidelity

The landscape of ICS testbeds encompasses a spectrum of three primary archetypes: purely virtual, purely physical, and hybrid systems.[1] Virtual testbeds, which rely on software simulation and emulation, offer a low-cost and flexible solution for reproducing network components and control logic. However, they often struggle to simulate the complex, non-linear dynamics of physical processes with high fidelity, a critical limitation when studying attacks designed to cause physical consequences.[1] At the other end of the spectrum, fully physical testbeds, composed of real industrial hardware, provide unparalleled realism in process behavior and network latencies. Their significant drawbacks include high cost, lack of

scalability, and the inherent risks associated with executing potentially destructive attack scenarios on tangible equipment.[1]

Consequently, academic and industry research has converged on hybrid testbeds as the optimal approach. These systems, often described as hardware-in-the-loop (HIL) or cyber-physical testbeds, combine physical devices with software simulations to achieve a balance between realism, safety, and cost-effectiveness.[1] The proposed project architecture, which integrates the

sintax1/scadasim physical process simulator with network protocol emulation capabilities inspired by CMU-SEI SCADASim, is a direct implementation of this hybrid philosophy.[3] This approach is strongly validated by prior work, such as the HIL Water Distribution Testbed (WDT) developed by Faramondi et al. (2021). Their testbed explicitly connects a real subsystem of tanks and pumps to a simulated one, enabling the generation of a rich dataset that captures the interplay between cyber events and their physical manifestations.[2]

The Digital Twin Paradigm in Critical Infrastructure Security

In recent years, the concept of the Digital Twin (DT) has become a central and powerful paradigm in ICS security research. A DT transcends simple simulation; it is a high-fidelity, real-time virtual model that is continuously synchronized with its physical counterpart through live data streams.[4] This persistent synchronization allows the DT to accurately mirror the physical asset's operational state, enabling advanced capabilities for monitoring, predictive analysis, and, most importantly, process-aware anomaly detection.[4]

The security application of a DT is particularly potent. By leveraging a physics-based model of the industrial process, the DT can generate an independent, real-time prediction of what the system's sensor readings *should* be under normal operating conditions. This prediction is then compared against the actual sensor data received over the SCADA network. The difference between the predicted state and the actual state, known as the "residual," serves as a highly sensitive indicator of an anomaly.[4] A significant, non-zero residual suggests that the physical process has deviated from its expected behavior or, more critically, that the sensor data being fed to the control system has been maliciously altered. This technique allows for the detection of stealthy false data injection attacks that are designed to appear plausible to traditional, threshold-based alarms but which violate the underlying physical laws of the system.[4]

The architectural design of the proposed simulation framework, while not explicitly named as such, is a practical and effective implementation of this Digital Twin security principle.[3] The plan involves using the

sintax1 physical process simulation as the source of "ground truth" data. The control logic, however, is intentionally decoupled and designed to read its inputs not from the simulation objects directly, but from a Modbus data store. This architectural choice creates an intentional "man-in-the-middle" vulnerability within the simulation itself. In this context, the sintax1 simulation functions as the high-fidelity model of the physical asset (the DT's core), while the Modbus data store represents the cyber-level data that is susceptible to manipulation. An attacker can alter the values in the Modbus store, but the ground-truth state of the simulated plant remains unchanged. Therefore, the delta between the state of the sintax1 objects (e.g., tank1.volume) and the corresponding values in the Modbus data store (e.g., holding register 40001) is functionally identical to the residual analysis performed in advanced DT-based intrusion detection frameworks. This realization elevates the project from a simple simulation exercise to the construction of a sophisticated, DT-based security testbed, aligning it with the cutting edge of academic research in the field.

Survey of Seminal Water Security Testbeds and Datasets

The academic landscape of water security is anchored by several influential testbeds that have produced benchmark datasets, enabling researchers worldwide to develop and compare intrusion detection methodologies. Among the most prominent is the Secure Water Treatment (SWaT) testbed at the Singapore University of Technology and Design. SWaT is a scaled-down but fully operational, six-stage water purification plant that has been the subject of extensive research and the source of a widely used public dataset.[5] Its companion, the Water Distribution (WADI) testbed, extends the scenario to model the challenges of a modern urban water distribution network.[6]

These testbeds are invaluable because they provide high-quality, publicly available datasets that meticulously log both network traffic (e.g., Modbus, EtherNet/IP) and physical process data (e.g., sensor readings, actuator states) under both normal operation and a wide variety of documented attack scenarios.[6] The WUSTL-IIOT-2018 dataset, generated from a different SCADA testbed, offers another critical resource, particularly for its realistic modeling of class imbalance between normal and attack data.[3] The proposed project, which aims to generate a new, custom dataset, will be situated within this existing body of work, and its results can be benchmarked against models trained on these established datasets. The following table provides a comparative analysis of these key testbeds.

Testbed/ Dataset	Testbed Type	Physical Process	Control System	Protocols	Dataset Availability	Key Attack Types in

						Dataset
SWaT	Physical	Water Purification	Allen-Bradley PLCs	EtherNet/IP, Modbus TCP	Public	FDI, Command Injection, DoS, Reconnaissance
WADI	Physical	Water Distribution	Allen-Bradley PLCs	EtherNet/IP, Modbus TCP	Public	Similar to SWaT, with distribution-specific scenarios
WDT (Faramondi et al.)	Hardware-in-the-Loop	Water Distribution	Siemens & Simulated PLCs	Modbus TCP	Public	FDI, DoS, MITM, Physical Faults (Leaks)
WUSTL-IIOT-2018	Physical	Generic Process	Allen-Bradley PLC	OPC UA, Modbus, etc.	Public	DoS, Command Injection, Reconnaissance
Proposed Project	Hybrid (Virtual + Physical Honeypot)	Water Treatment	Siemens/ Allen-Bradley (Physical), Simulated PLC (Virtual)	Modbus TCP/RTU	To be Generated	FDI, Command Injection, DoS, Reconnaissance

The Adversary's Perspective: Threat Intelligence from Honeypots and Real-World Incidents

To build an effective defense, one must first understand the offense. This section bridges the gap between simulated environments and the tangible threats facing water utilities. It begins by exploring the role of ICS honeypots as a proactive tool for gathering threat intelligence directly from adversaries. This academic approach is then grounded in reality through the analysis of seminal and modern case studies of attacks on water facilities, synthesized with macro-level trends from leading industry threat intelligence reports. This establishes the critical context for the research: the threats are real, they are evolving, and they have severe physical consequences.

ICS Honeypots: A Tool for Proactive Threat Intelligence

Honeypots are decoy computer systems intentionally deployed to be probed, attacked, and compromised, allowing security researchers to observe and analyze adversarial tactics, techniques, and procedures (TTPs) in a controlled environment.[9] In the specialized domain of ICS, honeypots are an indispensable tool for gathering intelligence on threats that specifically target Operational Technology (OT) protocols and devices, which are often opaque to traditional IT security tools.[12]

The literature distinguishes between low-interaction and high-interaction honeypots. Low-interaction honeypots, such as the widely used Conpot, emulate the network services and protocols of ICS devices (e.g., listening on Modbus TCP port 502) but do not replicate the underlying functionality or physical process.[14] They are effective for capturing widespread, automated scanning and initial probing activities. High-interaction honeypots, in contrast, provide a more realistic and engaging environment for attackers, often incorporating simulated physical processes or even real hardware components.[10] The plan to develop a physical water treatment honeypot represents a high-interaction approach, which, while more complex to implement, yields significantly higher-fidelity intelligence on targeted, hands-on-keyboard attacks.[3] Deployments of ICS honeypots consistently reveal a global landscape of reconnaissance activity, with automated scanners constantly probing for internet-exposed devices running protocols like Modbus, S7comm, and HTTP.[13] This data provides a crucial baseline of the "background radiation" of the internet that any real-world or

simulated honeypot will encounter.

Foundational Case Study: The Maroochy Shire Attack (2000)

The 2000 cyberattack on the Maroochy Shire wastewater treatment facility in Queensland, Australia, remains a seminal case study in ICS security, illustrating the devastating potential of a motivated insider with specialized knowledge.[17] Vitek Boden, a disgruntled former contractor for the company that installed the SCADA system, used a laptop, radio transmitter, and his intimate knowledge of the system to remotely seize control of 150 sewage pumping stations.[19] Over a period of three months, he repeatedly issued malicious commands that resulted in the release of approximately 800,000 liters of raw sewage into local parks, rivers, and the grounds of a hotel.[17]

The attack exploited vulnerabilities that are still prevalent in many control systems today. The primary vector was the use of unencrypted and unauthenticated radio communications, which allowed Boden to spoof legitimate control signals.[17] His commands, which included disabling pumps and silencing alarms, were implicitly trusted by the field devices. This case provides a powerful, real-world validation for focusing research on command injection attacks and the fundamental security weaknesses of legacy control protocols.[3]

A Modern Case Study: The Oldsmar Water Treatment Plant Incident (2021)

The 2021 incident at the Oldsmar, Florida, water treatment plant serves as a modern touchstone, bringing the threat to public water systems into sharp focus.[3] In this attack, an unauthorized actor gained access to the plant's control system via insecure remote access software (TeamViewer).[3] The attacker then used the Human-Machine Interface (HMI) to manipulate the process, attempting to increase the concentration of sodium hydroxide (lye) in the drinking water by a factor of over 100—from 100 parts per million to 11,100 parts per million.[3] The attempt was thwarted by an alert operator who noticed the unauthorized mouse movements and immediately reverted the setpoint.

The Oldsmar incident is a prime example of a high-impact, low-complexity attack. It did not require a sophisticated zero-day exploit but rather leveraged a common operational vulnerability—poorly secured remote access—to gain direct control of the physical process.[3] It represents a direct threat to public health and is a textbook case of the malicious command

injection attacks targeting chemical dosing systems that the proposed project plans to simulate.[3]

Synthesizing Industry Threat Intelligence (Dragos, Mandiant, Verizon)

Annual threat intelligence reports from leading cybersecurity firms provide a crucial macro-level view of the evolving threat landscape.

- **Dragos:** The Dragos OT Cybersecurity Year in Review reports consistently identify the water and wastewater sector as a significant target. The 2025 report highlights a marked increase in attacks on North American water utilities, primarily attributed to nation-state actors (specifically, the Russian group Sandworm, masquerading as the hacktivist collective CyberArmyofRussia_Reborn) and pro-Iran hacktivists.[25] These attacks often involved exploiting internet-exposed HMIs to directly manipulate operational processes, causing physical consequences such as tank overflows.[26]
- **Mandiant:** The M-Trends reports from Mandiant consistently show that the exploitation of known vulnerabilities in internet-facing devices is a dominant initial infection vector for attacks on critical infrastructure.[28] This reinforces the importance of the reconnaissance phase in the attack lifecycle and validates the simulation of network scanning and exploitation as a precursor to process manipulation.
- **Verizon:** The Data Breach Investigations Report (DBIR) provides broad statistical analysis across industries. For the Utilities sector, "System Intrusion" (often involving ransomware) is a dominant attack pattern.[31] While many ransomware attacks on OT systems cause disruption as a side effect of IT encryption, the potential for deliberate operational disruption remains a significant threat.

A striking conclusion emerges when analyzing these incidents chronologically. There is a direct, traceable lineage from the vulnerabilities exploited in the 2000 Maroochy attack to the TTPs used by nation-state actors in 2024. The core attack surface—the transmission of unauthenticated control commands over an insecure medium to manipulate a physical process—has remained fundamentally unchanged for over two decades. In the Maroochy case, an insider sent unauthorized radio commands to control pumps.[17] In the Oldsmar case, an external actor used insecure remote access to send unauthorized HMI commands to alter chemical setpoints.[3] In the 2024 attacks, state-sponsored actors compromised internet-exposed HMIs to send unauthorized commands to manipulate water flow.[26] In all three instances, spanning 24 years, the underlying control system implicitly trusted the command it received, leading to a negative physical outcome. This demonstrates that while the access methods have evolved from radio transceivers to the internet, the fundamental vulnerability in many SCADA systems remains the lack of authentication and integrity checking at the control protocol level. This validates the project's focus on simulating and detecting

attacks against an inherently insecure protocol like Modbus, as this addresses a persistent, long-standing, and actively exploited vulnerability in real-world water systems.

A Taxonomy of Cyber-Physical Attacks on Water Control Systems

To effectively simulate and defend against cyber-physical threats, a structured and formal classification of relevant attack vectors is necessary. This section synthesizes general ICS attack frameworks with specific, detailed taxonomies for the Modbus protocol, which is central to the proposed project. This creates a comprehensive threat model that can be directly implemented in the simulator and used to inform the design of the honeypot.

General ICS Attack Methodologies

Cyberattacks against ICS are rarely monolithic events; they are typically multi-stage campaigns that follow a logical progression, often modeled by frameworks like the ICS Cyber Kill Chain. A typical attack begins with an external **reconnaissance** phase, where the adversary actively or passively gathers intelligence about the target organization and its internet-facing infrastructure.[3] This can involve using public sources like Shodan to find exposed devices or actively scanning network ranges with tools like Nmap to identify live hosts, open ports, and running services.[3]

Following reconnaissance, the attacker seeks to gain initial access, often by exploiting a vulnerability in an internet-facing device or through social engineering. Once inside the network, they may engage in internal reconnaissance and lateral movement to reach the target OT environment. The final stage is the "act on objectives," where the adversary manipulates the control system to cause a physical impact.[34] The simulation of this entire chain, particularly the precursor reconnaissance activities, is essential for developing predictive security models that can forecast an attack before the final, damaging stage is executed.[3]

A Deep Dive into Modbus Protocol Attacks

The Modbus protocol, originally developed in 1979, is a de facto standard for communication in countless ICS environments due to its simplicity and open nature. However, it was designed for trusted, isolated networks and contains no native security mechanisms; it lacks authentication, authorization, and encryption, making it profoundly vulnerable to attack when exposed on modern IP networks.[36]

The academic literature provides exhaustive taxonomies of attacks against the protocol. The work by Huitsing et al. is a foundational reference, identifying 33 distinct attack types that leverage the protocol's standard functionality for malicious purposes.[37] These attacks can be grouped into categories directly relevant to the proposed simulation:

- **Reconnaissance:** An attacker can use legitimate Modbus function codes to map out a network and fingerprint devices. Sending a Report Slave ID (Function Code 17) or Read Device Identification (Function Code 43) request can reveal vendor information and device types.[40] Iterating through slave IDs and attempting to read various register addresses allows an attacker to map the memory of a PLC, revealing the structure of the control logic—a precursor to a stealthy attack.[3]
- **Manipulation (Integrity Attacks):** This is the core of False Data Injection (FDI) and Command Injection attacks. An attacker with network access can use standard protocol functions like Write Single Coil (FC 5), Write Single Register (FC 6), Write Multiple Coils (FC 15), and Write Multiple Registers (FC 16) to directly alter the state of actuators (e.g., turn a pump on) or overwrite sensor values being reported to the HMI or control logic.[3]
- **Denial of Service (DoS):** An attacker can render a PLC unresponsive through several means. This can include a simple network flood of TCP SYN packets to port 502, sending malformed Modbus packets that cause the protocol stack to crash, or using specific commands like Force Listen Only Mode that instruct the device to stop responding to legitimate requests.[3]

These vulnerabilities are so fundamental that a primary defense mechanism is the use of specialized industrial firewalls with Deep Packet Inspection (DPI). Unlike a standard firewall that only sees traffic on port 502, a DPI firewall inspects the application layer content of the Modbus packet, allowing it to enforce granular rules such as, "Allow Read Holding Registers (FC 3) from any HMI, but only allow Write Single Coil (FC 5) from the master engineering workstation".[38]

A crucial point to understand is that the vulnerabilities of the Modbus protocol are not necessarily bugs or flaws in the traditional sense; they are the protocol's features being used as designed, but by an unauthorized actor. The protocol was built on the fundamental assumption that any device able to send a command is authorized to do so. Consequently, at the protocol level, there is no distinction between a legitimate command from an HMI to start a pump and a malicious command from an attacker's laptop to do the same. The attack is not

an "exploit" that breaks the protocol; it is the legitimate use of the protocol's functionality in a context that breaks the physical process. This means the security challenge is primarily one of access control and contextual awareness. A successful detection system, therefore, must not only identify a write_coil command but also determine, based on context (source, timing, current process state), whether that command is legitimate or malicious.

Contextualizing Attacks for Water Treatment Processes

As emphasized in the project's foundational documents, high-quality cyber-physical security research requires the explicit mapping of abstract cyber events to concrete physical consequences.[3] This contextualization moves the analysis from a generic network security problem to a specific, domain-aware engineering challenge. The literature and project plans provide clear examples of this mapping:

- **False Data Injection (FDI)** on a tank's level indicator transmitter (LIT) can deceive the control logic into believing the tank is empty when it is full, leading to a **tank overflow**, loss of treated water, and potential environmental contamination. Conversely, faking a full reading when the tank is empty can cause a pump to run dry, leading to cavitation and permanent **equipment damage**.[3]
- **Denial of Service (DoS)** against the PLC controlling the chlorination process can halt the injection of disinfectant. If undetected, this could lead to inadequately treated water being distributed to the public, creating a significant **public health risk**.[3]
- **Malicious Command Injection** that forces a backwash valve on a membrane filter to open during a filtration cycle can waste enormous volumes of purified water and drastically reduce plant efficiency, leading to **operational disruption and economic loss**.[3]

The following table provides a practical threat model, synthesizing the Modbus attack taxonomy with specific, contextualized impacts on a water treatment plant.

Modbus Function Code(s)	Attack Category	Attacker's Action	Target Water Treatment Component (Example)	Potential Physical Impact
FC 5 (Write Single Coil), FC 15 (Write Multiple Coils)	Command Injection	Writes a True or False value to a coil address.	Pump control relay, automated valve actuator.	Forces a critical pump to shut down, or forces a

				valve to an incorrect state, bypassing treatment stages.
FC 6 (Write Single Register), FC 16 (Write Multiple Registers)	False Data Injection (FDI)	Overwrites a holding register with a false sensor value.	Tank level sensor register, pH sensor register.	Causes tank overflow/under flow; leads to incorrect chemical dosing, compromising water quality.
FC 1, 2, 3, 4 (Read Coils/Registers)	Reconnaissance (Memory Mapping)	Iteratively reads values from a wide range of addresses.	PLC data store (all registers and coils).	Reveals control logic parameters and system state, enabling the design of a stealthy, targeted FDI attack.
N/A (TCP SYN Flood)	Denial of Service (DoS)	Floods TCP port 502 with connection requests.	PLC network interface.	PLC becomes unresponsive to legitimate HMI commands, leading to loss of view and loss of control.

The Crux of the Matter: Generating High-Fidelity Datasets for Machine Learning

The performance of any machine learning model is fundamentally constrained by the quality and representativeness of the data on which it is trained. A primary contribution of the proposed research is the development of a systematic framework for generating a labeled, high-fidelity dataset from the simulated environment. This section surveys methodologies for data collection and feature engineering, then delves into the critical problem of class imbalance, comparing traditional mitigation techniques with the state-of-the-art generative approaches that form a core part of the project's methodology.

Data Logging and Feature Engineering in Cyber-Physical Systems

Creating a dataset suitable for cyber-physical security analysis requires the meticulous and synchronized logging of two distinct but complementary data streams.[3] The first is

physical process data, which captures the ground truth of the plant's state. This consists of time-stamped readings from all simulated sensors (e.g., tank levels, flow rates, pH) and the status of all actuators (e.g., pump on/off, valve open/closed).[3] The second is

network traffic data, captured using tools like tcpdump or Wireshark and saved into packet capture (PCAP) files. This stream records all cyber-level activity, particularly the Modbus TCP communications between control components and any attacking nodes.[3]

Raw data, however, is not directly suitable for machine learning. The process of **feature engineering** is required to extract meaningful, predictive information. This involves parsing the PCAP files to extract not only general network statistics (e.g., packet rate, byte counts) but also protocol-specific fields unique to Modbus, such as the Slave ID, Function Code, Register Address, and the data values being written.[3] The most powerful features are those that bridge the cyber and physical domains. For example, a highly predictive engineered feature can be the calculated delta between a value written to a Modbus register (the cyber event) and the actual ground-truth value of the corresponding physical sensor at that same time step (the physical state). During normal operation, this delta should be near zero; during an FDI attack, it becomes non-zero, providing a clear and immediate indicator of malicious activity.[3]

The Challenge of Class Imbalance in Security Datasets

A realistic simulation of a control system will naturally produce a dataset where data from normal operations vastly outnumbers data captured during attack scenarios. This

phenomenon, known as class imbalance, is a well-documented problem in machine learning for security applications.[3] If a model is trained on such a dataset, it can develop a strong bias towards the majority (normal) class, achieving high overall accuracy simply by always predicting "normal." This results in a model with a dangerously low detection rate for the minority (attack) class, rendering it ineffective for its primary purpose.[3]

Several established techniques exist to mitigate class imbalance. The simplest methods involve resampling the dataset, such as random over-sampling (randomly duplicating instances from the minority class) or random under-sampling (randomly removing instances from the majority class).[44] While simple, random over-sampling can lead to overfitting, as the model sees the exact same data points multiple times. A more sophisticated and widely adopted approach is the

Synthetic Minority Over-sampling Technique (SMOTE). Instead of duplicating samples, SMOTE creates new, synthetic minority samples by selecting an existing minority instance and interpolating between it and one of its nearest neighbors in the feature space.[3] This creates more diversity in the minority class and generally leads to better model performance. The Adaptive Synthetic Sampling (ADASYN) algorithm is a further refinement that generates more synthetic data for minority samples that are closer to the decision boundary, effectively forcing the model to focus on the most difficult-to-learn examples.[3]

State-of-the-Art: Generative Adversarial Networks (GANs) for Data Augmentation

While SMOTE and its variants are effective, a state-of-the-art approach for creating high-quality synthetic data involves using Generative Adversarial Networks (GANs). The proposal to use a **Conditional Tabular GAN (CTGAN)** represents a cutting-edge methodology for data synthesis in this domain.[3] A GAN framework consists of two neural networks—a Generator and a Discriminator—that are trained in an adversarial competition. The Generator learns to create realistic synthetic data, while the Discriminator learns to distinguish between real and synthetic data. The process continues until the Generator becomes so proficient that its output is statistically indistinguishable from the real data, meaning it has learned to approximate the true underlying data distribution.[3]

CTGANs are specifically architected to handle the mix of discrete and continuous variables found in the tabular data common to SCADA systems.[3] The "conditional" aspect is particularly powerful; it allows the Generator to be conditioned on a specific class label. This means the trained CTGAN can be explicitly instructed to generate new, high-fidelity data samples belonging to the underrepresented attack classes, enabling the creation of a perfectly

balanced dataset of any desired size.[3]

The choice between these augmentation techniques is not a mere implementation detail; it fundamentally dictates the type of "intelligence" the resulting machine learning model can acquire. SMOTE operates by creating synthetic samples along the linear paths connecting existing minority class examples.[46] This is effective for densifying the feature space around known attack patterns, essentially teaching a model to better recognize variations of what it has already seen. A CTGAN, by contrast, learns the entire joint probability distribution of the training data.[7] Its Generator does not simply interpolate but samples from this complex, learned distribution to create entirely novel instances. This means a well-trained CTGAN could learn the underlying

concept of a particular attack—for instance, a stealthy FDI attack characterized by a slow, linear drift in a sensor value. It could then generate new examples of this attack with different slopes, starting points, or noise profiles that were not present in the original data but are statistically consistent with the attack class. A model trained on such data is learning from a much richer, more diverse representation of the threat, potentially giving it a greater ability to generalize and detect novel or modified versions of known attacks.

To validate the quality of GAN-generated data, the "Train on Synthetic, Test on Real" (TSTR) methodology is critical. In a TSTR evaluation, a machine learning model is trained exclusively on the synthetic data and then evaluated on a held-out set of real data.[3] High performance in this test provides strong evidence that the synthetic data is a faithful and useful representation of the real-world scenarios. The planned empirical comparison of CTGAN against SMOTE using the TSTR framework is a significant methodological contribution that will yield valuable insights for the broader ICS security community.[3]

Machine Learning Frameworks for Securing Water Systems

With a high-quality, balanced dataset, the research can proceed to its analytical core: the development and evaluation of machine learning models for cyber-physical security. The literature supports a dual-pronged approach, mirroring the project's objectives: first, building classification models for real-time, reactive anomaly detection, and second, developing forecasting models for proactive, predictive security analytics.

Real-Time Intrusion and Anomaly Detection (Classification)

The primary machine learning task in SCADA security is binary or multi-class classification to distinguish between 'Normal' and 'Attack' system states in real-time.

- **Baseline and Ensemble Models:** The literature provides a strong precedent for evaluating a suite of well-established supervised learning algorithms. These include baseline models like Decision Trees, k-Nearest Neighbors (KNN), Naïve Bayes, and Support Vector Machines (SVM).[3] However, a consistent finding across numerous studies is the superior performance of ensemble methods, which combine the predictions of multiple individual models to achieve greater accuracy and robustness. In particular, **Random Forest (RF)**, an ensemble of decision trees, and **Gradient Boosting Machines (GBM)**, such as the highly efficient LightGBM and XGBoost implementations, are frequently the top-performing models on tabular SCADA datasets.[3] Their ability to capture complex non-linear interactions in the data without extensive feature engineering makes them exceptionally well-suited for this domain.
- **Temporal and Sequential Models (Deep Learning):** A key characteristic of SCADA data is its nature as a multivariate time series, where the state at one moment is highly dependent on previous states. While standard classifiers treat each time step as an independent event, Recurrent Neural Networks (RNNs) are specifically designed to process sequential data. The **Long Short-Term Memory (LSTM)** architecture, a specialized type of RNN, is particularly effective at learning long-term temporal dependencies.[3] By analyzing data in sliding windows, an LSTM can learn the normal dynamic behavior of the water treatment process. This enables it to detect sophisticated anomalies where the
sequence of events is abnormal, even if each individual sensor reading at any given moment falls within its normal range.[3]
- **Unsupervised and Zero-Day Detection:** A critical limitation of supervised models is that they can only detect the specific attacks present in their training data. To provide a defense against novel, or "zero-day," attacks, unsupervised learning methods are essential. **Autoencoders**, a type of neural network trained exclusively on normal operational data, are a prominent and effective approach.[6] The autoencoder learns to compress (encode) and then accurately reconstruct (decode) normal system behavior. When presented with anomalous data that deviates from the learned patterns of normality, the network will produce a high reconstruction error, which serves as a powerful signal for anomaly detection.[3] The combination of LSTMs with autoencoders (e.g., LSTM-Variational Autoencoder) is a particularly potent technique for unsupervised anomaly detection in time-series data from water systems.[6]

Predictive Analytics for Proactive Defense (Forecasting)

The ultimate goal of security analytics is to move beyond reactive detection ("An attack is happening now") to proactive prediction ("An attack is likely to happen soon"). This enables defenders to take mitigating action before physical consequences occur. The literature supports framing this advanced task not as a regression problem to predict the exact time of the next attack, but rather as a **time-series classification forecast**.[3]

The problem is structured as follows: given the sequence of system and network data from the last *N* time steps, predict the probability of each class label (e.g., Normal, DoS, FDI) occurring at the future time step *t+1*. The feasibility of this task is entirely contingent on the existence of detectable precursor events. Multi-stage attacks, which often begin with a reconnaissance phase before the main exploit is launched, provide exactly this kind of learnable temporal pattern.[3]

LSTM networks are the natural and most powerful choice for this forecasting task. Their internal memory cells are explicitly designed to capture long-term temporal dependencies, making them adept at identifying the subtle signatures in network traffic (e.g., a series of Nmap scans followed by specific Modbus read requests) that may signal an impending attack.[3] A successful predictive model can provide operators with a critical lead time—even just 60 seconds—to implement defensive measures, such as blocking a suspicious IP address, before the disruptive phase of the attack begins.

When selecting a model for deployment in a real-world Security Operations Center (SOC), a crucial trade-off exists between the raw predictive power of complex models and the practical requirements of interpretability and computational efficiency. Deep learning models like LSTMs may achieve the highest scores on offline datasets but are often computationally expensive and function as "black boxes," making it difficult for a human operator to understand *why* an alarm was triggered.[3] This can lead to a lack of trust and "alarm fatigue." In contrast, ensemble models like Random Forest are often only marginally less accurate but are significantly faster for real-time inference and offer greater interpretability.[3] The proposed research is well-positioned to provide valuable empirical evidence on this trade-off, evaluating models not just on F1-score but also on inference speed and, through the use of Explainable AI (XAI) techniques like SHAP, on model transparency.[3] This multi-faceted evaluation will lead to more practical and actionable conclusions about which models are truly suitable for deployment in critical infrastructure protection.

Model/Algorithm	Dataset Used	Reported F1-Score (or	Key Strengths	Key Weaknesses/Considera	Primary Reference(s

	(Example)	Accuracy)		tions	)
Random Forest	Custom Water Tank Testbed	99.9% Accuracy	High accuracy, robust to overfitting, relatively fast inference, some interpretability.	Treats data points independently, missing temporal context.	[49]
LightGBM	Custom SCADA Testbed	>99% F1-Score	Extremely fast training and inference, high accuracy, memory efficient.	Can overfit on small datasets, less interpretable than single trees.	[3]
Support Vector Machine (SVM)	Custom Water Tank Testbed	97.8% Accuracy	Effective in high-dimensional spaces, robust to outliers.	Computationally intensive to train on large datasets, sensitive to kernel choice.	[3]
LSTM	SWaT Dataset	High	Excellent for learning temporal dependencies and sequential patterns.	Computationally expensive, can be a "black box," requires more data to train.	[3]

LSTM-Autoencoder (Unsupervised)	SWaT, WADI Datasets	82.16% Anomaly Detection	Can detect novel "zero-day" attacks, does not require labeled attack data.	Performance depends heavily on threshold setting, can be sensitive to noise.	[6]

Synthesis, Gaps in the Literature, and Future Research Directions

This literature survey has traversed the key domains necessary for a comprehensive study of water treatment cybersecurity, from the foundational requirement of high-fidelity simulation to the advanced application of predictive machine learning. The synthesis of these findings reveals a clear and convergent research pipeline that defines the state-of-the-art in the field, while also highlighting existing gaps and opportunities for novel contributions.

Synthesis: The Convergence of a Cyber-Physical Security Pipeline

The body of academic work reviewed points to a de facto standard for advanced ICS security research, an end-to-end pipeline that the proposed project is structured to follow. This pipeline consists of four critical stages:

1. **High-Fidelity Emulation:** The process begins with the creation of a realistic experimental environment, with a strong trend away from purely virtual or purely physical systems toward hybrid, hardware-in-the-loop testbeds that embody the principles of the Digital Twin paradigm.
2. **Systematic Attack Injection:** This environment is then used as a platform for the systematic injection of context-aware, cyber-physical attacks to generate rich datasets that capture both network and process data.
3. **Advanced Data Augmentation:** Recognizing the inherent class imbalance of security data, state-of-the-art methodologies leverage generative models, such as Conditional Tabular GANs, to create balanced, high-fidelity, synthetic datasets for robust model training.
4. **Hybrid Machine Learning Frameworks:** Finally, these datasets are used to train and

evaluate hybrid machine learning frameworks that combine supervised models for known threat detection, unsupervised models for zero-day anomaly detection, and sequential models for both reactive and proactive analysis.

Identified Gaps and Challenges

Despite significant progress, several challenges and gaps remain in the literature:

- **Dataset Standardization and Diversity:** While benchmark datasets like SWaT have been invaluable, there remains a need for more diverse, large-scale, and continuously updated datasets from different types of water treatment and distribution systems. This lack of diversity can make it difficult to assess the generalizability of machine learning models.
- **The Zero-Day Detection Problem:** Unsupervised models are the primary defense against novel attacks, but their practical deployment is challenging. They can be sensitive to normal but previously unseen operational states (concept drift) and may generate false positives that erode operator trust.
- **Explainability and Operator Trust:** The "black box" nature of many high-performing deep learning models remains a significant barrier to their adoption in real-world control rooms. The field of Explainable AI (XAI) for SCADA security is still emerging, and more research is needed to provide operators with actionable, understandable insights into model decisions.[3]
- **The Honeypot-to-Defense Gap:** While honeypots are effective tools for gathering threat intelligence, there is a gap in research on methodologies for automatically and effectively translating the intelligence gathered from a honeypot into updated, concrete defensive rules for an IDS or firewall.

Future Research Directions and Project Contributions

The proposed project is exceptionally well-positioned to address several of these gaps and make novel contributions to the field:

- **Dataset Contribution:** The creation and public release of a new, labeled dataset from the hybrid simulator and physical honeypot would be a significant contribution, providing the research community with a new resource for model development and benchmarking.
- **Methodological Contribution:** A rigorous, empirical comparison of SMOTE and CTGAN for augmenting SCADA security data, validated using the TSTR methodology, would provide valuable evidence and guidance on the application of generative models in this

critical domain.

- **Analytical Contribution:** A detailed performance analysis that moves beyond simple accuracy metrics to include inference time, computational cost, and an XAI-based evaluation of model interpretability would yield highly practical and impactful conclusions regarding the real-world deployability of different machine learning models.
- **Predictive Contribution:** The successful demonstration of a model that can reliably forecast cyberattacks based on the detection of precursor reconnaissance activities would represent a state-of-the-art achievement in moving from a reactive to a proactive security posture for critical infrastructure.

By executing this comprehensive research plan, the project can deliver not only a functional and effective simulation and honeypot system but also valuable knowledge that advances the collective understanding of how to secure the world's most critical resource.

Works cited

1. Industrial Control Systems Testbed Survey - SPRITZ Group, accessed on September 11, 2025, https://spritz.math.unipd.it/projects/ics_survey/
2. (PDF) A Hardware-in-the-Loop Water Distribution Testbed Dataset ..., accessed on September 11, 2025, https://www.researchgate.net/publication/354255669_A_Hardware-in-the-Loop_Water_Distribution_Testbed_Dataset_for_Cyber-Physical_Security_Testing
3. Smart Water Treatment Cybersecurity Research.pdf
4. Digital Twin-Driven Intrusion Detection for Industrial SCADA: A ..., accessed on September 11, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC12390215/
5. Combined Anomaly Detection Framework for Digital Twins of Water ..., accessed on September 11, 2025, https://www.mdpi.com/2073-4441/14/7/1001
6. Lightweight Long Short-Term Memory Variational Auto-Encoder for ..., accessed on September 11, 2025, https://www.mdpi.com/1424-8220/22/8/2886
7. (PDF) Time-Series Generative Adversarial Networks for Cyber ..., accessed on September 11, 2025, https://www.researchgate.net/publication/387974602_Time-Series_Generative_Adversarial_Networks_for_Cyber-Physical_Systems
8. (PDF) Lightweight Long Short-Term Memory Variational Auto-Encoder for Multivariate Time Series Anomaly Detection in Industrial Control Systems - ResearchGate, accessed on September 11, 2025, https://www.researchgate.net/publication/359850345_Lightweight_Long_Short-Term_Memory_Variational_Auto-Encoder_for_Multivariate_Time_Series_Anomaly_Detection_in_Industrial_Control_Systems
9. Investigating Threats to ICS and SCADA Systems Via Honeypot Data Analysis and SIEM, accessed on September 11, 2025, https://www.researchgate.net/publication/382398394_Investigating_Threats_to_ICS_and_SCADA_Systems_Via_Honeypot_Data_Analysis_and_SIEM
10. Industrial Honeypots - Consensus Academic Search Engine, accessed on September 11, 2025, https://consensus.app/questions/industrial-honeypots/

11. ICSvertase: A Framework for Purpose-based Design and Classification of ICS Honeypots - Savio Sciancalepore, accessed on September 11, 2025, https://ssciancalepore.win.tue.nl/publications_pdf/2023_Kempinski_ARES.pdf
12. (PDF) Analysis of ICS and SCADA Systems Attacks Using Honeypots, accessed on September 11, 2025, https://www.researchgate.net/publication/372387888_Analysis_of_ICS_and_SCADA_Systems_Attacks_Using_Honeypots
13. Analysis of ICS and SCADA Systems Attacks Using Honeypots - MDPI, accessed on September 11, 2025, https://www.mdpi.com/1999-5903/15/7/241
14. SCADA honeypots: An in-depth analysis of Conpot | Request PDF - ResearchGate, accessed on September 11, 2025, https://www.researchgate.net/publication/310498432_SCADA_honeypots_An_in-depth_analysis_of_Conpot
15. Honeypot study of threats targeting critical infrastructure - DiVA portal, accessed on September 11, 2025, https://www.diva-portal.org/smash/get/diva2:1751352/FULLTEXT01.pdf
16. SimProcess: High Fidelity Simulation of Noisy ICS Physical Processes - arXiv, accessed on September 11, 2025, https://arxiv.org/html/2505.22638v1
17. Cyberattack On The Maroochy Shire (Australia) Wast | PDF | Computer Security - Scribd, accessed on September 11, 2025, https://www.scribd.com/document/867411150/Cyberattack-on-the-Maroochy-Shire-Australia-Wast-1
18. Malicious Control System Cyber Security Attack Case Study–Maroochy Water Services, Australia - Mitre, accessed on September 11, 2025, https://www.mitre.org/sites/default/files/pdf/08_1145.pdf
19. (PDF) Lessons Learned from the Maroochy Water Breach - ResearchGate, accessed on September 11, 2025, https://www.researchgate.net/publication/221654716_Lessons_Learned_from_the_Maroochy_Water_Breach
20. Malicious Control System Cyber Security Attack Case Study-Maroochy Water Services, Australia - DTIC, accessed on September 11, 2025, https://apps.dtic.mil/sti/html/tr/AD1107275/index.html
21. Lessons Learned from the Maroochy Water Breach., accessed on September 11, 2025, https://dl.ifip.org/db/conf/ifip11-10/cip2007/SlayM07.pdf
22. Malicious Control System Cyber Security Attack Case Study: Maroochy Water Services, Australia | MITRE, accessed on September 11, 2025, https://www.mitre.org/news-insights/publication/malicious-control-system-cyber-security-attack-case-study-maroochy-0
23. Mission Critical: CIRCIA's Regulations and the Race to Secure Critical Infrastructure, accessed on September 11, 2025, https://connectontech.bakermckenzie.com/mission-critical-circias-regulations-and-the-race-to-secure-critical-infrastructure/
24. 2022 Dragos ICS/OT Cybersecurity Year in Review – Insights on New Activity Groups, Industrial Ransomware, and ICS/OT Vulnerabilities | WaterISAC, accessed on September 11, 2025,

http://www.waterisac.org/2022-dragos-icsot-cybersecurity-year-
25. ICS/OT Cyber Resilience – Dragos' 2023 OT Cybersecurity Year in Review: Insights on New Activity Groups, Industrial Ransomware, and ICS/OT Vulnerabilities | WaterISAC, accessed on September 11, 2025, http://www.waterisac.org/icsot-cyber-resilience-%E2%80%93-dragos%E2%80%99-2023-ot-cybersecurity-year-review-insights-new-activity
26. 2025 OT Cybersecurity Report 8th Annual Year in Review - Dragos, accessed on September 11, 2025, https://www.dragos.com/ot-cybersecurity-year-in-review/
27. OT Cyber Threat Report - Waterfall Security Solutions, accessed on September 11, 2025, https://waterfall-security.com/wp-content/uploads/2025/03/2025-OT-Cyber-Security-Threat-Report.pdf?mc_cid=53c324e382&mc_eid=0069fc2d69
28. M-Trends 2025: Frontline insights for the public sector | Google Cloud Blog, accessed on September 11, 2025, https://cloud.google.com/blog/topics/public-sector/mandiant-m-trends-2025-3-key-insights-for-public-sector-agencies
29. M-Trends 2025: Data, Insights, and Recommendations From the Frontlines - Google Cloud, accessed on September 11, 2025, https://cloud.google.com/blog/topics/threat-intelligence/m-trends-2025
30. Mandiant's M-Trends 2024 Report on Targeted Attack Activity in 2023 | WaterISAC, accessed on September 11, 2025, https://www.waterisac.org/report-mandiants-m-trends-2024-report-targeted-attack-activity-2023
31. 2023 Data Breach Investigations Report - Verizon, accessed on September 11, 2025, https://www.verizon.com/business/resources/Ta5a/reports/2023-dbir-public-sector-snapshot.pdf
32. 2023 Data Breach Investigations Report - Verizon, accessed on September 11, 2025, https://www.verizon.com/business/resources/T34c/reports/2023-dbir-utilities-snapshot.pdf
33. OT Vulnerability Assessment and Penetration Testing - Secneural, accessed on September 11, 2025, https://www.secneural.com/ot-ics-scada-cybersecurity-consulting.html
34. Reconnaissance Techniques and Industrial Control System Tactics Knowledge Graph, accessed on September 11, 2025, https://www.researchgate.net/publication/371702598_Reconnaissance_Techniques_and_Industrial_Control_System_Tactics_Knowledge_Graph
35. I came, I saw, I hacked: Automated Generation of Process-independent Attacks for Industrial Control Systems - Hadjer Benkraouda, accessed on September 11, 2025, https://hbenkraouda.web.illinois.edu/wp-content/uploads/2025/08/ics-asiaccs20-1.pdf
36. A Comprehensive Security Analysis of a SCADA Protocol: From OSINT to Mitigation - SciSpace, accessed on September 11, 2025,

https://scispace.com/pdf/a-comprehensive-security-analysis-of-a-scada-protocol-from-fpqc88jm90.pdf
37. Attack taxonomies for the Modbus protocols | Request PDF, accessed on September 11, 2025, https://www.researchgate.net/publication/245478702_Attack_taxonomies_for_the_Modbus_protocols
38. Exploiting SCADA vulnerabilities using a Human Interface Device - The Science and Information (SAI) Organization, accessed on September 11, 2025, https://thesai.org/Downloads/Volume6No7/Paper_31-Exploiting_SCADA_vulnerabilities_using_a_Human_Interface_Device.pdf
39. ICSrank: A Security Assessment Framework for Industrial Control Systems (ICS) - LJMU Research Online, accessed on September 11, 2025, https://researchonline.ljmu.ac.uk/id/eprint/13480/1/2020AlhasawiPhD.pdf
40. (PDF) Dynamic Rule Generation for SCADA Intrusion Detection - ResearchGate, accessed on September 11, 2025, https://www.researchgate.net/publication/308143133_Dynamic_Rule_Generation_for_SCADA_Intrusion_Detection
41. DPI | Tofino Industrial Security Solution, accessed on September 11, 2025, https://www.tofinosecurity.com/resources/topics/dpi
42. SCADA Security & Deep Packet Inspection – Part 1 of 2, accessed on September 11, 2025, https://www.tofinosecurity.com/blog/scada-security-deep-packet-inspection-%E2%80%93-part-1
43. ON THE CHALLENGES OF ACHIEVING IEC 62443 SECURITY REQUIREMENTS IN TIME SENSITIVE INDUSTRIAL NETWORKS Co-authored by Student, accessed on September 11, 2025, https://ualberta.scholaris.ca/bitstreams/aa85d92d-a4a2-4821-96af-50198e3d90e0/download
44. Imbalance Datasets in Malware Detection: A Review of Current Solutions and Future Directions, accessed on September 11, 2025, https://thesai.org/Downloads/Volume16No1/Paper_126-Imbalance_Datasets_in_Malware_Detection.pdf
45. Intrusion Detection Model for Imbalanced Dataset Using SMOTE and Random Forest Algorithm - ResearchGate, accessed on September 11, 2025, https://www.researchgate.net/publication/356736575_Intrusion_Detection_Model_for_Imbalanced_Dataset_Using_SMOTE_and_Random_Forest_Algorithm
46. What is SMOTE? | Activeloop Glossary, accessed on September 11, 2025, https://www.activeloop.ai/resources/glossary/synthetic-minority-over-sampling-technique-smote/
47. A Conditional Tabular GAN-Enhanced Intrusion Detection System for Rare Attacks in IoT Networks - arXiv, accessed on September 11, 2025, https://arxiv.org/html/2502.06031v1
48. Conditional Tabular Generative Adversarial Based Intrusion ..., accessed on September 11, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC10301902/
49. SCAD System Testbed for Cybersecurity Research Using ... - arXiv, accessed on

September 11, 2025, https://arxiv.org/pdf/1904.00753
50. Improving SIEM for Critical SCADA Water Infrastructures Using Machine Learning - GitHub, accessed on September 11, 2025, https://github.com/AbertayMachineLearningGroup/machine-learning-SIEM-water-infrastructure
51. Improving Industrial Control System Cybersecurity with Time-Series Prediction Models, accessed on September 11, 2025, https://www.mdpi.com/2673-4591/101/1/4
52. Deep Learning Based Anomaly Detection in Water Distribution Systems - ResearchGate, accessed on September 11, 2025, https://www.researchgate.net/publication/346682571_Deep_Learning_Based_Anomaly_Detection_in_Water_Distribution_Systems
53. A Deep Learning Approach for False Data Injection Attacks Detection in Smart Water Infrastructure - CEUR-WS, accessed on September 11, 2025, https://ceur-ws.org/Vol-3962/paper24.pdf