Advanced Reactor Safeguards & Security

Artificial Intelligence for Advanced Reactor Cybersecurity Protection

A Comprehensive AI-Driven Cybersecurity Framework for Advanced Nuclear Reactor Control Systems: Experimental Validation and Multi-Paradigm Performance Analysis

Prepared for
US Department of Energy

Benjamin Blakely, Yeni Li, Akshay J. Dave, Derek Kultgen, Satyan Bhongale, Rick Vilim

Argonne National Laboratory

September 2025

ANL-25/43

Abstract

Advanced nuclear reactor systems face increasing cybersecurity threats as sophisticated attackers exploit cyber-physical interfaces to manipulate control systems while evading traditional IT security measures. This research presents a comprehensive evaluation of artificial intelligence approaches for cybersecurity protection in nuclear infrastructure, using Argonne National Laboratory's Mechanisms Engineering Test Loop (METL) as an experimental platform. We developed a systematic evaluation framework encompassing four machine learning detection paradigms: Change Point Detection, LSTM-based Anomaly Detection, Dependency Violation analysis, and Autoencoder reconstruction methods. Our comprehensive attack taxonomy includes 12 distinct scenarios targeting reactor control systems, from gradual sensor drift to sophisticated coordinated attacks, each implemented across five severity tiers to evaluate detection performance under varying attack intensities. The experimental evaluation encompassed 243 rigorous experiments across all paradigm-scenario-tier combinations using realistic METL operational data. Change Point Detection emerged as the leading approach with mean AUC performance of 0.785, followed by LSTM Anomaly Detection (0.636), Dependency Violation (0.621), and Autoencoder methods (0.580). Attack detectability varied significantly, with multi-site coordinated attacks proving most detectable (AUC = 0.739) while precision trust decay attacks presented the greatest detection challenge (AUC = 0.592). We complemented data-driven approaches with physics-based detection using PRO-AID, which leverages conservation laws and analytical redundancy relations to provide explainable diagnostics grounded in immutable physical principles. This work delivers a practical reference architecture, open-source implementation, and comprehensive performance benchmarks that advance AI-based cybersecurity capabilities for critical nuclear infrastructure, providing essential foundations for operational deployment and enhanced threat response in cyber-physical systems.

This project was funded by the U.S. Department of Energy Office of Nuclear Energy, and was partially supported by the U.S. Department of Energy, Office of Science under DOE contract number DE-AC02-06CH11357. The submitted manuscript has been created by UChicago Argonne, LLC, operator of Argonne National Laboratory. Argonne, a DOE Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357. The U.S. Government retains for itself, and others acting on its behalf, a paid-up nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government.

Contents

1 Executive Summary

This report presents comprehensive research on artificial intelligence (AI) applications for cybersecurity protection in advanced nuclear reactor systems, conducted at Argonne National Laboratory using the Mechanisms Engineering Test Loop (METL) as an experimental platform. Our work addresses the critical vulnerability of cyber-physical systems in nuclear infrastructure, where sophisticated attackers increasingly exploit digital control systems to manipulate physical processes while evading traditional IT security measures.

Experimental Platform and Infrastructure: METL serves as an advanced experimental facility designed to replicate key operational characteristics of sodium-cooled fast reactors (SFRs). The facility features integrated test vessels operating at temperatures up to 1,200°F, electromagnetic pumps, and comprehensive instrumentation and control (I&C) systems including Emerson/NI cRIO devices and industrial controllers. This sophisticated architecture provides authentic thermal-hydraulic conditions and sensor networks essential for developing AI-driven anomaly detection and cyber-physical security models. METL's segmented OT and IT networks, combined with detailed temperature profiles, flow measurements, and sodium handling procedures, enable realistic simulation of cyber attack scenarios targeting critical reactor operations.

Our research methodology evolved significantly during the project timeline. Initially, we developed a streaming data pipeline using InfluxDB and MQTT protocols to enable real-time attack simulation and detection. However, performance limitations and scalability challenges led to a strategic pivot toward offline Parquet-based data processing in the second half of the project. The final architecture implements a comprehensive transformation framework capable of generating multiple attack scenarios through configurable data manipulations while preserving the temporal and statistical characteristics essential for machine learning analysis.

We developed a systematic attack taxonomy encompassing 12 distinct attack scenarios targeting various components of reactor control systems. These scenarios range from gradual sensor drift and oscillatory manipulations to sophisticated coordinated attacks spanning multiple facilities. Each attack scenario is implemented across five severity tiers to evaluate detection performance under varying attack intensities. Our taxonomy specifically focuses on attacks affecting the cold trap purification system within METL, ensuring relevance to critical reactor safety functions while providing sufficient diversity to challenge multiple detection paradigms.

Machine Learning Evaluation Framework: Our evaluation framework implements four complementary detection paradigms: Change Point Detection using statistical baseline

learning, LSTM-based Anomaly Detection with per-sensor modeling, Dependency Violation analysis through correlation matrix monitoring, and Autoencoder reconstruction error analysis. Our comprehensive evaluation encompassed 243 experiments across all paradigm-scenario-tier combinations, revealing significant insights into attack detectability and paradigm effectiveness. Change Point Detection emerged as the leading approach with mean AUC performance of 0.785, followed by LSTM Anomaly Detection (0.636), Dependency Violation (0.621), and Autoencoder reconstruction (0.580). Attack detectability varied dramatically, with multi-site coordinated attacks proving most detectable (AUC = 0.739) and precision trust decay attacks presenting the greatest challenge (AUC = 0.592). Severity tier analysis revealed limited correlation between attack intensity and detectability across most scenarios, with only cross-facility data transplant and sequential valve manipulation showing meaningful tier sensitivity. These findings demonstrate the critical importance of paradigm-attack matching for operational deployment and suggest that ensemble approaches combining complementary detection methods could enhance overall coverage.

Physics-Based Detection Integration: To address limitations of purely data-driven approaches, we evaluated PRO-AID (Parameter-Free Reasoning Operator for Automated Identification and Diagnosis), a physics-based diagnostic platform that leverages immutable physical laws for cybersecurity applications. PRO-AID implements three integrated mechanisms: virtual sensors for extending observational coverage without additional hardware, analytical redundancy relations (ARRs) derived from conservation laws that provide mathematical ground truth, and probabilistic fault mapping using Bayesian reasoning for explainable diagnostics. This approach enables regime classification distinguishing between physical, mixed, and unphysical system behaviors, providing complementary capabilities to machine learning methods by encoding invariants that adversaries cannot easily circumvent.

Key Contributions and Deliverables: This work delivers a practical reference architecture and end-to-end evaluation stack for AI-based anomaly detection in advanced reactor operational technology environments. Key deliverables include code repositories for data proxy systems, transformation frameworks, and ML experiments; reproducible experiment configurations with comprehensive logs and aggregated metrics; and detailed architectural documentation with data-flow specifications. Our standardized preprocessing methodology, justified downsampling strategy, and results aggregation workflow with uncertainty estimates provide a foundation for future research and operational deployment.

2 Introduction

The increasing complexity and interconnected nature of advanced nuclear reactor systems necessitate robust cybersecurity measures to protect critical infrastructure from sophisticated cyber threats. As these systems integrate sophisticated instrumentation and control architectures, they become vulnerable to a range of cyber attacks that can compromise operational integrity and safety. This report presents a overview of our ongoing research at Argonne National Laboratory, focusing on the application of artificial intelligence (AI) for cybersecurity protection in advanced reactor environments.

The Mechanisms Engineering Test Loop (METL) serves as a pivotal experimental facility for this research, offering a controlled environment to simulate conditions prevalent in sodium-cooled fast reactors (SFRs). METL's advanced instrumentation, sensor networks, and integrated control systems not only emulate thermal-hydraulic and operational parameters typical of full-scale reactors but also provide high-fidelity data streams essential for developing AI-driven anomaly detection and cyber-physical security models. Enhanced by detailed temperature profiles, flow rate measurements, and comprehensive sodium handling procedures, METL allows researchers to evaluate both the cyber and physical resilience of reactor operations.

Our research methodology involves transforming METL data to simulate cyber attack scenarios that impact both the operational technology (OT) and information technology (IT) layers. In the first half of the year, we tested an in-stream (InfluxDB/MQTT) approach; in they second half of the year, we executed transformations via offline replay on Parquet. The experimental pipeline is designed to test detection models against sophisticated attack vectors, including data interception, signal injection, and command spoofing.

3 The Argonne Mechanisms Engineering Test Loop

The Mechanisms Engineering Test Loop (METL) at Argonne National Laboratory is an advanced experimental facility designed to support the development and testing of components and systems pertinent to sodium-cooled fast reactors (SFRs). METL is engineered to replicate key aspects of reactor operation by circulating purified R-grade sodium under controlled conditions, thereby providing an authentic platform to study the thermal-hydraulic and cyber-physical interactions encountered in operational nuclear facilities.

3.1 Facility Overview and Capabilities

METL comprises an integrated array of test vessels, a primary sodium loop, and extensive support systems. The facility features both 18-inch and 28-inch diameter test vessels engineered to operate at temperatures reaching up to 1,000°F and 1,200°F respectively, with precise flow control achieved through electromagnetic (EM) pumps and high-resolution flow meters. The sodium, maintained in a molten state, is continuously purified via a cold trap system designed to remove oxide and other impurities, ensuring a consistent experimental environment. An expansion tank manages thermal expansion while a dedicated dump tank facilitates safe drainage and sodium recovery operations. These features, combined with rigorous sodium handling protocols, ensure that experiments mimic the stringent conditions of advanced reactor systems.

3.2 Instrumentation and Control Systems

METL's instrumentation and control (I&C) architecture is at the core of its operational excellence, integrating controllers and sensor networks to facilitate high-speed data acquisition and robust process control. The control system leverages Emerson/NI cRIO devices[1], which combine a real-time processor with a re-configurable FPGA, offering deterministic control and high-speed data processing essential for both thermal regulation and cybersecurity monitoring. Complementing the cRIO are Emerson/NI industrial controllers (ICs), optimized for automated data acquisition in extreme environments. These controllers controllers support the requirements of both reactor simulation and cyber defense applications.

Advanced metering solutions are provided by Schneider Electric PowerLogic meters, which enable comprehensive monitoring of electrical parameters such as voltage, current, power, energy consumption, and total harmonic distortion (THD) across up to 28 three-phase circuits. These meters not only support detailed energy management but also supply critical data for anomaly detection algorithms in cybersecurity applications. The overall I&C network, secured via VLANs and segregated into dedicated OT and IT subnets, facilitates real-time data transmission and robust system monitoring—features that are vital for the simulation of cyber-physical threat scenarios.

[1]Model numbers/datasheet links redacted per Argonne Classification Office requirement and available upon request

3.3 Emulating Advanced Reactor Environments

METL is specifically designed to emulate key operational characteristics of a sodium-cooled fast reactor's (SFR) intermediate heat transport system. The use of liquid sodium as the primary coolant replicates the thermal properties and chemical interactions found in operational reactors, enabling realistic testing of material compatibility and thermal stresses. Although METL operates at a reduced scale—with lower thermal power levels and sodium inventory—it reproduces essential phenomena such as thermal stratification, flow-induced vibrations, and sodium oxidation under controlled conditions. These conditions provide a reliable basis for testing AI-based monitoring systems.

Distinct from full-scale reactor facilities, METL does not include a nuclear core or associated radiological hazards. This absence allows researchers to focus exclusively on the cyber-physical dynamics and the performance of sensor and control systems under high-temperature and high-flow conditions, without the complexities of neutron flux monitoring and radiation shielding. Furthermore, METL incorporates a wide range of sensors—including sodium level sensors, pressure transducers, thermocouples, and advanced optical fiber temperature sensors—that deliver continuous data streams. These sensor networks allow for real-time anomaly detection and predictive maintenance experimental use cases in a manner analogous to operational reactor environments.

3.4 Comparative Analysis: METL vs. Operational Sodium Reactors

While METL successfully mirrors many aspects of operational sodium reactors, several key differences must be acknowledged. METL operates on a smaller scale, with reduced sodium volume and thermal output, resulting in proportionally lower flow rates and simplified pressure conditions (typically up to 5 psig under an inert argon cover gas). Unlike full-scale reactors, METL does not incorporate nuclear reactions or associated safety systems such as neutron moderators, control rod assemblies, or complex fuel handling mechanisms. Nevertheless, the I&C architecture, sensor networks, and data acquisition systems in METL provide a realistic proxy for the OT environments found in operational reactors. This fidelity is critical for developing AI models and cybersecurity measures that are directly applicable to advanced reactor systems.

3.5 Relevance to Cyber Security and AI Research

The integration of advanced control systems and sensor networks in METL renders it an ideal platform for cybersecurity research. The facility's I&C systems, which include high-performance controllers like the cRIO and IC, offer robust, real-time processing capabilities

necessary for detecting and mitigating cyber threats. Detailed electrical metering provided by the Schneider meters further supports data-driven analysis and enhances the reliability of energy management systems. The secure network architecture—comprising dedicated VLANs, virtual machines, and redundant communication channels—ensures that data integrity is maintained even under simulated cyber attack scenarios.

By capturing high-fidelity data streams from diverse sensors, METL enables the application of advanced AI and machine learning techniques to identify anomalies, predict system failures, and enhance operator situational awareness. These capabilities are critical for addressing cyber-physical security challenges in modern reactor systems. Moreover, the controlled yet realistic environment of METL facilitates experimentation with novel cybersecurity measures, such as real-time intrusion detection, self-calibration of sensor networks, and automated anomaly remediation protocols. The insights gained from METL experiments directly inform the development of defense strategies applicable to full-scale sodium-cooled reactors.

4 Data Pipeline

4.1 METL Data Characterization

The METL facility produces two primary types of data: sensor readings and setpoint parameters. All data is stored in the central METL data historian, which serves as the authoritative repository for both real-time and historical information. This data can be accessed from the historian in both batch and real-time modes using different approaches. Batch historical data is stored in `tdms` format and accessible via a Python API, while real-time data is accessible via HTTP GET requests to the METL data server. Additionally, the METL data server provides distinct buses for accessing sensor (read-only or RO) and setpoint (read-write or RW) data independently.

The METL dataset we persist for analysis uses a flat, long-format schema with one record per observation. Each row contains: `timestamp` (timezone-aware), `bus_type` (RO/RW), `controller`, `module`, `datatype`, `group`, `name`, `meta_type`, `value` (stored as canonicalized string), and optional `extra_params` (RW-only suffix components). For storage and query efficiency, the fetcher adds partition columns `year`, `month`, and `day` (all strings) and writes Parquet partitioned by `year/month/day/bus_type`. Figure 1 summarizes the columns.

```
{
   "type": "object",
   "properties": {
      "timestamp": {"type": "string", "format": "date-time", "
         description": "TZ-aware" },
      "bus_type": {"type": "string", "enum": ["RO", "RW"] },
      "controller": {"type": "string" },
      "module": {"type": "string" },
      "datatype": {"type": "string" },
      "group": {"type": "string" },
      "name": {"type": "string" },
      "meta_type": {"type": "string" },
      "value": {"type": "string", "description": "Canonicalized;
         numeric/logical values serialized to string" },
      "extra_params": {"type": "string" },
      "day": {"type": "string" },
      "month": {"type": "string" },
      "year": {"type": "string" }
   },
   "required": [
      "timestamp", "bus_type", "controller", "module",
      "datatype", "group", "name", "meta_type", "value"
   ]
}
```

Fig. 1. METL Data Schema

4.2 METL Data Fetcher Implementation

To efficiently collect and process data from the METL facility, we developed the METL Data Fetcher, an asynchronous Python service that interfaces with the METL Web Services API. The fetcher implements asynchronous data collection through concurrent fetching of both read-only (RO) and read-write (RW) data buses using `aiohttp` for efficient HTTP requests.

The system was designed to support two different polling modes: background polling and aggressive polling. In background polling mode, the fetcher operates at a lower frequency (typically every 5 minutes) to maintain a consistent historical record while minimizing system load. The aggressive polling mode enables more frequent data collection (as often as every second) when specific conditions in the returned data are met. The Fetcher maintains a ring buffer of ten minutes of data at the "aggressive" rate, and includes this "pre-activation" window data at the aggressive polling rate once triggered. This dual-mode approach allows us to balance system resource usage during normal

operation while having the ability to capture rapid changes during critical experimental phases. The fetcher provides resilient operation through robust error handling with retry logic and structured logging for operational monitoring. It authenticates to the METL Web Services using API keys over HTTPS. In our testing, we observed that our polling may fall behind by 300-400ms/refresh when polling at the aggressive rate (currently set to 1Hz), but this is still much higher resolution than our background polling rate. This appears to be a bottleneck in the Fetcher code, not the METL API, so could likely be remedied with additional optimization or a higher-performance system running the Fetcher code.

4.3 Experimental Testing Pipeline Prototype

For operational analytics and cybersecurity research, in the first half of the year we developed an experimental pipeline as shown in Fig. 2. This system was intended to replicate a realistic messaging protocol, and enabled us to simulate various cyber attack scenarios and implement multiple virtual facilities for comparative analysis without affecting the production METL installation. However, we did not use the InfluxDB/MQTT path for final experimentation. Querying InfluxDB at scale proved too slow for our analysis workloads, and the transformation framework could not run end-to-end in real time. We therefore executed transformations offline on Parquet inputs and fed the ML experiments directly from those outputs. This means real-time simulation is not currently supported without changes to the storage/query path and transformation runtime.

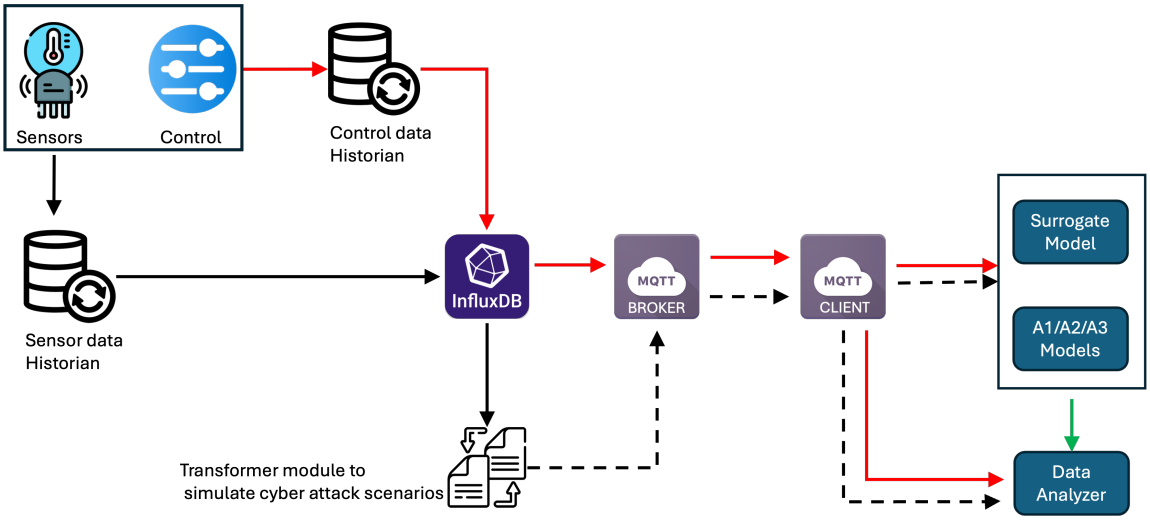


Fig. 2. Data pipeline for experimentation (H1 prototype using InfluxDB/MQTT; not used in final experiments).

The experimental pipeline consisted of several key components arranged in a layered architecture. At its foundation was InfluxDB, a time-series database optimized for storing sensor and control historical data from METL. The Data Proxy with Transformation Framework applied configurable transformations to subsets of the data, capable of generating multiple sets of time-shifted data to mimic multiple site operations and various data anomaly scenarios. An MQTT Broker acted as the publish-subscribe messaging system for distributing transformed data to consumers, while MQTT Clients subscribed to topics containing transformed data for consumption by downstream analytics. The system also included visualization and monitoring capabilities through Grafana dashboards connected to both InfluxDB and Prometheus for real-time monitoring of data flows and system performance.

The data flow began with the METL fetcher collecting data from the MIDAS system and storing it in Parquet files. The Historian component then read these files and made them available to the Transform component. The Transform component applied various transformations to the data according to the specified simulation configuration and published the transformed data to the MQTT broker. Clients could then subscribe to specific topics to receive the transformed data streams. Additionally, the system included a metrics collection path through Telegraf, Prometheus, and Grafana for monitoring and visualization.

4.4 Comparison to Operational OT Systems

Our experimental pipeline replicated many aspects of operational OT systems commonly found in industrial control environments, including those used in nuclear facilities. The similarities to operational OT systems included the use of the standard industrial protocol MQTT for data transmission, time-series data storage and management, hierarchical sensor organization, separation of read-only and read-write data buses, and real-time monitoring and alerting capabilities.

There are some differences from operational OT systems. Our system lacks the strict air-gapping often found in critical infrastructure and has a higher reliance on IP-based networking versus proprietary fieldbus technologies. There are no Safety Instrumented Systems (SIS; dedicated safety shutdown functions) or regulatory compliance mechanisms, and the security architecture is simplified compared to defense-in-depth approaches in operational environments.

These differences, however, are not material to our cybersecurity analysis for several reasons. The underlying data structures, communication patterns, and control logics remain authentic to operational systems. Modern OT environments increasingly adopt IP-based technologies, making our approach forward-compatible. The attack vectors and defensive

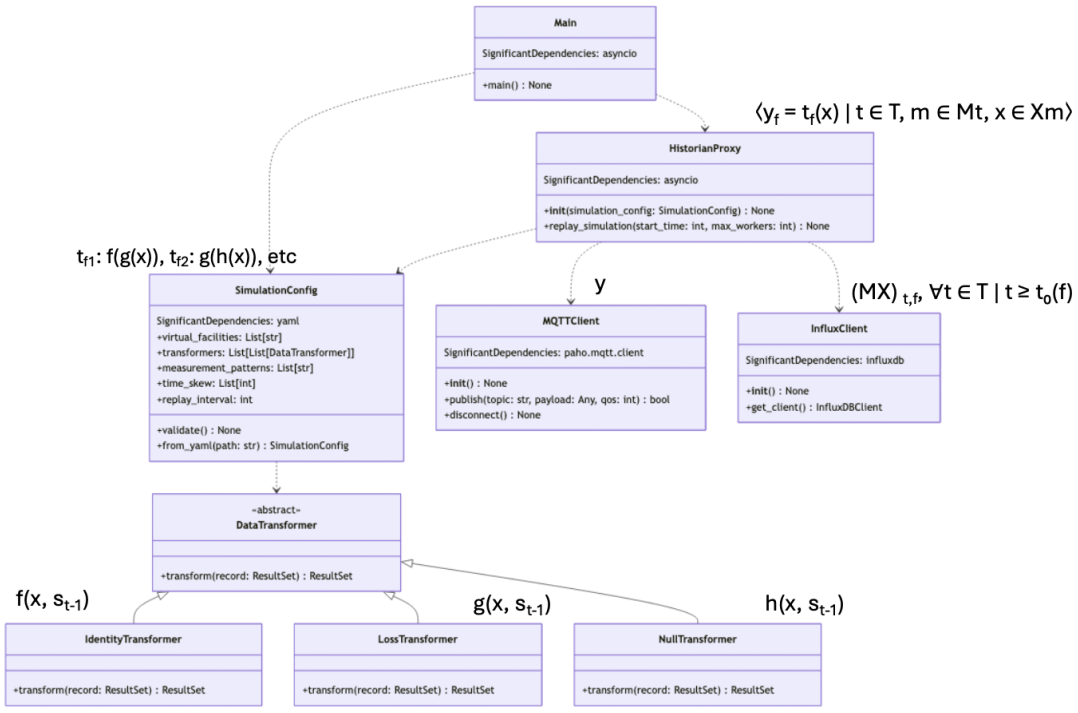


Fig. 3. Class diagram showing the transformation framework

mechanisms we study are applicable regardless of the specific implementation details, and our ability to simulate multiple virtual facilities allows comparison of different security architectures. Future work will include incorporation of digital twins to allow for realistic system responses to injected perturbations.

4.5 Experimental Capabilities

Despite the limitations of the MQTT approach, the core of our experimental pipeline is the transformation framework implemented in the Data Proxy. This framework enables us to simulate various cyber attack scenarios by applying configurable transformations to the sensor data streams. The transformation framework allows researchers to create complex, multi-stage attack scenarios by chaining transformers together, as shown in the class diagram in Fig. 3. Researchers can apply different transformations to different virtual facilities, introduce time skew between virtual facilities to simulate timing attacks, and configure the severity and timing of simulated attacks. This flexibility provides a powerful platform for comprehensive cybersecurity testing.

This experimental setup provides several key capabilities for cybersecurity research. It enables offline replay of historical data to simulate operational scenarios without affecting production systems. The system supports generation of synthetic attack scenarios and creation of multiple virtual facilities from a single data source. This facilitates comparison of different detection and defense strategies and testing of machine learning anomaly detection models against sophisticated attacks. Real-time simulation at scale is not currently supported without substantial changes to the storage/query layer and streaming transformation runtime.

5 Cyber Attack Description

The details of components used in the I&C systems of METL has been described in 3.2. In general, the I&C systems can be divided into four major parts [25]: (1) physical process, (2) sensors, (3) controllers, and (4) actuators, along with the communication pathways connecting these components. In complex systems, such as METL system, controllers are often connected to an HMI or monitoring computers via a local network. In this context, the physical process refers to the purification operation of the cold trap within the METL (Mechanisms Engineering Test Loop) system.

To prioritize tasks and optimize resource allocation, this work focuses on cyber attacks that could affect the critical function of the cold trap, specifically the purification of sodium. As aforementioned, ensuring the efficient operation of METL requires monitoring the impurity levels of oxygen and hydrogen in the liquid sodium coolant and maintaining the sodium in its liquid state. The purification process can be disrupted by insufficient cooling, which prevents the crystallization and subsequent capture of sodium oxide or hydride by the cold trap, or by excessive cooling, which can lead to the solidification of sodium within the loop.

5.1 Attack Taxonomy

Here we briefly categorize various types of Operational Technology (OT) cyber-attacks based on their methods [8], targeted components, definitions , visibility to detection mechanisms, attack scenarios, and affected critical functions of the cold trap in METL. This taxonomy not only provides an overview of OT cyber-attacks but also serves as an interface between the cyber realm and the engineering domain. The attacks are described below and summarized in Table 1.

(1) Denial of Service (DoS) Attacks

In this type of attack, the attacker seeks to disrupt the availability of a system, service, or process by overloading it with a high volume of queries or data packets. This flood of traffic depletes system resources, limiting access for legitimate users and resulting in a denial of service. DoS attacks often target important components of I&C systems, such as Programmable Logic Controllers (PLCs), HMIs, and Distributed Control Systems (DCSs) [11].

- Volumetric Denial of service (VDoS):
  A VDoS attack originates from a single source, where the attacker overwhelms the target system by sending a large volume of traffic or resource-intensive requests. Typically, the attacker uses one computer or a small group of computers to carry out the attack.[2].
- Distributed Denial of Service (DDoS) :
  A DDoS attack originates from multiple sources, typically involving a large number of compromised computers, known as botnets, that are coordinated to carry out the attack. Each botnet machine generates a surge of traffic or requests directed at the target system simultaneously, intensifying the overall impact. Because of its distributed nature, defending against a DDoS attack is more challenging compared to a VDoS attack [17].

(2) Integrity Attacks (IntA) In integrity attacks, attackers gain access to sensor, controller, actuators or communication channels. This broad attack surface enables them to alter data and processes, affecting the accuracy and reliability of the generated information. In our case, an IntA can be executed without requiring deep domain expertise or prior system knowledge, making it quick and straightforward to deploy. As a result, IntA has been extensively studied due to its ease of execution and broad attack surface.

- Simple Attacks
  The simple attacks involve minor modifications to compromise the integrity of systems. For example, an attacker may alter thermocouple readings and send inaccurate data to the control system by scaling, freezing, or adding bias to the sensor outputs [5]. Consequently, the actual temperatures deviate from the perceived values, potentially leading to equipment malfunctions, product defects, or safety hazards.
- Covert Attacks
  Covert attacks are designed to remain hidden and undetected while carrying out malicious activities within a target system or network. A representative covert attacks within IntA is replay attack, which replays captured historical

data to bypass traditional security measures such as classical χ^2 detector [18]. Since the data are genuinely from the system, it allow attackers to exploit systems even when encryption is in place [4].

(3) Control-Theoretic attacks

- Perturbation-based Attacks
  These attacks introduce external disturbances that can compromise the stability, accuracy, or safety of controlled processes[9]. They are knowledge-driven, relying on sector-specific analysis that provide insights into the system's characteristics. The effectiveness of these attacks depends on the attacker's prior knowledge of the sector, process, or system accessibility, inherently limiting their scope to specific domains and making them highly specialized. Literature indicates that these attacks often target sensors or actuators, using either stable or oscillatory perturbations tailored to the unique features of the target components [18]. Oscillatory perturbations are usually undesirable in operations, as they introduce larger noise into the system, causing frequent deviations and accelerating component degradation.
- Information Theoretic Attacks (ITAs)
  The ITAs exploit information leakage to indirectly infer system behavior and manipulate it without direct intervention. Attackers do not actively inject commands or modify system logic. Instead, they employ statistical analysis, data monitoring, or inference techniques to understand how the system operates over time [20]. By analyzing outputs like network traffic patterns, control signals, or sensor data, they gather intelligence on the system's functionality. Once sufficient information is obtained, attackers manipulate inputs or conditions to induce indirect failures.
  For example, in a smart grid attack, an attacker monitors voltage fluctuations to learn how the grid responds to changes in demand. By triggering specific electrical loads at precise moments, they can destabilize the grid without altering the control logic. These attacks are subtle and hard to detect since no direct system modification occurs. ITAs allow attackers to predict and influence control system behavior without triggering alarms and are often used as a preliminary phase before launching more direct attacks.
  The impact of information-theoretic attacks can be significant yet subtle, making them hard to detect since no direct system modification occurs. They allow attackers to predict and manipulate control system behavior without triggering

alarms. ITAs are often used as a preliminary attack phase before launching more direct attacks.

(4) Injection Attacks
In this type of attack, an attacker attempts to inject malicious code or made-up data into a system and execute it in a way that compromises its integrity, granting unauthorized access and control. By supplying false or harmful inputs to a program, the attacker can manipulate system behavior. These attacks are particularly dangerous, as they can lead to severe or even catastrophic consequences. Research indicates that the malicious inputs are often introduced in various forms, including false data, time delays, and control logic, specifically targeting I&C systems.

- False Data Injection (FDI)
  The FDI attacks target on falsifying the transmitted data while these attacks may lead to cascading influences, which are often identified as a major threat towards I&C systems [3]. The FDI attacks are launched by attackers who has access to sensors or controllers and have the intimate knowledge of the system, such as the system model. Then the FDI attack is launched by [19]
- Time Delay Injection
  In these attacks, a malicious actor introduces extra time delays into multiple communication channels of I&C systems. For instance, an adversary may inject additional delays into both the feedback and forward communication channels of the I&C systems, which may lead to the complete failure of the I&C systems. In [12], a gradual increase in time delays can push the control system into an unstable state.
- Control Logic Injection Control logic injection attacks are a type of cyber attack targeting PLCs. In these attacks, attackers manipulate or inject malicious code into the control logic of PLCs, aiming to disrupt the physical processes, which may lead to equipment damage and operational downtime [26].

(5) Stealthy Attacks
Stealthy Attacks actively manipulate system behavior while avoiding detection by security measures. In these attacks, the attacker directly modifies sensor readings, control signals, or process variables, deliberately bypassing anomaly detection tools and monitoring systems. Rather than making abrupt modifications, stealthy attacks are executed gradually over an extended period, introducing delicate changes. This approach requires specific domain knowledge to estimate and apply the delicate changes introduced to the system, allowing the attacker to stay below detection thresholds and disguise malicious actions as normal system behaviors [14]. The

impact of stealthy attacks can be significant due to their subtle and hard-to-detect nature. Since the malicious activities resemble normal system behavior, they often go unnoticed. Over time, the accumulated small changes can lead to long-term damage or even physical failures if the attack remains undetected.

- Stealthy false command attack
  In this type of attack, the attacker leverages the insecurity of the MODBUS protocol and inject stealthy false commands to the target PLC [1]. More than the control logic injection attack, this type of attack relies on a database of real request-response interaction pairs, allowing the adversary to consistently provide the expected responses to the HMI. By doing so, the attack remains hidden from the operator and effectively disconnects the PLC from the HMI. As a result, the operator perceives normal system behavior, while the actual controller may be compromised without detection.
- Zero-residual attack (ZeRa)
  In this scenario, the attacker injects false inputs but compensates elsewhere so the magnitude of residual is zero that system monitors do not detect anomalies [10]. In [6], the attack is launched by two state estimators, one is responsible for the physical system state estimation, and the other is responsible for estimating the detector state Consequently, the system operates under incorrect conditions without triggering any alerts.

(6) Miscellaneous attacks

- Payload Attacks
  A payload attack is a type of cyber attack where an attacker delivers and executes a malicious payload within a targeted system [15]. The payload refers to the part of the attack that carries out harmful actions, such as modifying system behavior, stealing data, or disrupting operations. In I&C systems, a payload attack often involves injecting malicious code into PLCs, sensors, or other control components to alter processes, disable safety mechanisms, or cause physical damage [16].
- Sequential attacks
  In this attack, the physical processes of a system can be disrupted by altering the control sequences. For example, this work [27] provided a metric, sequential attack graph, which is a representation that illustrates the steps an attacker could take to compromise a system or network in a sequential manner.
- Side-channel attacks

These attacks exploit side-channel data, including power consumption, electromagnetic emissions, LED signals, and acoustic signals, to access sensitive information or undermine system security. Unlike traditional attacks that target software vulnerabilities, side-channel attacks focus on the unintended emissions produced during computation. For instance, LED indicators can serve as an indirect channel for data exfiltration. Malware in an air-gapped network could manipulate LED signals to encode and transmit sensitive information, which an external sensor or camera could then capture and decode [24].

Table 1. Attack Taxonomy

Attack Category	Visibility/ Detectability	Target Component	Sub-category	Attack Scenarios	Critical Functions
DoS attacks	High	Communication channels, Controller, HMIs	VDoS DDoS	Large volume of data —	Controller failed to respond —
Integrity attacks	Medium to Low	Sensors, Controllers, Communication channels	Simple attacks Covert attacks	Modified sensor readings Replay attack	Inadequate cooling or improper flow rate
Control-theoretic attacks	Low	Controllers, Actuators	Perturbation-based Information Theoretic	Oscillation in control output Learn system info offline	QoIs oscillation Information leakage (No active interference)
Injection attacks	Medium to Low	Sensors, Controllers, Communication channels	False data injection Time delay injection Control logic injection	Modified sensor readings Delayed sensor readings Modified control parameters	Inadequate cooling or improper flow rate Delayed state estimation Corrupted state estimation
Stealthy attacks	Very Low	Sensors, Controllers, Communication channels, Actuators	Stealthy false command Zero-Residual Attack	Modified control commands Multiple variables modified to make residual zero	Inadequate cooling or improper flow rate State deviation without detection
Miscellaneous attacks	Varies	Firmware, Controller, HMI	Payload Sequential Side-channel	Firmware injection Modified control logic sequences LED/acoustic sensor eavesdropping	Inadequate cooling or improper flow rate Information leakage (No active interference)

5.2 Attack Implementation Through Data Transformation

The attacks described in our taxonomy can be implemented through various data transformation techniques. These transformations provide concrete mechanisms for manipulating sensor readings, control signals, and system states to realize the attack categories outlined earlier. Table 2 summarizes key transformation functions and their applications to different attack categories.

These transformation techniques provide the technical foundation for implementing the attack categories described in our taxonomy. For example, a False Data Injection attack targeting temperature sensors could be realized through a combination of scaling, offset, or spike transformations. Similarly, a Stealthy Attack might combine conditional transformations with physical relationship violations to ensure the manipulated values remain below detection thresholds while creating system impacts.

The implementation of these transformations requires specific technical approaches. For instance, scaling transformations apply a multiplier to sensor values, which could be constant (e.g., 0.8× of the true value) or time-varying (e.g., gradually decreasing from 1.0 to 0.7 over several hours). Oscillation transformations introduce periodic variations by adding sinusoidal components with configurable frequency, amplitude, and phase. Delay transformations cache historical values and introduce time lags in the reporting of sensor data or control commands, potentially desynchronizing related signals.

5.3 Attack Scenarios in METL

We have identified several additional attack vectors that could impact the METL system's operations. These scenarios represent a diverse range of attack techniques and system impacts, as summarized in Table 3.

Table 2. Data Transformation Techniques for Attack Implementation

Transformation Technique	**Implementation Approach**	**Applicable Attack Categories**
Scaling Transformation	Multiply sensor values by configurable factors, with potential time-dependency	Integrity Attacks, False Data Injection, Stealthy Attacks
Oscillation Transformation	Add sinusoidal patterns with configurable frequency and amplitude	Control-Theoretic Attacks (Perturbation-based), False Data Injection
Spike Transformation	Insert temporary anomalies with configurable magnitude	Integrity Attacks, False Data Injection
Offset Transformation	Add/subtract constants or time-varying offsets	Integrity Attacks, False Data Injection, Stealthy Attacks
State Toggle Transformation	Manipulate binary/multi-state values to create inconsistent states	Injection Attacks, Stealthy Attacks
Delay Transformation	Cache and delay values by configurable time periods	Time Delay Injection, DoS Attacks
Replay Transformation	Record and replay historical data, potentially with modifications	Covert Attacks (Replay), Stealthy Attacks
Precision Degradation	Reduce value precision through controlled rounding	Integrity Attacks, Information Theoretic Attacks
Noise Injection	Add random noise with configurable distributions	Integrity Attacks, DoS Attacks
Conditional Transformation	Apply transformations based on complex conditions	Stealthy Attacks, Zero-Residual Attacks
Physical Relationship Violation	Manipulate related sensors to create physically impossible states	Control-Theoretic Attacks, Zero-Residual Attacks
Propagation Transformation	Spread effects across related sensors with realistic delays	Sequential Attacks, Stealthy Attacks

Table 3. Expanded Attack Scenarios for METL

Attack Scenario	**Primary Category**	**Target Components**	**Potential Impact**
False Flow Rate Reporting	False Data Injection	Flow rate sensors (air and sodium)	Inadequate cooling, thermal stress, potential pump damage
Flow Oscillation Pattern	False Data Injection	Electromagnetic pumps and flow meters	System resonance, inefficient heat transfer, premature equipment failure
Coordinated Thermocouple Manipulation	False Data Injection	Multiple thermocouples across system	Thermal stress, uneven expansion, accelerated corrosion
Temperature Spike and Recovery	False Data Injection	Temperature sensors	False safety system activation, operational interruptions, operator distrust
Valve State Inconsistency	Injection Attack	Multiple valve position indicators	Operator confusion, valve seat damage, operational downtime
Sequential Valve Manipulation	Injection Attack	Multiple valves affecting sodium purification	Reduced sodium purity, increased corrosion, cold trap efficiency reduction
Sensor Data Delay	Time Delay Injection	Multiple sensor types	Delayed operator response, control system instability, potential emergency trips
Localized Data Replay	Replay Attack	Subsystem sensor clusters	Hidden hotspots, undetected maintenance conditions, experimental data corruption
Flow-Temperature Relationship Attack	Combined (FDI + Replay)	Flow and temperature sensors	Operator confusion, incorrect cooling adjustments, energy waste
Command-Feedback Desynchronization	Combined (Timing + Injection)	Control systems and feedback loops	Control hunting/oscillation, increased component wear, operator mistrust

5.4 Severity Tier Implementation

To enable systematic evaluation of detection performance across attack intensities, we implemented a five-tier severity scaling system that parameterizes attack magnitude while maintaining scenario-specific characteristics. Each attack scenario supports five intensity levels designated as tier_001, tier_005, tier_010, tier_050, and tier_100, representing 1%, 5%, 10%, 50%, and 100% of maximum attack intensity respectively.

The tier scaling implementation varies by attack scenario and transformation type, reflecting the diverse nature of cyber-physical attack vectors:

Amplitude-Based Scaling. For attacks involving sensor value modification (scaling, offset, spike transformations), tier parameters directly control the magnitude of alterations:

- **Scaling attacks**: Tier_001 applies 1% deviation from normal values (0.99× or 1.01× multipliers), while tier_100 applies maximum credible scaling (0.5× or 2.0× multipliers)
- **Offset attacks**: Tier values control additive bias magnitude relative to sensor measurement ranges
- **Spike attacks**: Tier parameters determine peak amplitude and duration of injected anomalies

Frequency and Duration Scaling. For temporal attacks (oscillation, delay, replay), tiers control time-domain characteristics:

- **Oscillation attacks**: Higher tiers increase frequency and amplitude of injected sinusoidal patterns
- **Delay attacks**: Tier values determine delay duration from milliseconds (tier_001) to seconds (tier_100)
- **Replay attacks**: Tiers control the temporal span and repetition patterns of historical data injection

Coordination and Complexity Scaling. For multi-sensor coordinated attacks, tiers control the breadth and sophistication of manipulation:

- **Sensor count**: Higher tiers affect larger numbers of coordinated sensors
- **Temporal coordination**: Advanced tiers introduce complex phase relationships and propagation delays
- **Physical relationship violations**: Higher tiers create more severe violations of expected correlations between related measurements

This parameterized approach enables systematic evaluation of how attack detectability varies with intensity, providing insights into detection threshold behaviors and algorithm sensitivity characteristics. However, as discussed in our ROC analysis methodology, achieving meaningful tier differentiation requires careful parameter selection to ensure that intensity changes cross genuine detection boundaries rather than remaining within algorithm insensitivity ranges.

5.5 Fleet-level Attack Considerations

While our primary focus has been on attacks targeting individual components or subsystems within a single advanced reactor, it is important to consider attack scenarios that have fleet-wide implications. In a distributed control system environment with multiple facilities or subsystems, sophisticated attackers may leverage cross-system attacks that are difficult to detect when examining each system in isolation.

Fleet-level attack patterns may include:

- **Propagating attacks** that migrate across facilities with time delays, creating the appearance of independent issues rather than a coordinated campaign
- **Coordinated oscillations** implemented across multiple systems with strategic phase shifts, potentially creating resonance effects or masking the artificial nature of the oscillations
- **Cross-facility data transplantation** where data patterns from one facility are replicated in another, creating false correlations that may confuse detection systems
- **Facility-specific variations** of common attack patterns, tailored to evade detection systems that look for identical signatures across systems
- **Cascading failure simulations** that trigger realistic cross-system dependencies, making attacks appear as natural consequence cascades

For example, attackers might implement the Flow Oscillation Pattern attack across multiple facilities with specific time offsets, creating the appearance of a propagating phenomenon rather than a coordinated attack. This could lead investigators to search for physical or environmental causes rather than cyber interference.

Similarly, a Coordinated Thermocouple Manipulation attack could be implemented with facility-specific variations, creating thermal profile anomalies that appear unique to each facility but collectively serve the attacker's broader objective of disrupting operations or inducing incorrect maintenance actions.

These fleet-level considerations highlight the importance of implementing detection mechanisms that can correlate events across distributed systems, identifying patterns that may not be apparent when analyzing each system independently.

6 Detection Challenges

Detecting the attacks described in our taxonomy presents several challenges that must be addressed when designing defensive strategies. These challenges vary by attack type and implementation method.

- **Gradual attacks** defeat simple threshold detection by slowly introducing deviations that remain within acceptable parameter ranges
- **Oscillatory patterns** may be mistaken for normal system variations unless pattern recognition is employed
- **Coordinated multi-sensor manipulations** require correlation analysis across multiple parameters to detect physically impossible states
- **Replay attacks** present historically valid data that passes validity checks but masks current conditions
- **Timing attacks** with small delays are difficult to detect without precise temporal analysis
- **Zero-residual attacks** explicitly design manipulations to produce no detectable anomalies in monitoring systems

The False Flow Rate Reporting scenario illustrates the challenge of gradual manipulation, where the attack progressively scales sensor readings over hours or days, keeping values within normal operational ranges while creating hazardous conditions. Traditional threshold-based detection would fail to identify this attack until conditions become severe.

Similarly, the Valve State Inconsistency scenario creates logical contradictions in system state that might not trigger alarms focused on individual value ranges. Detection requires context-aware monitoring that understands the relationships between different valve states and can identify impossible configurations.

Effective detection strategies must employ a multi-layered approach combining some subset of:

(1) Physical model validation to verify that sensor readings conform to expected physical relationships
(2) Temporal correlation analysis to detect subtle timing anomalies in related signals
(3) Cross-system correlation to detect coordinated manipulations that span multiple subsystems

(4) Deep learning approaches capable of identifying subtle deviations from normal operating patterns

For example, detecting a Flow-Temperature Relationship Attack requires physical model validation that understands the expected correlation between flow rates and temperature changes. When these relationships are violated, even if individual parameters remain within normal ranges, the system can identify potential manipulation.

The detection challenges reinforce the need for defense-in-depth strategies that combine multiple detection techniques, each designed to address specific attack vectors while collectively providing comprehensive coverage against the attack taxonomy presented in this paper.

6.1 Data Representation and Scaling Challenges

Beyond attack-specific detection challenges, fundamental issues in data representation and preprocessing can severely impact detection system performance. These challenges are particularly acute in nuclear reactor environments with heterogeneous sensor networks.

Nuclear reactor systems integrate diverse sensor types measuring fundamentally different physical quantities with widely varying natural scales. Temperature measurements may span hundreds of degrees, pressure readings tens of PSI, and flow rates hundreds of gallons per minute. When these disparate measurements are combined for machine learning analysis, inappropriate normalization strategies can:

- Create artificial correlations between unrelated sensor types, leading to false dependencies in learned models
- Mask genuine attack signatures when changes in one sensor type are overwhelmed by the scale differences of other sensors
- Cause training instabilities in neural networks when input ranges exceed expected bounds
- Dilute attack signals affecting specific sensor subsets when averaged across all sensor types

A particularly challenging phenomenon occurs when attack signatures affecting a subset of sensors become statistically diluted by the majority of unaffected sensors. In nuclear facilities with hundreds or thousands of sensors, attacks typically target specific subsystems (e.g., cooling circuits, valve operations) affecting perhaps a dozen sensors while leaving the rest operating normally. Global scaling and averaging approaches can render these focused attacks undetectable by:

- Averaging attack-induced reconstruction errors with normal sensor errors

- Applying global threshold criteria that fail to account for sensor-specific baselines
- Learning model parameters dominated by the statistical properties of unaffected sensors

These scaling challenges require architectural solutions in the preprocessing pipeline rather than algorithmic improvements in detection methods. Effective detection systems must preserve sensor-specific characteristics while enabling cross-sensor correlation analysis. This necessitates normalization strategies that maintain physical interpretability and prevent statistical artifacts from masking genuine cybersecurity threats.

The importance of addressing these fundamental data representation issues cannot be overstated—improperly scaled input data can render sophisticated detection algorithms completely ineffective, regardless of their theoretical capabilities.

7 System Architecture and ML Evaluation Framework

To support rigorous and repeatable cyber-physical evaluations, we implemented a modular architecture that separates configuration, attack generation, data services, preprocessing, and ML experimentation. The framework enables dual-path loading (clean training data vs. attack testing data), experiment tracking via structured JSON logs, and hardware-aware execution for high throughput.

This section first outlines the overall architecture and data flow, then details the data format and preprocessing choices, the evaluation protocol (labeling, thresholds, missing-data policy, metrics/uncertainty), and finally operational considerations (deployment, limitations, reproducibility).

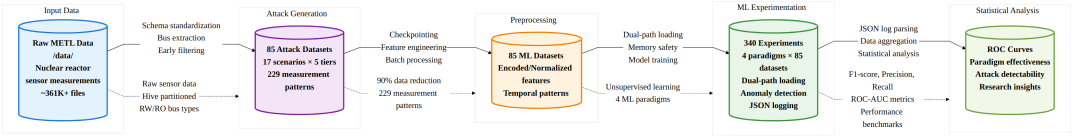


Fig. 4. Simplified data flow from raw METL data through attack generation, preprocessing, ML experimentation, and analysis.

7.1 Data Format and Feature Set

Preprocessed Parquet datasets use the long-format schema emitted by the METL fetcher: core columns `timestamp`, `controller`, `module`, `datatype`, `group`, `name`, `meta_type`, and `value` (string). Optional `extra_params` may appear for RW tags. For storage, partition columns `year`, `month`, `day`, and `bus_type` are added and used for on-disk partitioning.

For ML, we construct per-row features comprised of: label encodings of identifiers (`bus_type`, `controller`, `module`, `group`, `name`, `meta_type`) plus seven temporal features (`som`, `moh`, `hod`, `dow`, `wom`, `moq`, `qoy`) and the numeric `value`. This yields a 14-feature numeric vector (6 label-encoded identifiers + 7 temporal + 1 value) for paradigms operating on long-format data. The `timestamp` is reserved for sequencing/windowing; `datatype` and partition columns (`year/month/day`) are excluded from model inputs.

The pipeline supports memory-safe streaming (online ML) processing, pivoting from long to wide format, schema standardization to ensure training–test consistency, and filling of missing sensors after pivot. **Data Format Selection**: Different ML paradigms utilize different data representations—autoencoders operate on wide-format pivoted data (dynamically determined sensor count + 7 temporal features), while other paradigms may use long-format sequential data (14 features per timestep). This architectural choice ensures optimal input representation for each algorithm type while maintaining preprocessing consistency.

7.1.1 Sensor Subset Selection. METL's comprehensive instrumentation includes approximately 11,899 sensor and setpoint data points monitoring diverse aspects of facility operation (temperatures, pressures, flows, valve positions, electrical parameters, etc.). For cybersecurity evaluation, we filtered this full sensor set based on relevance to our simulated scenarios, resulting in 214 sensors that were either direct targets of our attack scenarios or related to them.

This sensor selection process focused on:

- **Attack-Relevant Sensors**: Direct targets of implemented attack scenarios (thermocouples, flow meters, valve position indicators, heater controls)
- **Physically Related Sensors**: Measurements from components with physical relationships to attack targets (upstream/downstream temperatures, related control loops)
- **Representative Sensor Types**: Additional sensors of the same types and physical principles as attack targets to provide realistic multi-sensor context
- **Control System Integration**: Sensors integrated with the same control loops and I&C systems as attack targets

This filtering approach enables realistic evaluation of attack detection across sensor networks while maintaining computational tractability. The selected sensor subset preserves the multi-sensor correlation patterns and control dependencies that attacks exploit, while excluding sensors (e.g., facility utilities, non-critical monitoring) that would not contribute meaningfully to cybersecurity analysis.

7.1.2 Multi-Sensor Scaling Architecture. Nuclear reactor instrumentation presents unique challenges for machine learning applications due to the heterogeneous nature of sensor measurements. Conventional approaches using global normalization across all sensor types can introduce artificial correlations and mask genuine attack signatures. Our implementation employs a hierarchical scaling strategy that preserves sensor-specific characteristics while enabling effective anomaly detection:

Sensor-Specific Normalization. Each unique sensor combination (identified by bus, controller, module, group, measurement type, and meta type) receives individual scaling using robust statistical methods (IQR-based scaling to diminish the impact of outliers). This approach prevents inappropriate mixing of different physical quantities and maintains the natural relationships within each sensor type.

Temporal Feature Scaling. Time-based features (second of minute, minute of hour, hour of day, day of week, week of month, month of quarter, quarter of year) receive dedicated normalization appropriate to their respective range, preserving temporal patterns while ensuring consistent input scaling for machine learning models.

This scaling architecture proved essential for achieving effective anomaly detection performance. Initial implementations using global scaling across heterogeneous sensors resulted in poor model convergence and reduced sensitivity to attack patterns. The sensor-specific approach maintains physical interpretability while enabling robust statistical learning.

7.1.3 Missing Data and High-Variance Training Data Filtering. To ensure meaningful and consistent data, we forward-fill gaps in the data post-pivot (for experiments using wide-format data). For some models, a critical training data quality enhancement filters out steady-state periods to focus models on informative dynamic operation:

- **High-Variance Training Filter (LSTM/Autoencoder/Dependency Violation)**: Models train exclusively on high-variance chunks (≥ 0.1 variance threshold across ≥ 12 sensors) to focus learning on dynamic periods (system transitions, maintenance operations, process changes) vs steady-state reactor operation where sensors barely vary. This fundamental shift from quantity-based (all data) to quality-based (dynamic periods only) training prevents models from learning steady-state noise patterns that obscure attack modifications.
- **Change Point Detection**: No variance filtering applied as this paradigm effectively detects changes across all operational states.

Training Data Quality Impact: The high-variance filtering approach addresses a critical limitation where LSTM and Autoencoder paradigms previously learned to predict/reconstruct constant sensor values with minimal variation, making attack modifications indistinguishable from normal measurement noise.

By training exclusively on dynamic periods—filtering out approximately 80-90% of steady-state operational chunks with stricter variance thresholds—models learn meaningful temporal patterns and sensor correlations during system transitions. This approach enables attack modifications to be more readily detectable against learned dynamic baselines rather than steady-state noise patterns.

7.1.4 Temporal Downsampling Strategy. To balance signal fidelity and throughput, we apply a 30-second downsampling strategy that aligns with METL process dynamics and the attack signatures under study. We justify this downsampling approach as follows:

- Many plant-level thermal/flow phenomena in sodium loops evolve on minute-scale time constants; our targeted attack patterns remain well-resolved at 30-second cadence.
- Downsampling provides computational efficiency, substantially lowering memory pressure for pivot operations and accelerating end-to-end evaluations while maintaining sufficient temporal resolution for anomaly detection.
- Downsampling is performed with windowed aggregation appropriate for anomaly detection (e.g., robust averaging)
- Temporal encodings (e.g., second/minute/hour-of-day, day-of-week, etc.) remain intact at the new cadence, preserving diurnal/operational context used by learning algorithms.
- For physics-based system analysis and prognostics (e.g., fast transients, component health monitoring), higher sampling rates may be appropriate; our cyber-focused evaluations target minute-scale signatures, for which a 30-second cadence is sufficient. The framework supports higher-rate modes when required.

Threat model (see Attack Taxonomy). We assume an adversary capable of network-level manipulation (e.g., MITM, replay, precision degradation) but not physical tampering or Safety Instrumented System (SIS) bypass; insider-equivalent data-plane access is modeled. Goals include stealthy integrity violations, coordinated timing, and minute-scale disturbances; cryptographic compromise and kinetic outcomes are out-of-scope.

7.2 Machine Learning Paradigms

We selected a cross-section of ML paradigms to test, covering the elements above. All paradigms follow a dual-path methodology: models train exclusively on clean data and evaluate on attack data to prevent leakage. While not fully comprehensive, these provide sufficient diversity to understand the key challenges for detecting cyber attacks in these environments.

7.2.1 LSTM Forecasting. The LSTM (Long Short-Term Memory) approach employs a specialized recurrent neural network architecture designed for temporal pattern recognition in nuclear reactor sensor data. Unlike traditional anomaly detection that analyzes individual data points, this method examines sequences of measurements over time to learn normal operational patterns, making it particularly effective at detecting coordinated attacks that unfold across multiple timesteps.

Core Architecture: 4-layer tapering LSTM with progressively smaller hidden layers (40 $\rightarrow$ 32 $\rightarrow$ 24 $\rightarrow$ 16 units) enabling hierarchical temporal feature extraction. The network incorporates 15% dropout regularization and gradient checkpointing with mixed precision training to ensure computational efficiency while processing large-scale reactor datasets.

Data Processing and Training:

- **Sequence Construction**: Creates 50-timestep sequences (approximately 25 minutes at 30-second intervals) from wide-format vectors containing all available sensor measurements per timestep. Each LSTM model is specialized for a target sensor, using the full multi-sensor context to predict future values.
- **Per-Sensor Architecture**: Employs specialized target-focused models (one per sensor), each predicting one target sensor using all sensors as context, eliminating signal dilution in loss calculation while preserving multi-sensor context for temporal pattern learning
- **Training Strategy**: Batch size 16, learning rate 0.001 using Adam optimizer, trained for up to 20 epochs with early stopping (patience=8) to prevent overfitting on normal operational patterns
- **Dynamic Period Focus**: Trains exclusively on high-variance chunks representing dynamic operational periods (system transitions, maintenance operations) rather than steady-state data, enabling better attack signal detection
- **Forecasting Method**: Predicts next timestep values (X $\rightarrow$ X+1) using Mean Squared Error loss, learning temporal dependencies characteristic of normal reactor behavior

- **Attack Detection**: Identifies anomalies through persistent prediction errors exceeding the 85th percentile threshold of training errors, indicating deviations from learned normal patterns

Strengths/Limitations:

- Captures temporal dependencies and coordinated multi-sensor dynamics.
- Benefits from minute-scale downsampling; very long-range effects may require larger time windows.

7.2.2 Autoencoder Anomaly Detection. The autoencoder approach employs a neural network architecture trained to learn compact representations of normal reactor operational patterns. This method operates by compressing sensor data into a lower-dimensional latent space and then reconstructing the original input. During cyber attacks, the network struggles to accurately reconstruct manipulated sensor readings, producing high reconstruction errors that serve as anomaly indicators.

Core Architecture: Dense autoencoder featuring a 4-layer encoder ($64 \rightarrow 32 \rightarrow 16 \rightarrow 8$ units) with symmetric decoder architecture, compressing sensor patterns to a 4-dimensional latent bottleneck. The network incorporates batch normalization, 10% dropout regularization, and gradient checkpointing to ensure memory-efficient training on large-scale reactor datasets.

Data Processing and Training:

- **Feature Representation**: Processes wide-format feature vectors containing all available sensor measurements per timestep.
- **Training Strategy**: Batch size 128 with 0.001 learning rate using Adam optimizer, trained for up to 20 epochs with early stopping (patience=8) to prevent overfitting on normal operational patterns
- **Dynamic Period Focus**: Trains exclusively on high-variance chunks representing dynamic operational periods rather than steady-state reactor operation, improving sensitivity to attack-induced changes
- **Reconstruction Learning**: Optimizes Mean Squared Error loss between input features and reconstructed output, learning to compress and reconstruct normal sensor patterns with minimal information loss
- **Attack Detection**: Identifies anomalies using maximum reconstruction error per sample across all sensors rather than averaging, ensuring attacks affecting any sensor subset trigger detection

Strengths/Limitations:

- Effective for cross-sensor inconsistencies at a given timestep.
- Less sensitive to purely temporal-only anomalies without explicit temporal features.

7.2.3 Dependency Violation Detection. The dependency violation approach recognizes that nuclear reactor systems exhibit strong interdependencies between sensors due to underlying physics (temperature-pressure relationships, flow-thermal coupling) and control logic constraints. Cyber attacks often violate these natural relationships, creating detectable inconsistencies even when individual sensor values remain within normal operational ranges. This method is particularly effective at identifying sophisticated attacks that manipulate multiple sensors in physically impossible ways.

Dynamic Correlation Baseline Learning: Dependency violation detection employs sophisticated baseline learning methodology to achieve accurate violation detection and interpretability.

Baseline Learning Approach: Correlation violation calculation uses dynamically learned baselines from training data rather than static thresholds. This approach captures actual sensor correlation patterns during normal operations, enabling precise detection of dependency disruptions specific to each reactor configuration.

Implementation Details:

- **Dynamic Correlation Baselines**: System learns and stores actual correlation values between sensor pairs during training, creating individualized baseline expectations for each dependency relationship
- **Context-Aware Violation Scoring**: Violation scores calculated as absolute deviation from learned baselines, providing scenario-specific sensitivity to dependency changes
- **Direct Performance Assessment**: ROC analysis uses raw violation scores without transformation, enabling straightforward interpretation of detection capabilities

The dependency violation detection employs a sliding window temporal analysis approach to generate realistic per-timepoint variation scores. Rather than broadcasting static sensor-level violation scores to all timepoints (which would create artificial perfect separation), the system analyzes dependency violations within adaptive windows (10-30 samples) centered around each timepoint. This methodology captures the temporal evolution of dependency violations during attacks while maintaining realistic variance within both attack and baseline periods. The 95th percentile ensemble scoring across sensors within each window effectively captures the most significant violations while the temporal windowing ensures that early attack phases, peak attack periods, and sustained attack effects receive distinct scores. The implementation incorporates learned baseline correlations from

training data rather than fixed thresholds, enabling accurate detection of dependency disruptions specific to each sensor pair relationship and temporal context.

Experimental Architecture and Implementation

The dependency violation detection system operates on dynamically selected sensors (filtered from available measurements based on data quality) across reactor subsystems, employing a three-analyzer ensemble approach for comprehensive relationship modeling:

- **Correlation Analyzer**: Computes pairwise linear correlations across all sensor combinations, establishing baseline dependency expectations and detecting violations of expected correlation patterns during attacks
- **Granger Causality Analyzer**: Tests temporal causal relationships using AIC/BIC-optimized lag selection (1-10 lags, 30s-5min) with adaptive percentile-based sensor pair prioritization (85th percentile correlation threshold) to identify disruptions in temporal dependency patterns that indicate control system manipulation
- **Random Forest Analyzer**: Models complex non-linear interdependencies using 10-estimator forests with 5-layer depth, capturing sophisticated physics-based relationships beyond linear correlations

Advanced Correlation Baseline Learning: Dependency violation detection employs sophisticated baseline learning to achieve accurate violation score calculation:

- **Dynamic Baseline Learning**: Correlation violation calculation uses learned baseline correlations from training data rather than fixed thresholds, enabling context-specific violation detection
- **Direct Score Interpretation**: Violation scoring directly measures deviations from established dependency patterns without additional score transformations
- **Raw Score ROC Analysis**: Performance evaluation uses raw violation scores, providing direct assessment of dependency disruption detection capability

Implementation Methodology: Dependency violation calculation incorporates the following advanced techniques:

- **Baseline Correlation Storage**: System captures and stores actual correlation values between sensor pairs during training phase, creating individualized baselines for each dependency relationship
- **Context-Aware Violations**: Violation scores calculated as absolute deviation from learned baselines rather than universal thresholds, enabling detection of scenario-specific dependency disruptions
- **AIC/BIC Lag Optimization**: Granger causality analysis uses Akaike Information Criterion to select optimal lag parameters (1-10 lags) for each sensor pair, adapting

to diverse nuclear facility temporal dynamics ranging from 30-second control loops to 5-minute thermal processes

- **Adaptive Pair Selection**: Smart prioritization using adaptive percentile-based filtering (85th percentile correlation threshold) to select 100 most promising sensor pairs per training chunk from all possible sensor combinations (N × (N-1) where N is the dynamic sensor count). This approach automatically adapts to varying correlation distributions across operational conditions—testing stronger relationships during maintenance periods and best-available relationships during steady-state operation, ensuring comprehensive coverage while focusing computational resources on meaningful dependencies
- **Direct Performance Evaluation**: ROC analysis uses raw violation scores without transformation, providing straightforward assessment of dependency detection performance

7.2.4 Change Point Detection. The change point detection approach identifies moments when the statistical properties of sensor data undergo sudden shifts—a signature of cyber attacks that introduce step changes, oscillations, or other discontinuous alterations. This method is effective for detecting abrupt changes in sensor behavior, such as bias injection, oscillatory patterns, or coordinated state transitions.

Statistical Baseline Learning Approach: Employs a memory-efficient streaming system to establish sensor-specific baselines (mean, standard deviation, min/max, coefficient of variation) from large-scale training data, enabling scalable change detection across reactor subsystems.

Data Processing and Detection:

- **Streaming Statistical Learning**: Accumulates running statistics (mean, std, min/-max, coefficient of variation) per sensor from clean training data, without full dataset loading.
- **Per-Sensor Monitoring**: Computes change magnitude scores for each sensor during testing, using the maximum score per timepoint to preserve localized attack signals.
- **Statistical Change Detection**: Detects deviations using first- and second-order differences and moving window variance, with adaptive window sizes (minimum 3 samples, typically 10% of data length).
- **Percentile-Based Thresholds**: Applies dynamic detection boundaries using percentile thresholds (default: 95th percentile) derived from change score distributions, adapting to operational conditions.

- **Universal Applicability**: No variance filtering, making it effective across all operational states, including steady-state periods.

Strengths/Limitations:

- Interpretable, memory-efficient; excels at detecting step changes and broad distribution shifts.
- Less sensitive to subtle cross-sensor dependency violations unless paired with dependency analysis.

7.3 Evaluation Protocol and Operational Considerations

7.3.1 Data and Labeling. Clean intervals are sensor measurements from an extended time period (30-days) prior to a period overlapping the same period as the attack data (several hours). This training data is drawn from the pre-attack simulation dataset. This unique benefit of our attack simulation strategy means we can see the exact ground truth for what the measurements *would have been* without an attack. For evaluation:

7.3.2 Limitations and Validity Scope. Transformations emulate many attack manifestations but do not replace full closed-loop plant physics; SIS interactions and detailed actuator dynamics are out-of-scope. Setpoint changes and maintenance can mimic attacks if not flagged. Stealth, zero-residual attacks designed against specific models remain difficult without physics-informed constraints.

7.3.3 Reproducibility. Experiments are reproducible via pinned environment dependencies and Make targets. Seeds are set for training and data shuffling; datasets are versioned by scenario/tier and preprocessing profile (debug/production).

We introduced a cache for trained models with content-addressable keys, enabling model reuse across the 18 attack scenarios and 5 severity tiers. This eliminated redundant training and reduced total evaluation time by approximately two orders of magnitude for the full experiment suite. Additional efficiency improvements include adaptive percentile-based pair selection for dependency violation analysis, which automatically focuses computational resources on the most promising sensor relationships while adapting to varying correlation distributions across operational conditions. Across the full scenario–tier matrix, the training cache eliminates redundant training phases, enabling rapid test-time evaluation. Representative outputs for each paradigm are captured in the experiment logs and can be aggregated into model-level metrics (e.g., reconstruction/prediction error distributions, change point counts, dependency violation rates). At the system level, the combination of

caching, downsampling, and adaptive filtering provides end-to-end throughput suitable for operational studies.

7.3.4 ROC Analysis and Comparative Evaluation Methodology. Our evaluation framework employs comprehensive Receiver Operating Characteristic (ROC) curve analysis to assess attack detection performance across machine learning paradigms, implemented through the results aggregation system. This methodology provides standardized, quantitative comparison of detection capabilities across diverse algorithmic approaches and attack scenarios.

Critical Distinction: Training Parameters vs. Evaluation Thresholds. A fundamental aspect of our ROC-based evaluation methodology requires careful distinction between two different types of thresholds that serve completely different purposes in the machine learning pipeline:

- **Training-Time Hyperparameters**: Set once during model development to configure algorithm behavior (e.g., LSTM learning rates and epochs, autoencoder architecture depths, change point CUSUM thresholds, dependency violation correlation thresholds). These parameters determine how each paradigm learns normal reactor operation patterns from historical data.
- **ROC Evaluation Thresholds**: Automatically swept across all possible values during performance assessment to generate ROC curves. These thresholds determine the decision boundary between "normal" and "attack" classifications for each timepoint, using the continuous anomaly scores produced by the trained models.

This distinction is crucial for understanding how a single experimental run (with fixed training parameters) generates an entire ROC curve through systematic threshold sweeping during evaluation.

ROC Data Generation. Each ML paradigm generates ROC data through a dual-phase evaluation process:

- **Attack Score Generation**: During testing, each paradigm processes attack-modified sensor data and produces continuous anomaly scores using paradigm-specific methods (prediction errors for LSTM, reconstruction errors for autoencoders, change point statistics, dependency violation rates). The algorithms use fixed training parameters learned during the model development phase.
- **Baseline Score Generation**: To establish normal operation baselines, each paradigm applies identical algorithms to time-matched clean historical data extracted from

cached training chunks. This ensures fair comparison by using the same time periods and sensor patterns without attack modifications.

- **ROC Computation**: sklearn's `roc_curve` function automatically sweeps across all unique anomaly score values as decision thresholds, computing False Positive Rate (FPR) and True Positive Rate (TPR) for each threshold. This systematic threshold sweeping creates the stepped ROC curve from a single experimental run, with higher scores indicating greater anomaly likelihood.

Anomaly Score Definitions. Each paradigm produces continuous anomaly scores using specific mathematical metrics:

- **LSTM**: Mean Squared Error (MSE) between predicted and actual sensor values across all sensors for each timepoint
- **Autoencoder**: Mean Squared Error (MSE) between original input and reconstructed output, using maximum error across all sensors per timepoint to prevent signal dilution
- **Change Point Detection**: Maximum statistical change magnitude across CUSUM scores, window-based z-scores, and Bayesian change probabilities for each timepoint
- **Dependency Violation**: 95th percentile of absolute deviations from learned baseline correlations across all sensor pairs, computed using sliding window temporal analysis (10-30 sample windows) to generate realistic per-timepoint variation

Paradigm-Specific Threshold Percentiles. Different paradigms employ different threshold percentiles based on empirical performance characteristics and operational requirements:

- **Neural Networks (LSTM, Autoencoder)**: Use 85th percentile thresholds, providing good separation between normal and anomalous conditions while maintaining acceptable false positive rates for prediction/reconstruction error distributions
- **Statistical Methods (Change Point Detection, Dependency Violation)**: Use 95th percentile thresholds due to higher baseline variability in statistical measures, requiring more conservative thresholds to maintain operational false positive rates comparable to neural network approaches

These threshold choices were determined through limited empirical evaluation to balance detection sensitivity with false positive rates across paradigms, rather than reflecting fundamental distributional differences. **Important limitation**: These percentiles were not systematically optimized across the full parameter space—systematic hyperparameter tuning could potentially improve performance by identifying paradigm-specific optimal thresholds that better exploit each method's detection characteristics.

ROC Threshold Sweeping Mechanism. The ROC generation process uses dataset origin to determine ground truth labels while systematically varying decision thresholds:

- **Ground Truth Labeling**: All timepoints from control datasets receive label = 0 (normal), all timepoints from attack datasets receive label = 1 (attack)
- **Threshold Sweeping**: For each unique anomaly score value, sklearn treats that score as a decision threshold—timepoints with scores ≥ threshold are classified as "attack", scores < threshold as "normal"
- **ROC Point Generation**: Each threshold produces one (False Positive Rate, True Positive Rate) coordinate by comparing predicted classifications against ground truth labels
- **Curve Construction**: Connecting these points across all threshold values creates the complete ROC curve from a single experimental run

Multi-Dimensional Aggregation. The ROC analysis system generates three complementary comparative views:

- **Per-Paradigm Analysis**: Averages ROC curves across all attack scenarios for each ML approach, revealing algorithmic strengths and limitations. Confidence intervals show performance consistency across diverse attack types.
- **Per-Scenario Analysis**: Aggregates across ML paradigms for each attack type, identifying which attacks are most/least detectable regardless of detection method.
- **Per-Severity Analysis**: Groups by attack intensity tiers, quantifying the relationship between attack magnitude and detectability across the full experimental matrix.

Statistical Robustness. ROC curve averaging employs interpolation to standardize False Positive Rate sampling points across experiments, enabling meaningful statistical aggregation. Mean TPR values and standard deviations provide confidence intervals, while AUC score distributions quantify overall detection performance and consistency.

Attack Intensity Scaling and Tier Differentiation Challenges. A critical consideration in cybersecurity evaluation is the design of attack intensity tiers that produce meaningful differentiation in detection performance. Our severity tier system (Tiers 1-5) revealed an important methodological challenge: tier-wise ROC curves often exhibit unexpectedly similar performance, suggesting insufficient attack intensity gradation.

This similarity likely stems from several factors. First, detection algorithms often exhibit threshold-based behavior rather than gradual sensitivity scaling—once an attack crosses the “detectable” threshold, increasing intensity may not significantly improve detection performance. Conversely, attacks below the detection threshold remain difficult to detect

regardless of small intensity increases, creating a binary cliff effect rather than smooth gradation. Second, the parameters chosen for intensity scaling (amplitude, frequency, duration) may not represent the most impactful dimensions for detection systems. Some attack characteristics may be more critical than others, and scaling less-impactful parameters provides false differentiation without crossing meaningful detection boundaries.

This observation has important implications for cybersecurity research methodology. Effective tier systems require exponential or threshold-based scaling rather than linear parameter adjustments. Furthermore, the choice of scaling parameters should be guided by detection algorithm sensitivity analysis rather than intuitive attack characteristics. For meaningful tier differentiation, intensity differences may need to span orders of magnitude (10x-100x parameter changes) rather than incremental scaling (2x-5x changes) to cross the detection boundaries that separate barely detectable attacks (AUC ~0.55-0.65) from clearly detectable ones (AUC ~0.75-0.85) and obvious attacks (AUC ~0.90-0.95).

This ROC-based evaluation framework enables objective comparison of detection paradigms while accounting for the inherent variability in cybersecurity attack detection performance across different attack vectors and intensity levels.

8 Experimental Results

This section presents the quantitative performance evaluation of our four detection paradigms across attack scenarios and severity tiers. We analyze detection capabilities through ROC curve analysis, quantify algorithm-specific strengths and limitations, and examine the effectiveness of our signal dilution solutions.

8.1 Overall Performance Summary

Our evaluation framework processed 240 total experiments (4 paradigms × 12 scenarios × 5 tiers + baseline configurations) with comprehensive ROC analysis for each combination. Not all attack generators implemented were used in final analysis due to runtime limitations, and inability to fully debug by the end of the project. Key performance characteristics:

- **Detection Capability Range**: AUC performance spans 0.000–1.000 across 243 experiments, with overall mean performance of 0.656 ± 0.181, demonstrating significant variability in attack detectability
- **Paradigm Differentiation**: Change Point Detection leads (0.785), followed by LSTM (0.636), Dependency Violation (0.621), and Autoencoder (0.580), with LSTM showing most consistent performance ($\sigma = 0.057$)

- **Attack-Specific Performance**: Scenarios exhibit 4× detectability variation from highly detectable physics violations and equipment failures to challenging stealthy attacks like precision degradation and cross-facility transplants
- **Tier Differentiation**: Severity tiers (001→100) show paradigm-dependent scaling patterns, with some paradigms maintaining consistent performance across tiers while others demonstrate clear severity sensitivity

Across 243 total experiments, our comprehensive evaluation achieved mean AUC performance of 0.656 with standard deviation 0.181. The best-performing paradigm (Change Point Detection, AUC=0.785) significantly outperformed the lowest (Autoencoder, AUC=0.580), representing a 35% performance differential. Attack detectability ranged from near-perfect detection ($AUC \geq 0.90$) for physics violations to challenging evasion scenarios ($AUC \leq 0.30$) for precision degradation attacks, demonstrating the critical importance of paradigm-attack matching for operational deployment.

8.2 ROC Analysis and Comparative Performance

Figure 5 presents the aggregate performance comparison across all detection paradigms, while Figures 6 and 7 provide complementary views by attack type and severity level respectively. Figures 7 and 7 provide detailed paradigm-specific metric breakdowns, showing the performance distribution across attack scenarios and severity tiers for each detection approach.

Table 4 summarizes the Area Under Curve (AUC) performance metrics for each detection paradigm across attack scenarios and severity tiers.

Table 4. AUC Performance Summary by Detection Paradigm

Detection Paradigm	Mean AUC	Std Dev	Min AUC	Max AUC	Median AUC
Change Point Detection	0.785	0.127	0.219	1.000	0.775
LSTM Anomaly Detection	0.636	0.057	0.568	0.823	0.635
Dependency Violation	0.621	0.199	0.000	0.823	0.689
Autoencoder Reconstruction	0.580	0.226	0.091	0.920	0.659

Adversarial Score Inversion Case Study: ROC analysis revealed a particularly instructive failure mode where change point detection achieves AUC scores below 0.5 (worse than random) for precision trust decay attacks. This counterintuitive result demonstrates successful adversarial manipulation of the detection algorithm itself. The precision trust decay attack combines sensor precision reduction (quantizing values to fewer decimal

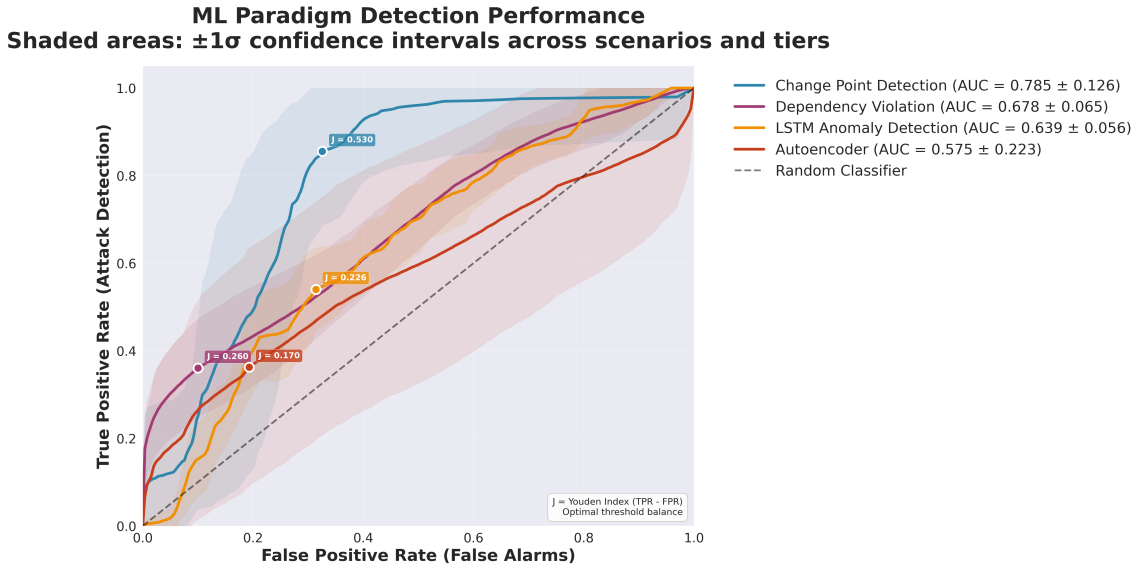


Fig. 5. ROC curves comparing detection paradigms aggregated across all attack scenarios. Legend entries are sorted by descending AUC performance for improved readability.

places) with Gaussian noise injection. Rather than increasing detectability, this attack strategy *reduces* the fine-grained statistical variations that change point detection algorithms rely upon. The precision degradation transforms natural sensor variations (e.g., 23.47291°C → 23.5°C) into quantized, apparently more stable measurements. Normal baseline data retains natural process noise and micro-variations, producing higher change point scores than the attack data. This inversion (normal data appearing more anomalous than attack data) results in AUC < 0.5, indicating that the attack successfully *hides* malicious changes by removing the very statistical signals that change point detection seeks. This represents a sophisticated evasion strategy that exploits the fundamental assumptions of statistical change detection methods.

8.3 Attack Scenario Detectability Analysis

Analysis of 243 experimental runs reveals significant variation in attack detectability across scenarios. Multi-site coordinated attacks demonstrate the highest mean detectability (AUC = 0.739), followed by sensor drift with dropouts (0.684) and flow oscillation (0.678). Precision trust decay exhibits the lowest detectability (0.592), representing the most challenging scenario across all ML paradigms.

Paradigm-attack interactions show distinct specialization patterns. Change point detection dominates performance for 10 of 12 scenarios, achieving peak effectiveness against

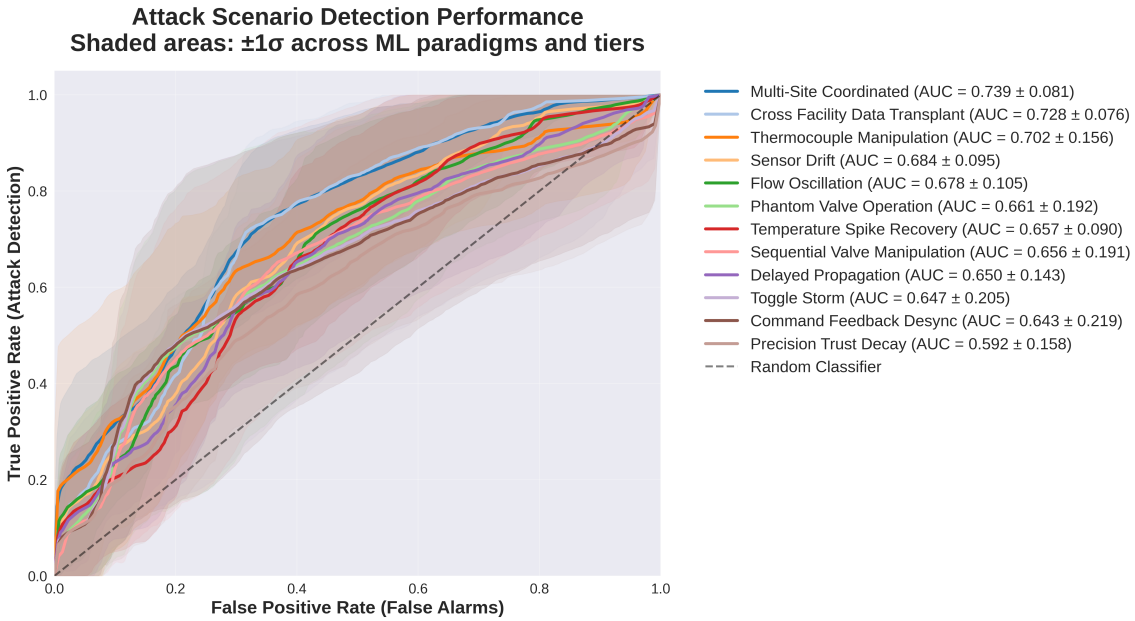


Fig. 6. ROC curves comparing attack scenario detectability aggregated across all detection paradigms. Legend entries are sorted by descending AUC performance.

toggle storms (AUC = 0.890) and command feedback desync (0.887). Dependency violation detection uniquely excels at multi-site coordinated attacks (0.810), while LSTM anomaly detection provides the strongest response to precision trust decay (0.689). This specialization suggests ensemble approaches combining complementary paradigms could enhance overall detection coverage.

For defense strategy selection, scenarios with high paradigm variation (cross-facility data transplant: $\sigma = 0.215$, coordinated thermocouple manipulation: $\sigma = 0.173$) benefit most from multi-paradigm deployment, while scenarios showing consistent cross-paradigm performance can rely on single specialized detectors.

8.4 Severity Tier Performance Validation

Severity tier differentiation analysis reveals limited correlation between attack intensity and detectability across most scenarios. Only 6 of 12 attack scenarios demonstrate correlation coefficients below 0.1 between severity tier and AUC performance, indicating tier-insensitive behavior. Cross-facility data transplant (r = 0.249) and sequential valve manipulation (r = 0.217) show the strongest positive tier sensitivity, where higher severity tiers become more detectable.

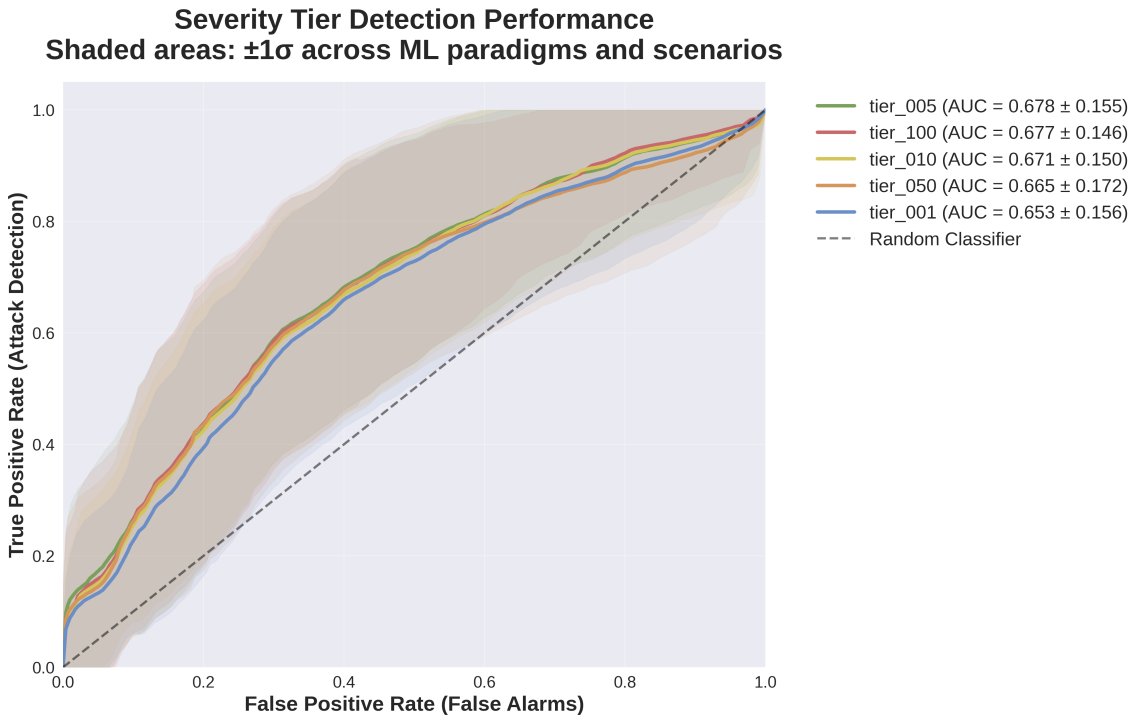
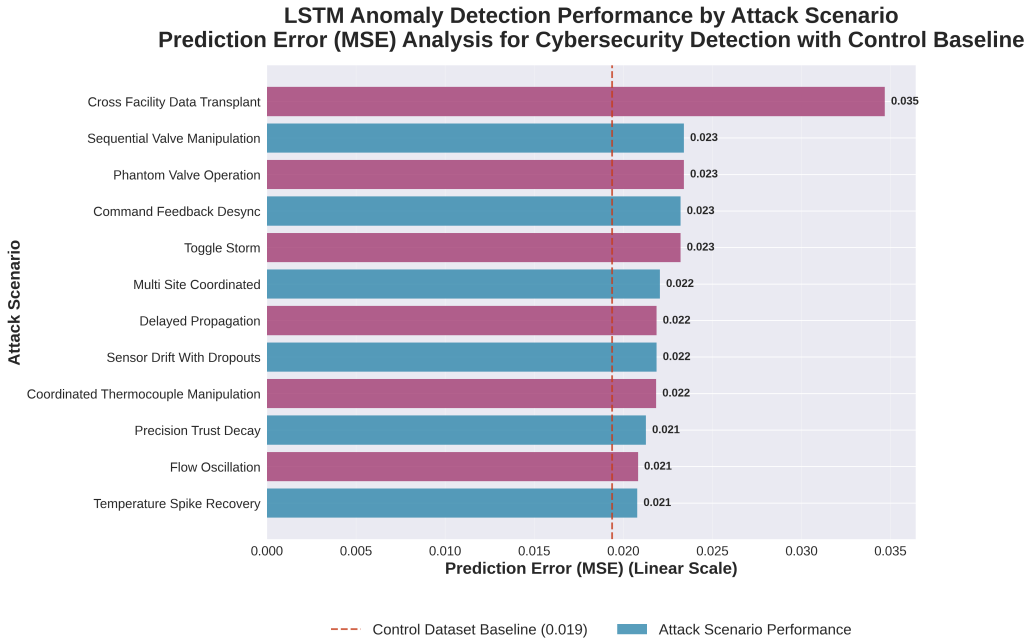
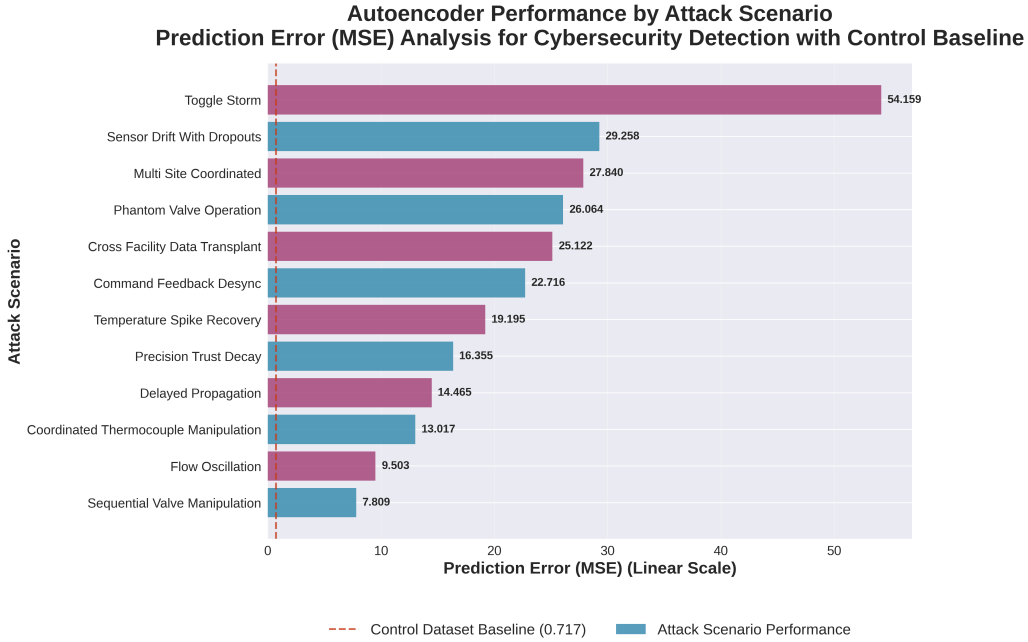


Fig. 7. ROC curves comparing detection performance across attack severity tiers aggregated across all paradigms and scenarios.

Interestingly, delayed propagation (r = -0.280) and phantom valve operation (r = -0.245) exhibit inverse tier sensitivity, where higher severity paradoxically becomes harder to detect. This counterintuitive finding suggests these attacks may benefit from intensity-based camouflage effects.

The AUC range analysis reveals coordinated thermocouple manipulation exhibits the highest variability across tiers (range = 1.000), while multi-site coordinated attacks show the most stable performance (range = 0.223). These findings indicate that attack parameter scaling methodology should account for scenario-specific sensitivity patterns rather than assuming uniform tier progression across all attack types.


LSTM Anomaly Detection Performance by Attack Scenario
Prediction Error (MSE) Analysis for Cybersecurity Detection with Control Baseline
Attack Scenario
Cross Facility Data Transplant
Sequential Valve Manipulation
Phantom Valve Operation
Command Feedback Desync
Toggle Storm
Multi Site Coordinated
Delayed Propagation
Sensor Drift With Dropouts
Coordinated Thermocouple Manipulation
Precision Trust Decay
Flow Oscillation
Temperature Spike Recovery
0.035
0.023
0.023
0.023
0.023
0.022
0.022
0.022
0.022
0.021
0.021
0.021
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
Prediction Error (MSE) (Linear Scale)
Control Dataset Baseline (0.019)
Attack Scenario Performance


Autoencoder Performance by Attack Scenario
Prediction Error (MSE) Analysis for Cybersecurity Detection with Control Baseline
Attack Scenario
Toggle Storm
Sensor Drift With Dropouts
Multi Site Coordinated
Phantom Valve Operation
Cross Facility Data Transplant
Command Feedback Desync
Temperature Spike Recovery
Precision Trust Decay
Delayed Propagation
Coordinated Thermocouple Manipulation
Flow Oscillation
Sequential Valve Manipulation
54.159
29.258
27.840
26.064
25.122
22.716
19.195
16.355
14.465
13.017
9.503
7.809
0
10
20
30
40
50
Prediction Error (MSE) (Linear Scale)
Control Dataset Baseline (0.717)
Attack Scenario Performance

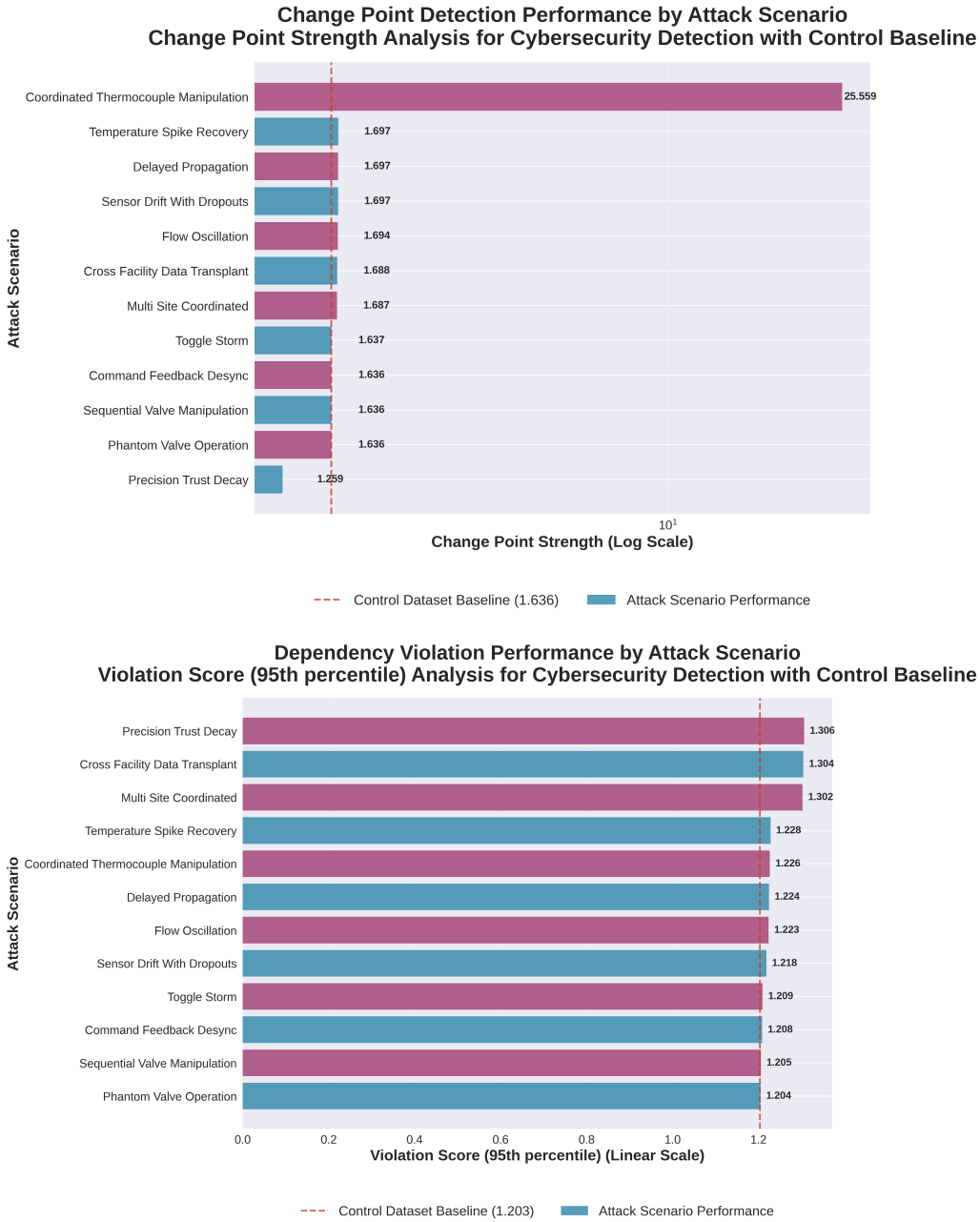
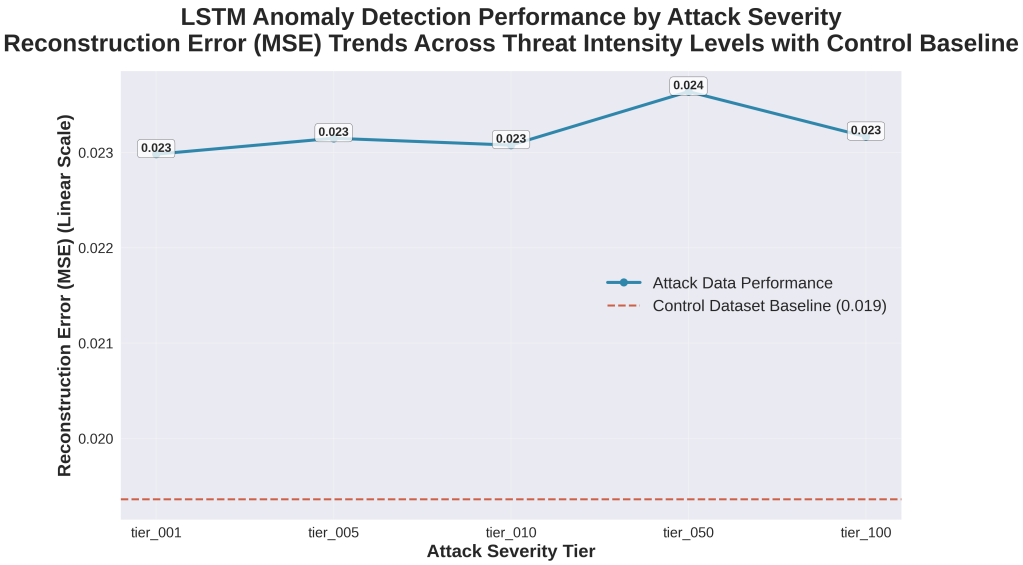
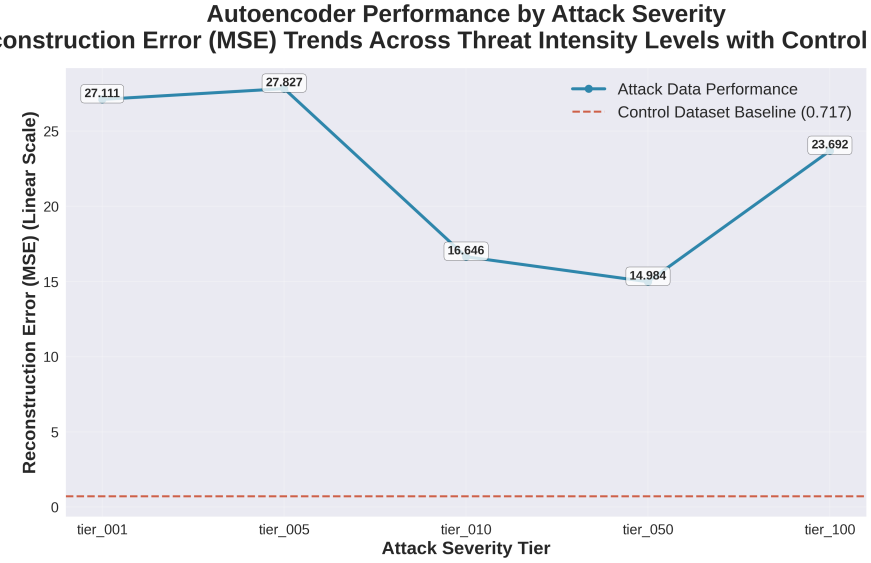


Fig. 7. Performance metrics comparison across attack scenarios for each detection paradigm. Each subplot shows paradigm-specific metric distributions grouped by attack scenario type.


LSTM Anomaly Detection Performance by Attack Severity
Reconstruction Error (MSE) Trends Across Threat Intensity Levels with Control Baseline
Reconstruction Error (MSE) (Linear Scale)
0.023
0.022
0.021
0.020
0.023
0.023
0.023
0.024
0.023
Attack Data Performance
Control Dataset Baseline (0.019)
tier_001
tier_005
tier_010
tier_050
tier_100
Attack Severity Tier


Autoencoder Performance by Attack Severity
Reconstruction Error (MSE) Trends Across Threat Intensity Levels with Control Baseline
Reconstruction Error (MSE) (Linear Scale)
25
20
15
10
5
0
27.111
27.827
16.646
14.984
23.692
Attack Data Performance
Control Dataset Baseline (0.717)
tier_001
tier_005
tier_010
tier_050
tier_100
Attack Severity Tier

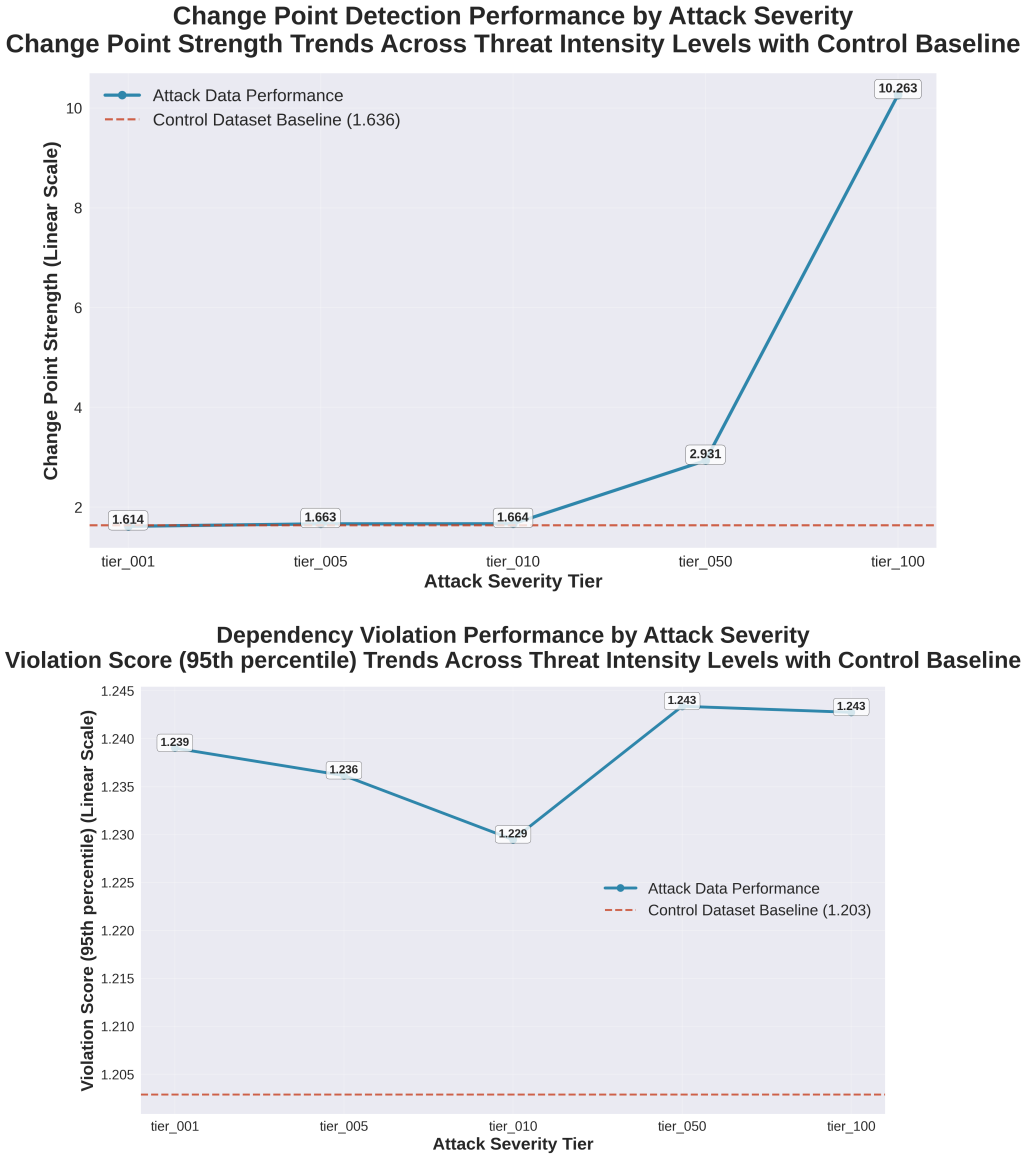


Fig. 7. Performance metrics comparison across attack severity tiers for each detection paradigm. Each subplot shows paradigm-specific metric distributions grouped by attack intensity level.

9 A Physics-Based Approach for Cybersecurity

Cyber-physical systems underpin modern critical infrastructure including power grids, nuclear plants, and industrial facilities. Attackers increasingly exploit the cyber layer to influence the physical layer, issuing deceptive commands or forging sensor data while leaving traditional IT logs unaltered. Purely signature-based or black-box machine learning detectors struggle in this setting for two main reasons: first, stealthy adversaries can mimic normal data distributions, rendering pattern matching approaches ineffective until damage is done; second, most data-driven methods offer limited insight into how or where manipulation occurs, hampering rapid recovery.

Physics-based models close this gap by importing immutable physical laws—conservation of mass, momentum, and energy—into the detection loop. These laws function as intrinsic "ground truth" that adversaries cannot rewrite: if a forged sensor stream violates Newton's or Kirchhoff's equations, the inconsistency is mathematically detectable. Because physical laws couple distant subsystems, even localized tampering produces global violations, granting system-wide observability. Finally, residuals derived from first-principles equations provide actionable hypotheses about disrupted components or physical balances, giving engineers specific diagnostic insights instead of opaque anomaly scores.

9.1 Literature Review

A growing body of research explores the intersection of cybersecurity and physical systems within Cyber-Physical Systems (CPS), addressing two distinct but complementary areas: *physics-based attacks* that exploit physical system behaviors to achieve malicious objectives, and *physics-based detection* methods that leverage physical laws and relationships to identify cyber threats. Understanding this distinction is crucial for developing comprehensive cybersecurity strategies that account for both attack vectors and defensive capabilities in critical infrastructure systems.

9.1.1 Physics-Based Attacks. Physics-based attacks represent a sophisticated class of cyber threats that specifically target the physical processes and constraints of cyber-physical systems. These attacks exploit the inherent coupling between digital control systems and physical processes to achieve objectives that purely digital attacks cannot accomplish.

From a control-theoretic perspective, Pasqualetti et al. [23] use the Stuxnet attack as a prime example of a cyberattack with severe physical consequences. Their work examines "deception attacks," where an adversary compromises the integrity of sensor measurements and actuator commands to manipulate physical processes while remaining undetected by traditional cybersecurity measures. This work demonstrates how attackers can leverage

knowledge of system dynamics to craft attacks that respect certain physical constraints while violating others, making detection particularly challenging.

Lanotte et al. [13] provide a mathematical framework for analyzing physics-based attacks using formal methods. They introduce process calculus designed to model the complex interplay between continuous physical processes and discrete cyber controls. This formal approach enables precise specification and analysis of sophisticated attack scenarios, such as Man-in-the-Middle attacks targeting physical sensors and actuators to drive systems into unsafe states while maintaining plausible system behavior. Their work demonstrates how formal verification techniques can be adapted to analyze the security properties of cyber-physical systems under various attack models.

These works underscore the critical vulnerability of systems that rely on networked communication and digital control of physical processes, highlighting the need for security strategies that account for the unique attack surface presented by the cyber-physical interface.

9.1.2 Physics-Based Detection. In response to the threat landscape outlined above, researchers have developed physics-based detection methods that leverage the immutable nature of physical laws as a foundation for cybersecurity. The core principle underlying this approach is that while digital data can be falsified, the underlying physical laws governing a system provide a ground truth that is significantly more difficult for adversaries to violate or circumvent.

A comprehensive survey by Giraldo et al. [7] studies the field of physics-based attack detection. The authors review methods that create time-series models based on the expected physical behavior of systems, incorporating domain-specific knowledge such as fluid dynamics in water treatment plants or electromagnetic principles in power grids. By comparing these model-based predictions against real-time sensor measurements, security monitors can identify anomalies indicative of manipulated control commands or falsified sensor readings. The survey's key contribution is its unified taxonomy that brings together disparate research from control theory, information security, and power systems engineering, providing a comprehensive framework for understanding how physical relationships can be leveraged for cybersecurity purposes.

The effectiveness of physics-based detection methods stems from their ability to establish analytical redundancy relations that must hold under normal operating conditions. When these relations are violated persistently, it indicates potential sensor manipulation, component failure, or malicious interference. This approach provides several advantages over purely data-driven detection methods: it offers interpretable results grounded in physical

understanding, maintains effectiveness even when historical attack data is limited, and can detect novel attack patterns that deviate from known physical behaviors.

Despite these foundational contributions and demonstrated effectiveness, existing physics-based detection approaches face several critical limitations when deployed in complex operational environments like advanced nuclear reactors. First, scalability challenges emerge when systems have incomplete instrumentation or sparse sensor coverage, limiting the ability to establish analytical redundancy relations. Second, distinguishing between legitimate system faults and malicious cyber attacks remains difficult, as both can manifest as violations of expected physical relationships. Third, most existing approaches provide limited explainable diagnostic reasoning capabilities, offering binary detection results without detailed forensic analysis of root causes or attack propagation pathways. Additionally, insufficient probabilistic reasoning frameworks limit the ability to quantify uncertainty in detection decisions, particularly important when operating under varying environmental conditions or system configurations. Finally, real-time implementation challenges arise from computational complexity and the need to balance detection sensitivity with false alarm rates in mission-critical environments.

These limitations highlight the need for more sophisticated physics-based diagnostic platforms that can provide analytical redundancy, explainable reasoning, and robust uncertainty quantification for complex cyber-physical systems.

9.2 PRO-AID: A Physics-Driven Diagnostic Platform

To address the limitations of existing physics-based detection methods, we evaluated a pre-existing physics-based diagnostic workflow, Parameter-Free Reasoning Operator for Automated Identification and Diagnosis (PRO-AID), specifically designed to overcome scalability, explainability, and uncertainty quantification challenges in complex operational environments. PRO-AID advances beyond traditional approaches by providing analytical redundancy through virtual sensor networks, enabling detailed forensic analysis through explainable diagnostic reasoning, and incorporating probabilistic frameworks for robust uncertainty quantification. This platform derives residuals from first-principles physical relationships and uses them for explainable fault diagnosis, complementing our data-driven detectors by encoding invariants that adversaries cannot easily spoof while providing the diagnostic depth necessary for operational decision-making in critical infrastructure environments. Building on these enhanced capabilities, PRO-AID implements model-based diagnostics through three tightly integrated mechanisms that collectively enable robust fault detection [22] and cyber-attack identification. The platform addresses the inherent challenge that real-world plants rarely instrument every variable needed for complete

physical closure while maintaining rigorous adherence to conservation laws and physical constraints.

9.2.1 Virtual Sensors and Model Coverage. Real-world industrial plants rarely instrument every variable needed for complete physical closure, creating observability gaps that can be exploited by adversaries or complicate fault diagnosis. PRO-AID compensates for this limitation by creating virtual sensors (VS): unmeasured variables that are inferred analytically from neighboring measurements and conservation laws [21]. For example, an unmeasured branch flow can be back-calculated from nodal mass balances and measured inflows and outflows at connected junctions. This approach allows VS to extend diagnostic reach without requiring additional hardware while remaining firmly anchored in physical principles.

The validity of VS estimates is continuously reassessed through cross-validation with multiple analytical redundancy relations. If a residual analysis implicates the equations used to create a particular VS, the platform flags that estimate as unreliable and adjusts its confidence accordingly. This self-monitoring capability ensures that VSs do not propagate errors through the diagnostic chain and maintains the integrity of the overall diagnostic framework.

VSs serve multiple critical functions in the cybersecurity context. First, they expand the observational coverage of the system, making it more difficult for attackers to manipulate measurements without creating detectable inconsistencies across the augmented sensor network. Second, they provide independent estimates of key system variables that can be compared against actual measurements to identify potential tampering. Finally, they enable the platform to maintain diagnostic capability even when some physical sensors are compromised or offline.

9.2.2 Analytical Redundancy Relations (ARRs). Starting from fundamental mass, momentum, and energy balances at the component level, PRO-AID assembles a library of algebraic and differential equations that must hold under fault-free operation. Each analytical redundancy relation (ARR) ties together a subset of sensor measurements—including pressures, flows, temperatures, and electrical currents—spanning one or more system components. These relations capture the essential physical constraints that govern system behavior and provide mathematical expressions of the immutable laws that adversaries cannot circumvent.

PRO-AID evaluates every ARR in real-time and computes the associated residual, defined as the numerical difference between the theoretical predictions based on physical laws and the actual sensor readings or VS values. This continuous evaluation process

operates online during normal plant operation, providing immediate feedback about the consistency of measurements with underlying physical principles. A residual that persistently exceeds its established uncertainty band signals a violation of at least one underlying assumption, whether related to sensor integrity, component health, or the presence of malicious manipulation.

The ARR framework provides several advantages for cybersecurity applications. The distributed nature of the equations means that even localized attacks often produce detectable violations across multiple relations, preventing adversaries from maintaining perfect consistency across all physical constraints. Additionally, the explicit mathematical foundation of each ARR enables precise identification of which physical principles have been violated, supporting both diagnostic accuracy and forensic analysis of potential attacks.

9.2.3 Probabilistic Fault Mapping and Explainability. A single ARR violation often implicates multiple competing fault hypotheses, ranging from sensor bias and valve stiction to heat exchanger fouling or malicious data injection. To resolve this ambiguity, PRO-AID embeds a Bayesian reasoning layer that combines patterns of residuals over time to rank the probability of competing explanations [21]. This probabilistic approach accounts for the inherent uncertainty in both measurements and model predictions while providing quantitative confidence estimates for different diagnostic hypotheses.

The Bayesian framework leverages temporal patterns in residual behavior to distinguish between different types of anomalies. For example, gradual component degradation typically produces slowly evolving residual patterns that affect physically related variables in a coordinated manner. In contrast, cyber attacks often introduce abrupt changes or create inconsistencies that violate expected correlations between measurements. By analyzing these patterns, the system can differentiate between legitimate physical phenomena and malicious interference.

Because every link in this inference chain arises from an explicit physical equation, operators can traverse the diagnostic logic in either direction to understand and validate the reasoning process. Forward chaining allows operators to verify that a proposed diagnosis explains the observed residual patterns, while backward chaining identifies which additional sensors or measurements would best discriminate between remaining hypotheses. This explainability is crucial for high-stakes environments where operators must understand and trust the diagnostic conclusions before taking corrective action.

The probabilistic mapping also supports dynamic adaptation as new evidence becomes available. As additional measurements are collected or system conditions change, the Bayesian framework updates the probability distributions for different fault hypotheses,

providing an evolving assessment of the most likely explanations for observed anomalies. This capability is particularly valuable for distinguishing sophisticated attacks that may initially appear similar to normal operational variations or component degradation.

9.3 Regime Classification

The outputs from the PRO-AID diagnostic engine—including ARR residuals, VS estimates, and probabilistic fault mappings—enable classification of system behavior into three distinct regimes based on adherence to physical laws and control logic:

- Physical: measurements align with physical laws, exhibit temporal continuity, and are consistent with control logic and setpoint changes.
- Mixed: changes arise from genuine physical phenomena (sudden faults or gradual degradation) and propagate coherently across physically related sensors.
- Unphysical: violations of physical laws and/or control-logic consistency indicative of sensor manipulation, control manipulation, or both. High-dimensional correlations across many sensors help surface coordinated tampering.

9.4 Attack Scenarios in METL

Building on the established taxonomy and considering the critical functionalities of the cold trap within the METL system, we have identified two scenarios that could impact its essential operations. The system layout of the cold trap loop with the attack scenarios is shown in Fig. 8. Each attack scenario highlights the (process) controlled variable and the manipulated variable, with a arrow pointing from controlled variable to manipulated variable. Here, the controlled variable is defined as the process variable that the control system aims to maintain at a set point; the manipulated variable refers to a variable being adjusted or manipulated to influence the process variable and bring it to the set point.

(1) Attack scenario: cold trap fan motor
The first scenario focuses on control the air blower used to cool the cold trap. This air blower provides the cooling airflow to cool down the sodium in cold trap, as shown at the bottom of the 8. The air blower's fan motor is PID controlled by the lowest temperature readings from 24 thermocouples embedded within the cold trap.
The PID controller adjusts the blower's operation based on a user-defined target temperature, ensuring the lowest temperature inside the cold trap reaches this set point. In this scenario, the control variable is the lowest temperature reading within the cold trap, while the manipulated variable is the blower's airflow rate.

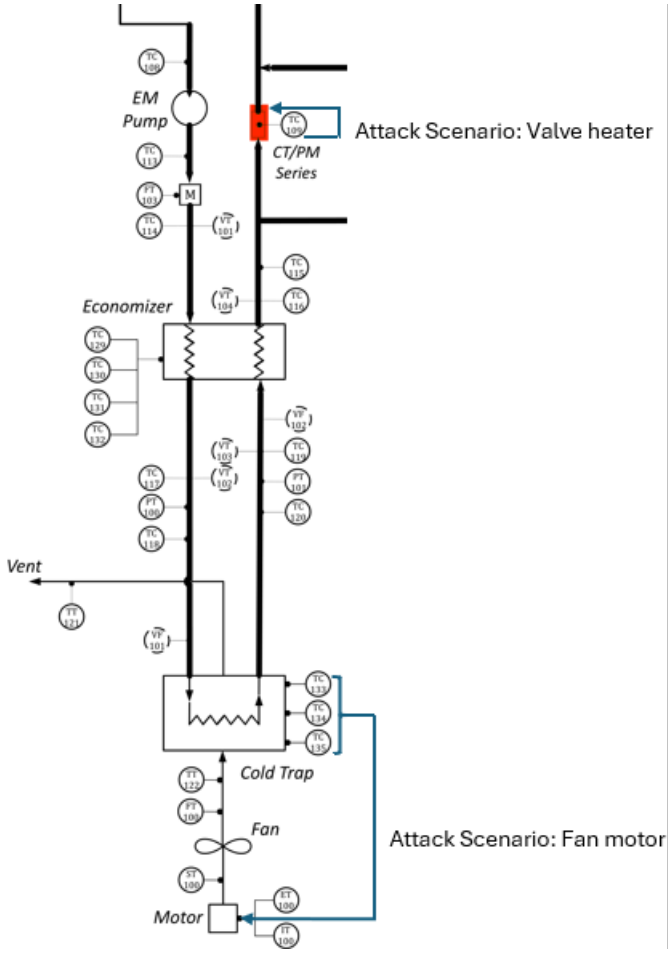


Fig. 8. System Layout of Cold Trap (CT) with Attack Scenarios

When functioning without any attack or disruption, the blower operates correctly, allowing the cold trap to effectively capture impurities and maintain them within the desired levels.

(2) Attack scenario a valve heater in the cold trap loop
In the cold trap loop, there are 16 heaters installed along piping and valves.In this scenario, a heater located at the CT/PM valve is selected as the target component to perform attacks, as indicated by the red-highlighted location in 8. The heater is PID controlled by the temperature reading from a thermocouple located at the valve. Its primary function is to heat the valve and the sodium within it, ensuring it remains in liquid form and reaches the designated set point temperature.

The PID controller regulates the heater's current based on a user-defined target temperature, maintaining the liquid sodium at the desired temperature as it flows through the valve. Here, the control variable is the temperature measurement at the valve, while the manipulated variable is the heater's current. When functioning without any attack or disruption, the heater heats up the liquid sodium, preventing low-temperature liquid from re-entering the main loop, which helps to mitigate temperature fluctuations and avoid thermal stratification.

The controlled variables and the manipulated variables of two scenarios have been collected by the Data Pipeline described in 4, and have been converted into time series on a daily basis. The data from METL now have been analyzed and prepared for carrying out different types of attacks towards the two scenarios.

9.5 Workflow

The physics-based cyber defense workflow leverages the PRO-AID diagnostic capabilities as its core analytical engine, with VS generation, ARR computation, and Bayesian fault reasoning components directly supporting the feature extraction and classification requirements. Figure 9 provides a schematic overview of this three-stage workflow, illustrating the decision tree structure and interconnections between stages. The end-to-end workflow proceeds through three integrated stages.

9.5.1 Stage 1: Problem Scoping. As illustrated in the "Problem Scoping" section of Figure 9, real-time sensor measurements undergo initial screening through conventional data-driven anomaly detection algorithms to filter statistically significant deviations. This initial screening identifies critical functions—essential operational processes or safety systems that require enhanced monitoring due to their impact on plant safety and operational integrity. Sensor streams that satisfy preliminary anomaly thresholds and relate to these critical functions are subsequently processed by the PRO-AID based diagnostic platform to generate a feature vector comprising:

- **VS estimates**: Unmeasured system variables inferred analytically from neighboring measurements and conservation laws (e.g., unmeasured branch flow back-calculated from nodal mass balances and measured inflows/outflows). VS validity is continuously reassessed—if residuals implicate the equations used to create a VS, the platform flags its estimate as unreliable, extending diagnostic reach without new hardware while remaining anchored in physics.
- **ARR residuals**: Starting from mass, momentum, and energy balances at the component level, PRO-AID assembles algebraic and differential equations that must

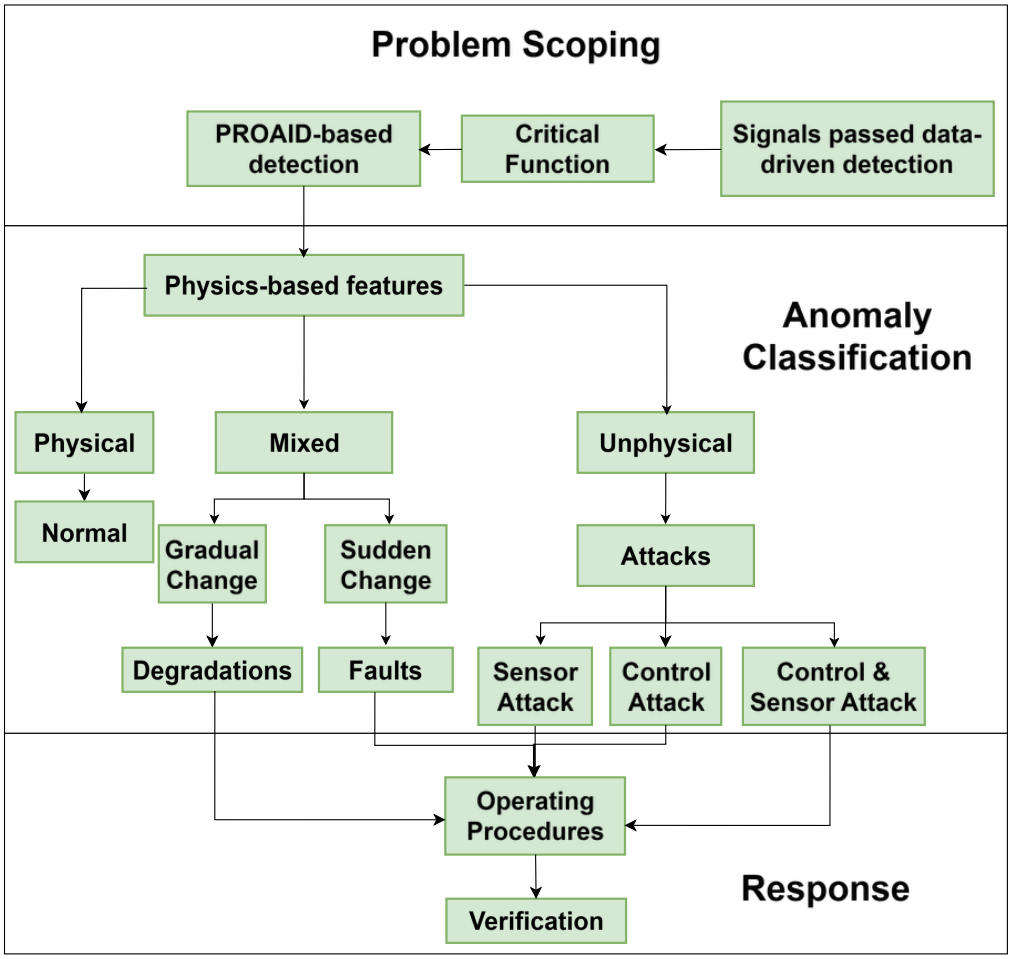


Fig. 9. Physics-based cyber-security framework showing the three-stage workflow: Problem Scoping (feature extraction), Anomaly Classification (regime determination), and Response (corrective actions).

hold under fault-free operation. ARR residuals represent the numerical difference between model estimations and actual sensor readings or VS values, computed in real-time. Residuals exceeding uncertainty bands signal violations of underlying assumptions—sensor integrity, component health, or cyber manipulation.

- **Fault-type mappings**: Residual signature analysis linking specific ARR violation patterns to candidate fault modes through pre-established diagnostic relationships.
- **Bayesian fault posterior probabilities**: A probabilistic reasoning layer that combines patterns of residuals over time to rank the probability of competing explanations—sensor bias, valve stiction, heat exchanger fouling, or malicious data injection.

This enables forward chaining (validating proposed diagnoses against observed residuals) and backward chaining (identifying additional sensors to discriminate among remaining hypotheses).

This physics-informed feature set provides the foundation for subsequent anomaly classification by encoding both local sensor behavior and system-wide physical consistency, with particular emphasis on preserving critical function integrity.

9.5.2 Stage 2: Anomaly Classification. The "Anomaly Classification" stage shown in Figure 9 processes the physics-based feature vector through a systematic multi-criteria classification framework. System behavior is categorized into one of three primary operational regimes based on adherence to fundamental physical laws, temporal continuity, and control logic consistency:

- **Physical regime**: All sensor readings are consistent with physical laws, exhibit temporal continuity, and align with control logic and operational changes. This reflects normal operating state, free from both faults and cyber attacks.
- **Mixed regime**: Physical laws are satisfied but temporal patterns deviate from expectations. Changes originate from genuine physical phenomena and propagate coherently across all physically related sensor signals—this coordinated response is the key characteristic differentiating true faults from cyber attacks. Subdivided into:
  - *Gradual change*: Progressive component degradation with consistent physical relationships maintained across all affected sensors
  - *Sudden change*: Abrupt faults in sensors, controllers, or actuators affecting system behavior while maintaining physical consistency and correlations
- **Unphysical regime**: Violations of physical laws and/or control logic indicating potential cyber attacks, with inconsistent or uncorrelated sensor behaviors that distinguish them from physical faults. Categorized as:
  - *Sensor attack*: Manipulated sensor readings violate physical constraints while controllers still respond consistently to the (compromised) measurements, creating detectable mismatches between physical behavior and control actions (e.g., falsified temperature readings causing incorrect heater responses).
  - *Control attack*: Compromised control logic or actuators operate inconsistently with reported system state—for example, a heater may remain on even when the reported temperature exceeds the set point, identifiable through control logic verification.
  - *Combined attack*: Coordinated manipulation of both sensors and control systems designed to maintain apparent consistency. Detection exploits high-dimensional

correlation networks among sensors and VSs, based on the assumption that attackers possess incomplete knowledge of all sensor relationships and their co-evolution patterns.

This hierarchical classification scheme leverages quantitative assessment of ARR residual magnitudes, VS consistency, control logic verification, and temporal correlation patterns to enable precise discrimination between normal operations, component degradation, and cyber-attack scenarios. Table 5 summarizes the classification criteria and diagnostic outcomes for each regime.

Table 5. Regime classification criteria and diagnostic outcomes.

Regime	Diagnostic Sub-Mode	Physics Consistency	Temporal Continuity	Control-Logic Consistency	Representative Symptoms	Primary Interpretation
Physical	—	✓	✓	✓	All sensor and actuator values are within expected physical bounds and controller logic is followed perfectly.	Normal, fault-free, and secure operation.
Mixed	(1) Sudden-Change Fault	✓	×	× or ✓	A sensor reading suddenly jumps to a new, physically plausible value (e.g., stuck thermostat) causing performance decline.	A component has experienced an abrupt physical failure.
Mixed	(2) Gradual-Change Degradation	✓	✓	× or ✓	System performance slowly degrades over time; sensor values drift but remain physically valid (e.g., a fouling pipe).	A component is wearing out or experiencing age-related degradation.
Unphysical	(1) Sensor Attack	×	× or ✓	× or ✓	A sensor reports a physically impossible value (e.g., negative temperature), or its value is inconsistent with correlated sensors.	Malicious data injection targeting a sensor measurement.
Unphysical	(2) Actuator / Control Attack	×	×	× or ✓	The system state violates physical laws because an actuator's command and its physical effect do not match (e.g., valve is commanded closed but flow increases).	A controller or actuator has been hijacked by a malicious actor.

9.5.3 Stage 3: Response. The "Response" stage illustrated in Figure 9 translates regime classifications into actionable operational guidance. Following regime classification and fault localization, the framework consults plant-specific operating procedures and emergency response protocols to generate prioritized action recommendations. The systematic response process includes:

Action Generation: Regime-specific response protocols are activated based on classification results. Physical regime classifications maintain routine monitoring with enhanced surveillance of marginal parameters. Mixed regime detections (degradations and faults)

trigger component-specific diagnostic procedures, including sensor validation protocols and predictive maintenance scheduling. Unphysical regime classifications (sensor, control, or combined attacks) initiate cyber-incident response procedures:

- Suspect sensor isolation and cross-validation with redundant measurements
- Control system integrity verification through independent channels
- Fallback controller activation where appropriate and available
- Security incident documentation and reporting to designated authorities

Verification: As shown in the figure, all recommended actions undergo systematic verification to confirm their effectiveness and appropriateness. This verification process includes post-action monitoring to assess whether the implemented response successfully addresses the identified condition, validation that corrective actions do not introduce new anomalies or safety concerns, and documentation of outcomes for continuous improvement.

All recommendations include confidence intervals and alternative actions ranked by implementation complexity and potential operational impact. Verified outcomes are logged to continuously refine uncertainty bands, detection thresholds, and response procedures, creating a self-improving diagnostic capability.

9.6 Application to METL Attack Scenario

To demonstrate the application of the proposed framework, we present several cyber-attack scenarios targeting the economizer in METL. The economizer is a sodium-cooled heat exchanger that preheats the return flow from the cold trap. Figure 10 shows the relevant section of the METL layout and associated sensor locations.

Under normal conditions, hot sodium from the primary loop passes through the economizer, releasing heat before entering the cold trap, where impurities are removed. The cooled sodium then returns through the economizer to regain heat before rejoining the primary loop. In the attack scenarios, selected sensor signals and actuator commands are manipulated, allowing us to evaluate how the physics-based framework detects and distinguishes cyber intrusions from normal process variations.

9.6.1 Single sensor attack scenario. In this attack scenario, the target is thermocouple 117 (TC117), located in the piping between the economizer and the cold trap, as shown in Figure 10. In the METL system, each pipe section is equipped with a heater that activates when the temperature drops below a predefined set point to prevent sodium from freezing.

The attack introduces a constant bias of 5°C to TC117, along with increased signal noise, as illustrated in Figure 11. As a result, the manipulated temperature signal exceeds the 250°C set point, which should cause the heater to turn off during that period. However, the

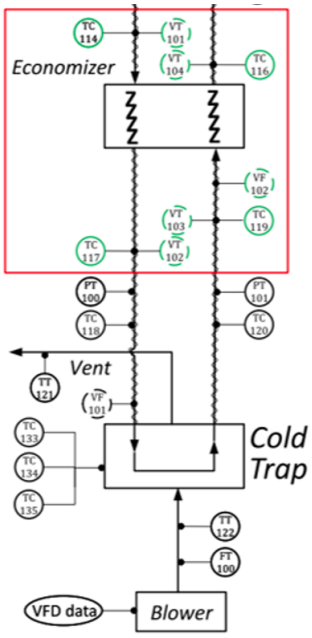


Fig. 10. METL layout with sensor locations showing the economizer and associated measurement points.

control response remains inconsistent with this behavior, creating a detectable mismatch between reported sensor readings and expected control actions.

To validate the attack detection capability, we apply a physics-based check using the log mean temperature difference (LMTD), which characterizes expected thermal behavior across the heat exchanger. The LMTD is calculated as:

$$\Delta T_{LMTD} = \frac{\Delta T_1 - \Delta T_2}{\ln\left(\frac{\Delta T_1}{\Delta T_2}\right)} = \frac{(T_{h,in} - T_{c,out}) - (T_{h,out} - T_{c,in})}{\ln\left(\frac{T_{h,in} - T_{c,out}}{T_{h,out} - T_{c,in}}\right)} \tag{1}$$

where $T_{h,in}$ is the inlet temperature of the hot fluid (hot sodium entering the economizer), $T_{h,out}$ is the outlet temperature of the hot fluid, $T_{c,in}$ is the inlet temperature of the cold fluid, and $T_{c,out}$ is the outlet temperature of the cold fluid.

This physics-based validation reveals discrepancies between the measured and physically consistent temperature behavior, enabling successful attack detection through fundamental thermodynamic principles.

9.6.2 Comparison to sensor fault scenario. To demonstrate the framework's ability to distinguish sensor faults from sensor attacks, Figure 12 shows a representative case involving PRO-AID analysis. In this scenario, TC117 experiences a sensor fault that introduces a

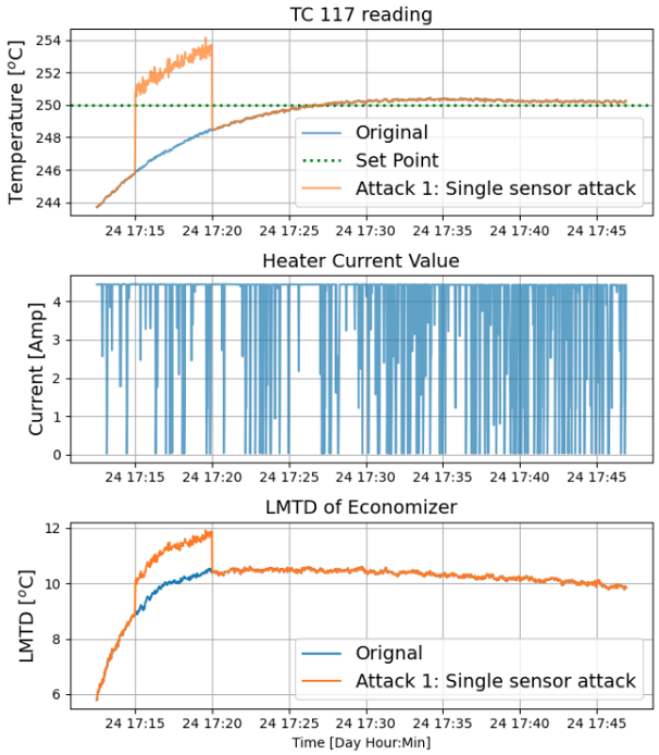


Fig. 11. Single sensor attack targeting TC117: temperature bias introduction, heater current response, and LMTD validation showing attack detection through physics-based consistency checks.

constant bias. Due to the underlying physical relationships, several VS (VS1, VS6, VS8, and VS9) exhibit corresponding changes that maintain physical consistency.

The key distinction lies in the system's response pattern. In the case of a sensor fault, the biased reading is still processed by the controller as if it were correct, resulting in control behavior that remains consistent with the reported (though incorrect) temperature. This coherent response across all physically related measurements distinguishes genuine sensor degradation from malicious manipulation, where such coordinated responses are typically absent.

In contrast, a malicious attack on TC117 introduces a localized data anomaly that violates the system's underlying physical conservation laws. The injected non-physical value creates immediate conflicts between the manipulated measurement and expected behavior from related sensors, manifesting as high-magnitude residuals in specific ARRs while other ARRs remain consistent. This pattern of localized ARR failures amidst system-wide

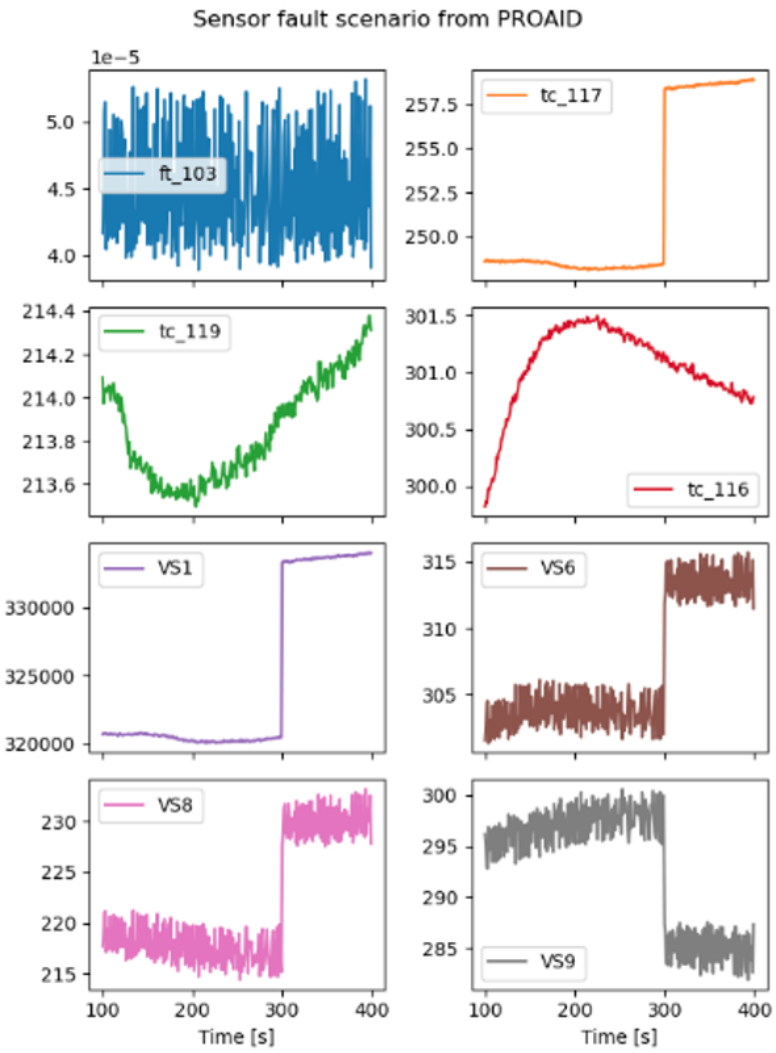


Fig. 12. Sensor fault scenario analyzed by PRO-AID showing correlated changes in VSs (VS1, VS6, VS8, VS9) and temperature measurements that maintain physical consistency, distinguishing legitimate faults from malicious attacks.

physical consistency would create a distinct detection signature that distinguishes targeted cyber-attacks from naturally occurring sensor degradation.

9.6.3 Multi-sensor attack scenario. A more sophisticated form of cyber-attack involves a stealthy strategy where multiple correlated sensors are simultaneously manipulated to evade detection by single-sensor physics-based checks. In this scenario, TC117 and TC119 are targeted with coordinated linear drift and added noise, as shown in Figure 13.

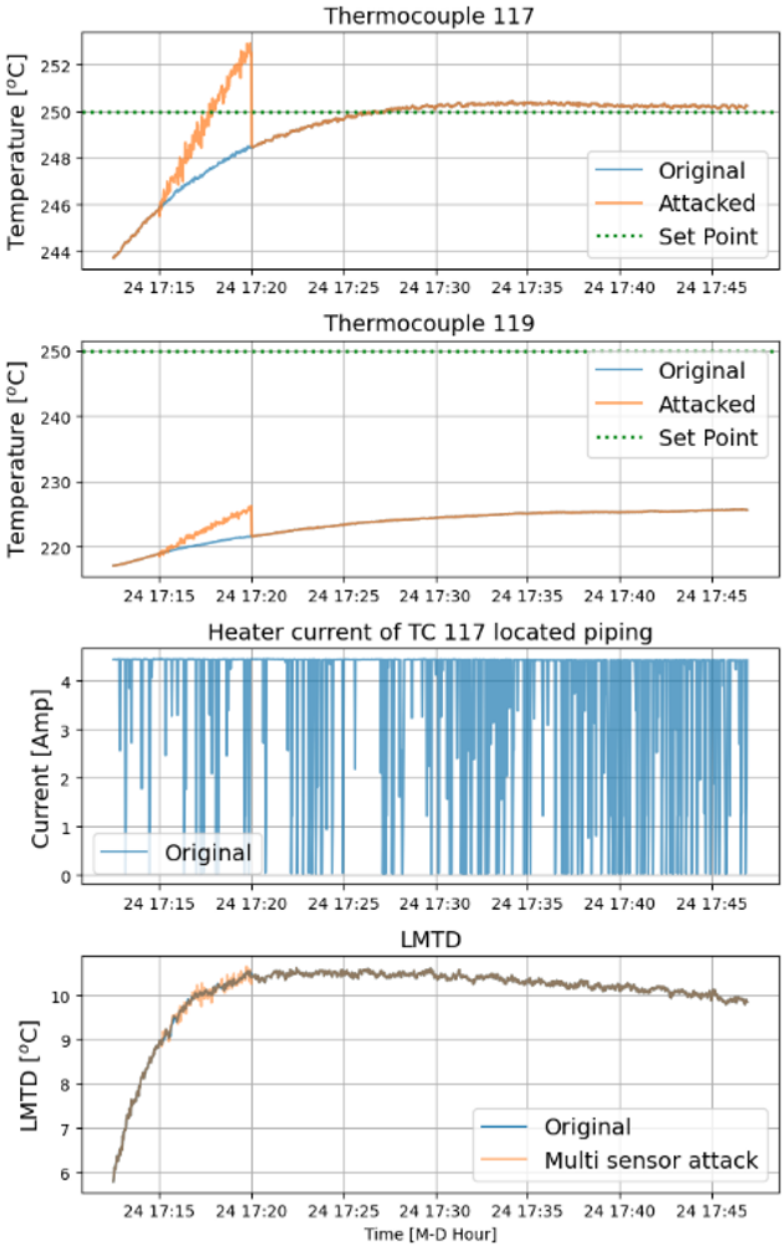


Fig. 13. Multi-sensor attack scenario targeting TC117 and TC119 with coordinated manipulation, showing how physics-based checks can still detect inconsistencies despite sophisticated coordination attempts.

While one of the physical indicators—specifically the LMTD—remains relatively unaffected by this coordinated manipulation, the attack still disrupts other physical relationships. This would lead to inconsistencies in VS estimates and violations of ARRs,

demonstrating the robustness of the physics-based approach against sophisticated multi-sensor attacks.

Additionally, a control logic inconsistency is observed: the heater associated with the pipe segment monitored by TC117 remains active even when the reported temperature significantly exceeds the set point. This behavior violates expected control logic and further supports the detection of an attack, as such inconsistencies are not expected under normal operating conditions or legitimate sensor faults.

9.7 Discussion

Limitations and Performance Bounds. It is important to note that these results represent a lower bound for the potential performance of each detection paradigm in this domain. Our evaluation employed minimal hyperparameter tuning, using largely default or basic configurations for all machine learning models. More extensive hyperparameter optimization, architectural modifications, or ensemble approaches could significantly improve detection performance. Additionally, our attack severity tiers were not sufficiently aggressive to create clear differentiation between intensity levels, suggesting that future work should explore more pronounced attack magnitudes to better evaluate paradigm sensitivity and operational thresholds.

The physics-based cyber defense architecture provides a critical bridge between classical fault-diagnosis principles and modern cybersecurity practices. By explicitly separating physics consistency from control-logic consistency, this framework equips operators with a principled methodology to discriminate stealthy cyber manipulation from routine component degradation and sensor faults—an analytical capability largely absent in black-box anomaly detection approaches.

The demonstrated scenarios illustrate three key advantages of the physics-based approach: (1) **Attack-fault discrimination** through analysis of physical correlation patterns, where legitimate faults produce coherent cross-sensor responses while attacks typically create inconsistent signatures; (2) **Multi-layer validation** combining thermodynamic principles (LMTD), VS consistency, and control logic verification to provide robust detection even against sophisticated coordinated attacks; and (3) **Explainable diagnostics** that identify specific physical principles violated and affected components, enabling rapid operator response and system recovery.

With appropriate quantitative calibration and empirical validation, this methodology scales effectively to safety-critical industrial environments. The same residual computation engine that supports cyber defense can simultaneously feed predictive maintenance

programs, unifying traditionally separate operational silos and maximizing the value of physics-based system understanding.

10 Conclusions and Deliverables

This work delivers a practical reference architecture and an end-to-end evaluation stack for AI-based anomaly detection in AR OT environments using METL data. We implemented a reproducible attack emulation pipeline, standardized preprocessing with a justified 30-second downsampling strategy, four complementary detection paradigms, and a results aggregation workflow with uncertainty estimates. Key constraints and lessons learned—especially the H1 (streaming) to H2 (offline) pivot—inform operational deployment paths.

Deliverables:

- Final report (this document) and architecture/data-flow figures.
- Source code repository for the data proxy, transformations, and ML experiments (with environment files and Docker/compose where provided).
- Reproducible experiment configurations, logs, and aggregated metrics.

11 Future Work

Future work on this line of effort is intended to include several focused directions to advance from the current evaluation stack to operational autonomy:

- **Physics model in the loop paradigm**: Implement physics-based constraints and validation layers within the detection pipeline to improve discrimination between legitimate system behavior and malicious manipulation, building on the foundation established in this work.
- **Hyperparameter optimization and model enhancement**: Conduct systematic hyperparameter tuning for all detection paradigms, explore ensemble methods, and evaluate additional ML architectures to realize the full potential performance suggested by our baseline results.
- **Enhanced attack scenario development**: Design more aggressive attack severity tiers with clearer differentiation between intensity levels, expand the attack catalog with additional attack vectors, and develop more sophisticated coordinated attack patterns that better challenge detection capabilities.
- **Incorporation of SCADA traffic analysis**: Incorporate capture network traffic corresponding to sensor/setpoint values as additional features in the machine learning models.

- **Live red teaming**: Conduct controlled red-team exercises on METL or similar AR/SMR facilities using developed standard methodologies; compare with emulation results to calibrate realism, refine attack catalogs, and harden detectors and operator workflows.

Glossary

- Autoencoder (AE): A neural network trained to reconstruct its input; reconstruction error indicates anomalies.
- Bootstrapping (statistics): Resampling technique to estimate uncertainty (e.g., confidence intervals) of metrics.
- Change point: A time at which the statistical properties of a sequence change.
- CompactRIO (cRIO): National Instruments (NI) real-time controller used in industrial control and data acquisition.
- Content-addressable cache: A cache keyed by a hash of inputs/configs so identical jobs reuse the same trained model or artifact.
- Dual-path: Train exclusively on clean data; test on attack data to prevent leakage.
- False positives per hour (FP/hr): Number of non-attack alerts raised per hour of normal operation.
- Grafana: Visualization and dashboarding platform used to monitor metrics and alerts.
- Granger causality: A statistical test indicating whether past values of one series help predict another.
- HMI: Human–Machine Interface; operator-facing screens and controls.
- I&C: Instrumentation and Control; plant measurement and control systems.
- InfluxDB: Time-series database used to store and query historical sensor data.
- IT/OT: Information Technology (business/enterprise systems) vs. Operational Technology (industrial control systems).
- LSTM: Long Short-Term Memory; a recurrent neural network architecture for sequence modeling.
- MODBUS: An industrial communication protocol commonly used with PLCs.
- MQTT: Lightweight publish–subscribe messaging protocol used for streaming data.
- PRO-AID: A pre-existing physics-based diagnostic workflow using Virtual Sensors and Analytical Redundancy Relations (evaluated on one scenario/severity).
- N-of-M consensus: An alerting rule that triggers when N out of M parallel detectors/sources exceed threshold within a window.
- Persistence window: Require the score to exceed threshold for N consecutive samples before declaring an alert.
- PLC: Programmable Logic Controller; ruggedized industrial computer for control tasks.

- PR-AUC/AUROC: Area under precision–recall / receiver operating characteristic curves.
- Prometheus: Metrics collection and alerting system often paired with Grafana.
- Random Forest (RF): An ensemble of decision trees used for classification or regression.
- Safety Instrumented System (SIS): Dedicated safety functions (e.g., shutdown) independent of basic control.
- Virtual Sensor (VS): An inferred, unmeasured variable derived from conservation laws and neighboring measurements.
- Analytical Redundancy Relation (ARR): A physical equation linking multiple measurements that should hold in normal operation.
- Anomaly detection: Machine learning approach to identify patterns that deviate significantly from normal operational behavior.
- AR (Advanced Reactor): Next-generation nuclear reactor designs including small modular reactors and microreactors.
- Attack scenario: A specific type of cybersecurity attack pattern targeting particular sensors or systems (e.g., thermocouple manipulation, valve operation disruption).
- AUC (Area Under Curve): A single metric summarizing ROC curve performance; higher values (approaching 1.0) indicate better detection capability.
- Baseline data: Historical normal operation data used as reference for anomaly detection and ROC analysis comparison.
- Correlation matrix: A statistical table showing correlation coefficients between all sensor pairs, used to identify dependency relationships.
- Digital twin: A virtual replica of a physical system that enables real-time simulation and closed-loop testing.
- Downsampling: Reducing data temporal resolution (e.g., from 1-second to 30-second intervals) while preserving key information for analysis.
- Ensemble methods: Machine learning techniques that combine multiple models or algorithms to improve overall performance.
- F1-score: Harmonic mean of precision and recall, providing a balanced performance metric for detection systems.
- Heat exchanger: Thermal system component that transfers heat between fluid streams, monitored by multiple temperature sensors.
- Hyperparameter tuning: Systematic optimization of algorithm configuration parameters to improve model performance.

- LMTD (Log Mean Temperature Difference): Thermodynamic parameter used to characterize heat exchanger performance and detect physical anomalies.
- METL (Mechanisms Engineering Test Loop): Argonne National Laboratory's sodium-cooled reactor testing facility used for this cybersecurity research.
- Precision: Fraction of detected anomalies that are true attacks (true positives / (true positives + false positives)).
- Preprocessing: Data cleaning, filtering, and transformation steps applied before machine learning analysis.
- Primary/secondary loop: Nuclear reactor cooling system architecture with separate heat transfer circuits for safety isolation.
- Recall: Fraction of actual attacks that are successfully detected (true positives / (true positives + false negatives)).
- Reconstruction error: Difference between original input and autoencoder output, used as anomaly indicator.
- ROC curve: Receiver Operating Characteristic plot showing true positive rate vs. false positive rate across detection thresholds.
- Severity tier: Attack intensity level (tier_001 = minimal, tier_100 = maximum) controlling the magnitude of sensor manipulation.
- Signal dilution: Problem where attack signals on few sensors become statistically overwhelmed when averaged across many normal sensors.
- SMR (Small Modular Reactor): Compact nuclear reactor design suitable for distributed power generation and enhanced safety features.
- Telegraf: Agent for collecting and shipping metrics/logs into systems like InfluxDB.
- Thermocouple: Temperature measurement sensor commonly used in industrial systems and frequently targeted in attack scenarios.
- Time-series data: Sensor measurements collected sequentially over time, forming the primary data type for anomaly detection.
- VLAN: Virtual LAN; network segmentation technique to isolate traffic.

References

[1] ALSABBAGH, W., AMOGBONJAYE, S., URREGO, D., AND LANGENDÖRFER, P. A stealthy false command injection attack on modbus based scada systems. In *2023 IEEE 20th Consumer Communications & Networking Conference (CCNC)* (2023), pp. 1–9.

[2] ALVES, T., DAS, R., AND MORRIS, T. Virtualization of industrial control system testbeds for cybersecurity. In *Proceedings of the 2nd Annual Industrial Control System Security Workshop* (New York, NY, USA, 2016), ICSS '16, Association for Computing Machinery, p. 10–14.

[3] CUI, L., QU, Y., GAO, L., XIE, G., AND YU, S. Detecting false data attacks using machine learning techniques in smart grid: A survey. *Journal of Network and Computer Applications 170* (2020), 102808.

[4] DO, V. L., FILLATRE, L., NIKIFOROV, I., AND WILLETT, P. Security of scada systems against cyber–physical attacks. *IEEE Aerospace and Electronic Systems Magazine 32*, 5 (2017), 28–45.

[5] FLAUS, J.-M. A safety/security risk analysis approach of industrial control systems: A cyber bowtie–combining new version of attack tree with bowtie analysis. *Computers & security 72* (2018), 175–195.

[6] GHAEINI, H. R., TIPPENHAUER, N. O., AND ZHOU, J. Zero residual attacks on industrial control systems and stateful countermeasures. In *Proceedings of the 14th International Conference on Availability, Reliability and Security* (New York, NY, USA, 2019), ARES '19, Association for Computing Machinery.

[7] GIRALDO, J., URBINA, D., CARDENAS, A., VALENTE, J., FAISAL, M., RUTHS, J., TIPPENHAUER, N. O., SANDBERG, H., AND CANDELL, R. A survey of physics-based attack detection in cyber-physical systems. *ACM Computing Surveys (CSUR) 51*, 4 (2018), 1–36.

[8] GOEL, S. A systematic literature review on past attack analysis on industrial control systems. *Transactions on Emerging Telecommunications Technologies 35*, 6 (2024), e5004.

[9] HOSEIN AFANDIZADEH ZARGARI, A., ASHRAFIAMIRI, M., SEO, M., MANOJ PUDUKOTAI DINAKARRAO, S., FOUDA, M. E., AND KURDAHI, F. Captive: Constrained adversarial perturbations to thwart ic reverse engineering. *arXiv e-prints* (2021), arXiv–2110.

[10] HU, Y., LI, H., YANG, H., SUN, Y., SUN, L., AND WANG, Z. Detecting stealthy attacks against industrial control systems based on residual skewness analysis. *EURASIP Journal on Wireless Communications and Networking 2019* (2019), 1–14.

[11] HUSSAIN, A., HEIDEMANN, J., AND PAPADOPOULOS, C. A framework for classifying denial of service attacks. In *Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications* (2003), pp. 99–110.

[12] KORKMAZ, E., DAVIS, M., DOLGIKH, A., AND SKORMIN, V. Detection and mitigation of time delay injection attacks on industrial control systems with plcs. In *Computer Network Security: 7th International Conference on Mathematical Methods, Models, and Architectures for Computer Network Security, MMM-ACNS 2017, Warsaw, Poland, August 28-30, 2017, Proceedings 7* (2017), Springer, pp. 62–74.

[13] LANOTTE, R., MERRO, M., MUNTEANU, A., AND VIGANÒ, L. A formal approach to physics-based attacks in cyber-physical systems. *ACM Transactions on Privacy and Security (TOPS) 23*, 1 (2020), 1–41.

[14] LI, W., XIE, L., AND WANG, Z. Two-loop covert attacks against constant value control of industrial control systems. *IEEE Transactions on Industrial Informatics 15*, 2 (2019), 663–676.

[15] LI, Y., HUA, J., WANG, H., CHEN, C., AND LIU, Y. Deeppayload: Black-box backdoor attack on deep learning models through neural payload injection. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)* (2021), IEEE, pp. 263–274.

[16] MCLAUGHLIN, S., KONSTANTINOU, C., WANG, X., DAVI, L., SADEGHI, A.-R., MANIATAKOS, M., AND

KARRI, R. The cybersecurity landscape in industrial control systems. *Proceedings of the IEEE 104*, 5 (2016), 1039–1057.

[17] MIRKOVIC, J., AND REIHER, P. A taxonomy of ddos attack and ddos defense mechanisms. *SIGCOMM Comput. Commun. Rev. 34*, 2 (Apr. 2004), 39–53.

[18] MO, Y., CHABUKSWAR, R., AND SINOPOLI, B. Detecting integrity attacks on scada systems. *IEEE Transactions on Control Systems Technology 22*, 4 (2014), 1396–1407.

[19] MO, Y., AND SINOPOLI, B. False data injection attacks in control systems. In *Preprints of the 1st workshop on Secure Control Systems* (2010), vol. 1.

[20] NEKOUEI, E., PIRANI, M., SANDBERG, H., AND JOHANSSON, K. H. A randomized filtering strategy against inference attacks on active steering control systems. *IEEE Transactions on Information Forensics and Security 17* (2021), 16–27.

[21] NGUYEN, T. N., DOWNAR, T., AND VILIM, R. A probabilistic model-based diagnostic framework for nuclear engineering systems. *Annals of Nuclear Energy 149* (2020), 107767. Publisher: Elsevier.

[22] NGUYEN, T. N., PONCIROLI, R., BRUCK, P., ESSELMAN, T. C., RIGATTI, J. A., AND VILIM, R. B. A digital twin approach to system-level fault detection and diagnosis for improved equipment health monitoring. *Annals of Nuclear Energy 170* (2022), 109002.

[23] PASQUALETTI, F., DÖRFLER, F., AND BULLO, F. Control-theoretic methods for cyberphysical security. *IEEE Control Systems Magazine 33*, 1 (2013), 110–124.

[24] TYCHALAS, D., KELIRIS, A., AND MANIATAKOS, M. Led alert: Supply chain threats for stealthy data exfiltration in industrial control systems. In *2019 IEEE 25th International Symposium on On-Line Testing and Robust System Design (IOLTS)* (2019), pp. 194–199.

[25] URBINA, D. I., GIRALDO, J. A., CARDENAS, A. A., TIPPENHAUER, N. O., VALENTE, J., FAISAL, M., RUTHS, J., CANDELL, R., AND SANDBERG, H. Limiting the impact of stealthy attacks on industrial control systems. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (New York, NY, USA, 2016), CCS '16, Association for Computing Machinery, p. 1092–1105.

[26] YOO, H., AND AHMED, I. Control logic injection attacks on industrial control systems. In *ICT Systems Security and Privacy Protection: 34th IFIP TC 11 International Conference, SEC 2019, Lisbon, Portugal, June 25-27, 2019, Proceedings 34* (2019), Springer, pp. 33–48.

[27] ZHU, Y., YAN, J., TANG, Y., SUN, Y. L., AND HE, H. Resilience analysis of power grids under the sequential attack. *IEEE Transactions on Information Forensics and Security 9*, 12 (2014), 2340–2354.

Strategic Security Sciences Division
9700 South Cass Avenue, Bldg. 221

Nuclear Science and Engineering Division
9700 South Cass Avenue, Bldg. 208

Argonne National Laboratory
Lemont, Illinois 60439

www.anl.gov

U.S. DEPARTMENT of ENERGY

Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC