Review

Not peer-reviewed version

# Intrusion Detection Datasets for IIoT and ICS: A Taxonomic Review with a Decision-Aid Scoring Rubric

Ayman Termanini , Hadj Bourdoucen * , Dawood Al-Abri , Ahmed Al-Maashri

Posted Date: 2 June 2026

doi: 10.20944/preprints202606.0145.v1

Keywords: CPS/IIoT security; ICS security; OT intrusion detection; dataset; MITRE ATT&CK; anomaly detection; machine learning for security

 doi:10.20944/preprints202606.0145.v1



*Review*

# Intrusion Detection Datasets for IIoT and ICS: A Taxonomic Review with a Decision-Aid Scoring Rubric

**Ayman Termanini, Hadj Bourdoucen *, Dawood Al-Abri and Ahmed Al-Maashri**

Department of Electrical and Computer Engineering, Sultan Qaboos University, Muscat, Oman

* Correspondence: hadj@squ.edu.om

**Abstract**

Dataset quality significantly affects the effectiveness of a machine learning (ML) model for an intrusion detection system (IDS) for cyber-physical industrial control systems (CPS/ICS) and Industrial Internet of Things (IIoT). Existing surveys compare datasets qualitatively or along limited dimensions, while this review introduces quantitative documentation and decision-aid scoring across 23 ICS/OT/IIoT datasets. These datasets are analyzed along seven measurable axes, with their attacks mapped to MITRE ATT&CK for ICS tactics. We introduce a checklist for documentation completeness (0–7) and a decision-aid rubric (0–15) covering realism, attack diversity, class imbalance, documentation, and reproducibility. Protocol coverage in datasets is skewed toward Modbus, while many other protocols (such as Profinet and OPC UA) are underrepresented relative to their deployment in industry. Available datasets show structural gaps in capturing multi-stage adversary behavior. In practice, dataset selection should pair a realism-anchored aspect with a high-reproducibility one, and account for protocol diversity and APT representation.

**Keywords:** CPS/IIoT security; ICS security; OT intrusion detection; dataset; MITRE ATT&CK; anomaly detection; machine learning for security

## 1. Introduction

Industrial Control Systems (ICS) are basic elements of critical infrastructure, used in different areas to control operations through automation. They are composed of many different technologies such as SCADA, PLC, and field components. The ICS networks offer the infrastructure that can link these components, thereby allowing real-time monitoring and control of industrial plants. The typical system design of an industrial network, following the Purdue Reference Model (Figure 1), defines five levels starting from the process and ending with the enterprise layer. The Industrial Internet of Things (IIoT) extends ICS connectivity with additional heterogeneity in connectivity, vendors, protocols, and cloud-based services. The convergence of IT and OT environments and the adoption of IIoT have widened the cyber-attack surface of industrial environments [1].

The Stuxnet attack [2] was one of the initial alerts of the potential damage cyber threats can inflict on industrial systems, illustrating that their consequences can extend to cause real-world physical damage. Successful cyber-attacks on ICS can result in physical damage and disruption of industrial plants. It can even lead to loss of human lives, which is a serious issue for industries and governments [3]. The frequency and complexity of attacks on critical infrastructure require security strategies that combine technical defenses with organizational practices and operator training. IDS is one such solution that detects potential threats to the network by inspecting and analyzing network traffic. It can be deployed as a software solution or as a physical network device on a firewall appliance.

 doi:10.20944/preprints202606.0145.v1



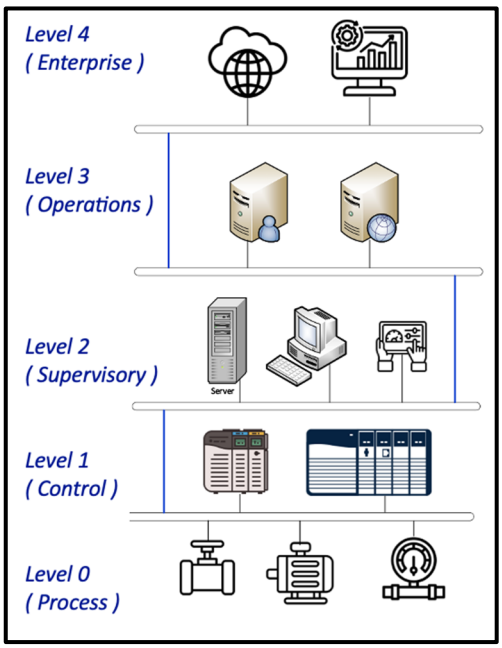


**Figure 1.** Standard ICS Interfacing Diagram.

Figure 2 shows two different IDS deployments, which can be installed in line with traffic and provide a passing control capability for the traffic called IPS. There are two main IDS mechanisms. One is signature-based, which compares network packets against a known database of malicious traffic, while the other is anomaly-based, which builds a baseline of normal network traffic and inspects for any deviations. Different ML approaches are gaining attention in IDS technologies and are becoming increasingly important in industrial and academic domains.

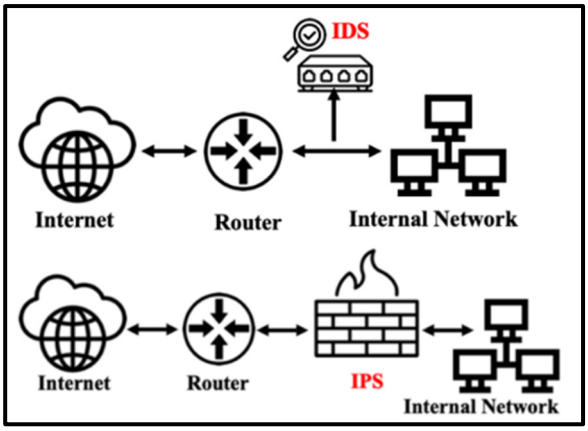


**Figure 2.** Intrusion Detection/Prevention Deployments.

Two factors explain the recent uptake of ML-based IDS for ICS: an increasing threat volume against ICS infrastructure, and the lowered engineering barrier provided by open-source ML frameworks (PyTorch, TensorFlow, scikit-learn) and pre-trained models.

Three ML learning families are used in ICS-IDS research, each with different dataset requirements. Supervised learning involves training a classifier using labeled data of normal and attack instances. It has good detection accuracy on the dataset, but poor generalization performance to new attacks or datasets. This type is limited by the quality of the labels. Most of the ICS-IDS research in the literature is based on supervised techniques (Random Forest, SVM, deep neural networks). Unsupervised learning requires unlabeled instances and is promising for ICS applications where there are few labeled instances, but it is sensitive to non-malicious operational changes and is prone to high false-positive rates. Semi-supervised learning uses limited labeled data, while the

 doi:10.20944/preprints202606.0145.v1



majority of data is unlabeled. It combines the advantages of the two former learning types and is of growing interest due to the high labeling cost [4–6].

Datasets shape three stages of the ML pipeline: training, validation, and testing. Training sets are used to fit model parameters by minimizing the loss between predictions and labels. Validation sets are held out during training and used to tune hyperparameters; they expose overfitting before the model meets unseen data. Test sets, used only after training is complete, provide the unbiased performance estimate that will generalize to deployment. Each stage depends on the next: poor partitioning of any one of the three undermines the others. Dataset quality, therefore, propagates into every downstream IDS performance metric [7].

This study offers a multi-axis taxonomic analysis of cybersecurity datasets for ICS network intrusion detection. A seven-dimensional taxonomy is applied to 23 ICS-specific datasets, and a quantitative documentation-completeness score is used in place of the qualitative framing adopted by prior surveys. A five-criterion decision-aid scoring rubric is proposed and yields per-research-direction dataset recommendations. The rest of the paper is organized as follows: Section 2 reviews related work in the ICS IDS literature; Section 3 describes the methodology used to study and analyze the datasets; Section 4 surveys the selected datasets; Section 5 analyzes the datasets according to the seven taxonomic dimensions; Section 6 presents the results and findings, including the decision-aid scoring rubric; Section 7 discusses future research directions; and Section 8 concludes the paper.

## 2. Related Work

Conti et al. [1] examined a large number of testbeds/datasets and verified the performance of the algorithms on them to provide a baseline. The period covered by our paper is extended to Q1 2026 (including the ICS-NAD, ICS-ADD, HiTar, and EDS datasets), whereas Conti et al. cover up to 2020. We introduce an explicit per-dataset documentation-completeness scoring, and add a use-case ranking rubric.

The protocol and feature taxonomy presented here is complementary to the taxonomy of attack paths used by Choi et al. [8] for comparing ICS datasets. Mubarak et al. [9] presented an ICS testbed and ML attack-detection results but did not provide a taxonomy analysis. Martins et al. [10] compared ICS datasets, specifically regarding composition and attack-scenario coverage, and found that attack range and class imbalance were limitations. Mitseva et al. [11] investigated representativeness properties and dataset constraints in a limited way. Both works compare datasets as objects, but do not provide quantitative documentation, scoring, or per-use-case rankings. Pinto et al. [6] conducted a survey of ML-based IDS for critical infrastructure, which also included the datasets used to train the ML-based IDS, while our paper focuses on ICS-related datasets. Koay et al. [5] conducted a survey of ML in ICS security and found that dataset diversity is an open challenge. Dehlaghi-Ghadim [12] points out that the common datasets do not contain realistic ICS network data, which results in shortcomings in the evaluation of the effectiveness of different ML algorithms for ICS-IDS. Hu et al. [13] also pointed out that another challenge to IDS methodologies is the lack of comprehensive and realistic datasets that accurately reflect real ICS network traffic.

The quality of the data has a significant impact on the feature engineering performed in the pre-processing phase of model training. Anwar et al. [14] discuss attribute extension to improve anomaly detection in SCADA networks. They conclude that the ability of ML algorithms to detect attacks can be greatly improved by modifying and extending the dataset attributes. Sun et al. [15] propose a hybrid ML model to overcome the high dimensionality and class imbalance issues frequently encountered in ICS datasets by using feature selection. Their approach shows that the use of different ML techniques in conjunction with good feature selection can result in more powerful and reliable IDS in ICS environments.

Yang et al. [16] systematically reviewed anomaly-based network intrusion detection methods and datasets, without the ICS-process focus adopted here; Kheddar et al. [17] surveyed deep transfer learning for ICS-IDS, with dataset coverage as a secondary axis. Mubarak et al. [9] provided a focused analysis of public ICS datasets using EDA and a Random Forest baseline. In addition to the

 doi:10.20944/preprints202606.0145.v1



taxonomies listed above, this work differs from previous surveys in three specific ways: (i) coverage of releases from 2024 to Q1 2026; (ii) the introduction of a quantitative documentation-completeness score in Section 6; and (iii) the production of per-use-case dataset rankings.

## 3. Study Methodology

### *3.1. Search Strategy and Selection Procedure*

We searched for datasets using structured queries in international publishers (Springer, IEEE Xplore, ScienceDirect, MDPI, and arXiv). Search terms were combined using the OR operator for "intrusion detection", "anomaly detection", and "IDS"; the AND operator for "industrial control system", "ICS", "SCADA", "cyber-physical system", and "IIoT"; and the AND operator for "dataset", "testbed", and "data collection". Only datasets which satisfy the following criteria were retained: (a) the dataset represents ICS/OT/SCADA/IIoT network traffic or process data that is captured or simulated; (b) the dataset is documented in a peer-reviewed publication or readily accessible technical report; and (c) the dataset is publicly available or available on request, or a labeled subset is referenced in subsequent IDS research. Records were excluded for: IT-only traffic without an OT process or industrial-protocol data (the most common reason). Finally, 23 datasets are screened to represent the surviving set after these filters. Datasets containing only IT without OT traffic are mentioned for reference but excluded from the study scope. This is a narrative taxonomic review, not a systematic review; we did not run a PRISMA-style screening pipeline. The 23 surviving datasets were selected by the criteria above, applied independently by the first author, and reviewed by the co-authors.

### *3.2. Analytical Axes*

Seven measurable dimensions are considered to analyze each dataset. First, we check the environment where the dataset is captured: whether from a physical testbed, by simulation of a process model, or from a real operational network. Then we identify the data type and distinguish process data from networking traffic. Third, we highlight how the dataset is delivered (PCAP, CSV, or other file formats). Next, the OT protocols are checked (Modbus TCP/RTU, S7Comm, and others). The attack representation is studied and analyzed for the diversity of attacks. It is also highly important to identify the dataset features/classes to check whether any imbalance ratio exists between the majority class (normal traffic) and the anomaly minority class. Finally, the documentation is assessed using a novel approach, and we assign a quantitative score to each dataset using a 0–7 checklist, awarding one point when each of the following elements is documented: system architecture, protocol, attack or anomaly methodology, record or class counts, feature description, labeling or ground-truth method, and availability conditions. We then convert the documentation score into three classes, from Mild to Comprehensive. The aim is to provide a comparison of documentation quality across datasets to support a scoring approach and a reproducibility assessment in the decision-aid rubric.

### *3.3. Dataset Characteristics*

This study evaluates one quality characteristic based on the originating publication: the balance between normal and attack traffic (reported as the imbalance ratio in Table 5). Examining the features in these datasets is also important, as they characterize the data and determine what patterns ML models can learn to identify. The protocols represented in the datasets are additionally investigated, since different industrial communication stacks expose distinct vulnerabilities and attack vectors. The various attack types are categorized to assess the comprehensiveness of each dataset's coverage of realistic adversary behavior. We map the identified attacks to the MITRE ATT&CK for ICS tactics [18].

 doi:10.20944/preprints202606.0145.v1



## 4. Datasets Survey

### *4.1. Datasets Used in ICS-IDS*

Researchers in ICS-IDS initially had to rely on IT datasets for model training. Then ICS/OT-specific datasets began to emerge in the mid-2010s. We list the datasets widely used in the literature in Table 1, along with their types and publication years.

**Table 1.** Survey Used Datasets in ICS IDS.

| Sr | Name | OT | Publisher | Year | Ref. |
|---|---|---|---|---|---|
| 01 | SWaT | Yes | iTrust Centre at Singapore University of Technology and Design | 2015 | [19–21] |
| 02 | WADI | Yes | iTrust Centre at Singapore University of Technology and Design | 2017 | [22,23] |
| 03 | EPIC | Yes | iTrust Centre at Singapore University of Technology and Design | 2018 | [24] |
| 04 | BATADAL | Yes | iTrust Centre, Technion, KIOS | 2018 | [25] |
| 05 | S317 | Yes | iTrust Centre at Singapore University of Technology and Design | 2017 | [26] |
| 06 | MSU-GP | Yes | Power/Energy Lab, Mississippi State University | 2014 | [27,28] |
| 07 | MSU-PWR | Yes | Power/Energy Lab, Mississippi State University | 2014 | [29–31] |
| 08 | ICS-Flow | Yes | RISE Institute, Mälardalen University, Sweden | 2023 | [12] |
| 09 | Lemay | Yes | École Polytechnique de Montréal | 2016 | [32] |
| 10 | Electra | Yes | Department of Computer Engineering, University of Murcia, Spain | 2019 | [33] |
| 11 | Rodofile | Yes | Queensland University of Technology, Australia | 2017 | [34,35] |
| 12 | WUSTL-IIoT | Yes | CSE, McKelvey Engineering School, Washington University | 2021 | [36–38] |
| 13 | Edge-IIoTset | Yes | Department of CS, University of Guelma, Algeria | 2022 | [39] |
| 14 | HIL-WDT | Yes | University Campus Bio-Medico of Rome, Italy | 2021 | [40] |
| 15 | X-IIoTID | Yes | University of New South Wales, Australia | 2022 | [41] |
| 16 | TEP | Yes | ONR, Pacific Science & Engineering Group | 2017 | [42] |
| 17 | TLIGHT | Yes | Department of Computer Science, University of Hong Kong | 2017 | [43] |
| 18 | IUNO | Yes | German Research Center for AI (DFKI), Kaiserslautern | 2019 | [44] |
| 19 | ICS-NAD | Yes | Zhejiang University | 2026 | [45] |
| 20 | ICS-ADD | Yes | University of Genoa, Project RAISE | 2024 | [46] |
| 21 | HiTar | Yes | Tunisia Polytechnic, University of Carthage | 2025 | [47] |
| 22 | EDS | Yes | Zhejiang University | 2024 | [48] |
| 23 | HAI | Yes | Affiliated Institute of ETRI, South Korea | 2020 | [49,50] |
| 24 | KDD-CUP99 | No | University of California, Information and CS Department | 1999 | [51] |
| 25 | NSL-KDD | No | University of New Brunswick, Canada | 2009 | [52] |
| 26 | UNSW-NB15 | No | University of New South Wales, Australia | 2015 | [53] |
| 27 | CICIDS | No | University of New Brunswick, Canada | 2017 | [54] |
| 28 | ISCX | No | University of New Brunswick, Canada | 2012 | [55] |
| 29 | BoT-IoT | No | Australian Centre for Cyber Security | 2021 | [56] |

The temporal distribution of OT/ICS vs IT-only dataset releases is illustrated in Figure 3.

 doi:10.20944/preprints202606.0145.v1



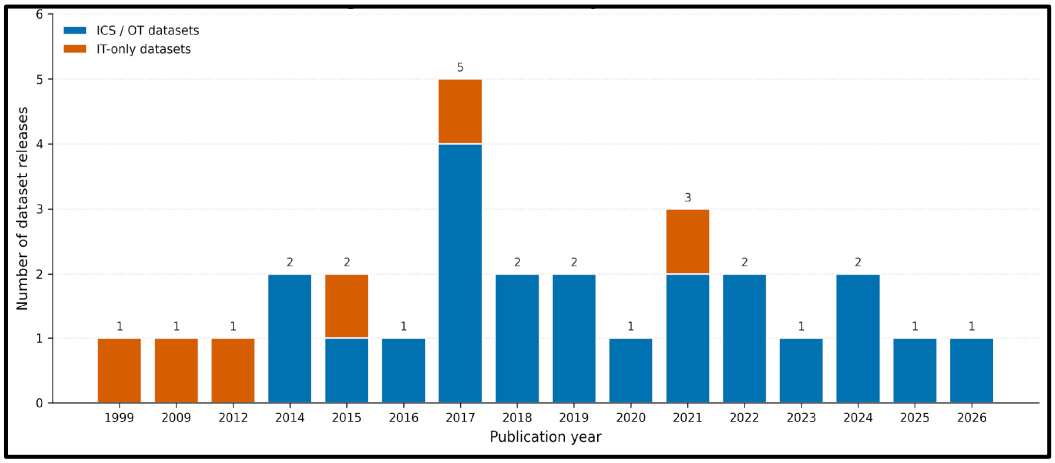


**Figure 3.** Dataset Releases by Year (1999–2026), grouped by OT/ICS vs IT-only.

*4.2. Selected Datasets for This Study*

Our study focuses only on datasets related to ICS network traffic. In Table 2, the selected datasets are presented with general information to provide researchers with insights into them. The industrial domain spans water, gas, electricity, and other sectors. The environment clarifies how the data is captured or generated. The citation statistics indicate academic impact; we relied on Google Scholar as of April 2026. It is worth understanding the availability of these resources, and we classify them into Public (directly downloadable), Request (accessible upon request), and Restricted (not found or unavailable).

**Table 2.** Selected datasets for comparative study.

| Sr | Name | Domain | Environment | Citations | Availability |
|---|---|---|---|---|---|
| 01 | SWaT [19–21] | Water Treatment | Physical Testbed | 1296 | Request |
| 02 | WADI [22,23] | Water Distribution | Physical Testbed | 755 | Request |
| 03 | EPIC [24] | Electrical Power Grid | Physical Testbed | 116 | Request |
| 04 | BATADAL [25] | Water Distribution | Simulation | 325 | Public |
| 05 | S317 [26] | Water Treatment | Physical Testbed | 55 | Request |
| 06 | MSU-GP [27,28] | Gas Pipeline | Physical Testbed | 362 | Public |
| 07 | MSU-PWR [29–31] | Electrical Power Grid | Physical Testbed | 327 | Public |
| 08 | ICS-Flow [12] | Manufacturing | Simulation | 103 | Public |
| 09 | Lemay [32] | Electrical Power | Simulation | 169 | Public |
| 10 | Electra [33] | Railway | Real Application | 155 | Restricted (1) |
| 11 | Rodofile [34,35] | Mining Refinery | Physical Testbed | 40 | Public |
| 12 | WUSTL-IIoT [36–38] | Water Storage / IIoT | Physical Testbed | 80 | Public |
| 13 | Edge-IIoTset [39] | IoT / IIoT | Physical Testbed | 1223 | Public |
| 14 | HIL-WDT [40] | Water Distribution | Physical Testbed (4) | 111 | Public |
| 15 | X-IIoTID [41] | IIoT | Physical Testbed | 331 | Public |
| 16 | TEP [42] | Chemical Process | Simulation | 166 | Public |
| 17 | TLIGHT [43] | Traffic Light Control | Simulation | 22 | Restricted (2) |
| 18 | IUNO [44] | Water Storage | Physical Testbed | 23 | Restricted (3) |
| 19 | ICS-NAD [45] | Power Generation & STP | Real Application | 1 | Public |
| 20 | ICS-ADD [46] | Water Treatment | Simulation | 51 | Public |
| 21 | HiTar [47] | Manufacturing / IIoT | Simulation | 4 | Public |
| 22 | EDS [48] | Ethanol Distillation | Physical Testbed | 38 | Public |
| 23 | HAI [49,50] | Power Generation | Physical Testbed + HIL | 320 | Public |

 doi:10.20944/preprints202606.0145.v1



*(1) The Electra dataset (Perales Gómez et al., 2019) was released, but no current public repository is found as of 2026. (2) Yau & Chow (2015) published a system specification only for TLIGHT, and no downloadable dataset is found. (3) The IUNO dataset (Duque Antón et al., 2019) was released as part of the BMBF IUNO project, and no public repository is found as of 2026. (4) HIL-WDT can be considered a hybrid of a physical testbed plus a simulated MiniCPS.*

According to the chart in Figure 4, the majority of the OT datasets, 14 out of 23 (60.9%), were created with physical testbeds. This means that much of the comparative study is conducted with datasets produced in a controlled laboratory setting with real industrial parts; the realism of these data relative to synthetic data is higher. The second group of datasets is simulation-based, comprising 7 datasets (30.4%), demonstrating that simulation is used to safely model attacks and complex industrial processes. This shows that there is a lack of datasets collected from operational industrial infrastructures due to privacy, safety, and security constraints, except for Electra and ICS-NAD (2 datasets, 8.7%), which are the only ones to have traffic from a real application.

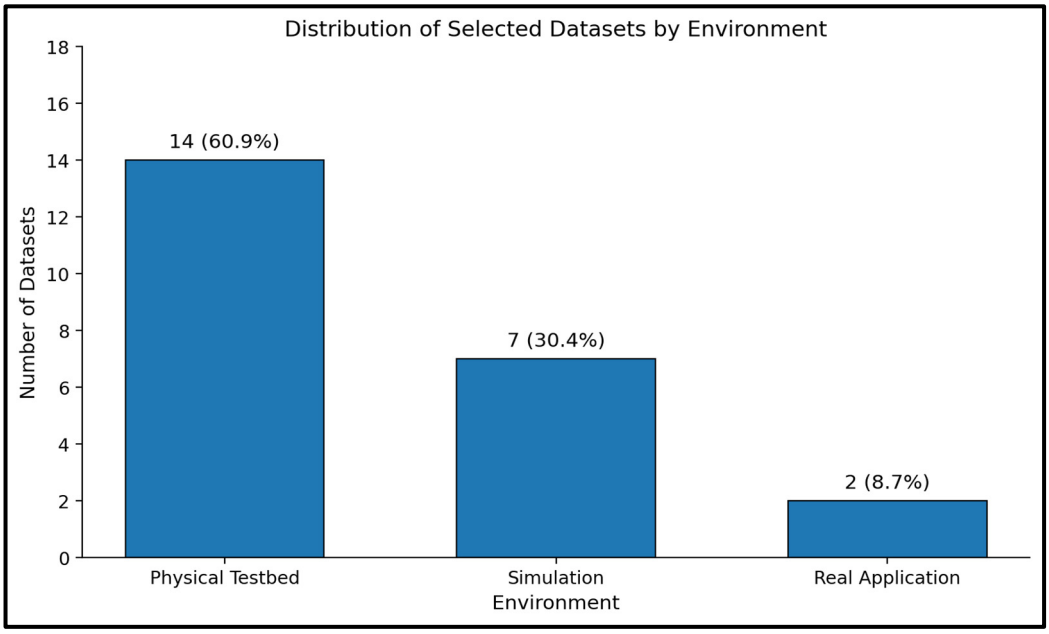


**Figure 4.** Environment distribution of studied datasets.

### *4.3. Dataset Generation Across the Selected Datasets*

#### 4.3.1. SWaT Dataset

The iTrust Center for Research in Cyber Security at Singapore University of Technology and Design (SUTD) has developed datasets that contain both normal and malicious samples [57]. The most widely used datasets are SWaT and WADI. The SWaT testbed is a scaled-down version of a modern water treatment facility. It consists of multiple components forming a modern six-stage water treatment process. It comprises various electrical, electronic, physical, and mechanical components. This six-stage system is operationally capable of treating and purifying 19 liters of water per minute [19].

The testbed includes sensors and actuators, six Rockwell Automation PLC systems, six Schneider remote I/O units, and HMI, SCADA, and Historian systems. The system architecture and interfaces of the testbed is available at webpage of iTrust Testbeds (https://www.sutd.edu.sg/itrust/itrust-labs/testbeds). The SWaT dataset was collected over continuous operation for eleven days. It recorded 7 days of normal operations and 4 days during which various attacks were executed. The CSV process dataset contains readings from 51 sensor/actuator tags. Network traffic is stored in separate packet-level files [20].

 doi:10.20944/preprints202606.0145.v1



Researchers performed 41 attack scenarios as adversarial operators and manually recorded the time of the attacks, the sensor or actuator targeted, and the occurrences of physical consequences. The network traffic was collected simultaneously on the process network (EtherNet/IP) and on the corporate LAN, resulting in packet-level PCAP files and sensor value CSV logs [21].

#### 4.3.2. WADI Dataset

The Water Distribution testbed is the downstream extension of SWaT and comprises raw-water tanks, elevated reservoir tanks, six consumer-tank assemblies, a return tank, pumps, motorized valves, solenoid valves, pipelines, flow meters, pressure transducers, and water-quality analyzers to measure pH, conductivity, and turbidity [22].

The system architecture and interfaces of the testbed is available at webpage of iTrust Testbeds (https://www.sutd.edu.sg/itrust/itrust-labs/testbeds), and consists of 3 Rockwell PLCs using EtherNet/IP and 2 Schneider RTUs using Modbus TCP, bridged through Moxa Modbus/Eth IP gateways. The first PLC controls the primary reservoir, the second the consumer tank, and the last the return water. The historian stores all process values with a 1-second sampling rate.

The dataset was acquired in normal operation and attack operation in 2017. 15 attacks were applied related to the malicious manipulation of valves, pumps, flow transmitters, level readings, quality readings, chemical dosing, consumer-supply valves, leakage paths, and booster pressure set points. Some attacks are coordinated or stealthy, e.g., draining elevated reservoirs while tricking the sensors to hide the actual impact [23].

#### 4.3.3. EPIC Dataset

As a cyber-physical systems testbed, the Electrical Power and Intelligent Control testbed was developed at SUTD iTrust to facilitate smart-grid security research. It contains four major subsystems starting from generation, along transmission, passing the micro-grid, and finally the smart home. It aims to facilitate experiments on cyber-physical attacks, the cascading effects, and the testing of attack-detection and defense solutions. The testbed components comprise motor-driven generators, photovoltaic panels, a battery system, and load demand. The control system consists of PLCs, IEDs (Intelligent Electronic Devices), switches, SCADA, and historian functions [24].

The system architecture and interfaces of the testbed is available at webpage of iTrust Testbeds (https://www.sutd.edu.sg/itrust/itrust-labs/testbeds), which comprises supervisory control, local control, protection, and network communication. The entire system is monitored by the SCADA workstation and is used for supervisory control, with protection and control functions accomplished by the IEDs. The EPIC paper mentions the use of WAGO PLCs for breaker control and generator sync logic, and Siemens SIPROTEC relays as IEDs. Communication is based on IEC 61850 as the primary standard, which facilitates the transfer of events and system data.

GOOSE and MMS are utilized in the IEC 61850 communication landscape to transfer data between IEDs and the SCADA workstation. GOOSE is used as the IEC 61850 mechanism for fast event/protection-related messaging, while MMS can be considered the IEC 61850 communication service for monitoring and controlling data exchange between supervisory entities and field/control devices [58]. The paper also mentions that communication between SCADA and PLCs can be wired or wireless, and that in some PLC configurations Modbus TCP/IP may be used in addition to IEC 61850. EPIC is capable of supporting a variety of attack scenarios applicable to smart-grid security. The author outlines four possible classes: power supply interruption, nuisance tripping, physical damage, and economic-advantage attacks. Two of these were demonstrated: a power supply interruption attack and a physical damage attack. EPIC can be said to capture realistic cyber-physical behavior of a physical smart-grid testbed in normal operation and selected attack scenarios.

 doi:10.20944/preprints202606.0145.v1



#### 4.3.4. BATADAL Dataset

The Battle of the Attack Detection Algorithms dataset is the result of a community-wide competition to compare cyber-physical attack detection algorithms for water distribution networks. The medium-sized, real-world, C-Town water distribution system was controlled by programmable logic controllers (PLCs) and a SCADA system. Participants were given simulated observations of the SCADA and asked to identify whether it was operating normally or being attacked, with performance evaluated using classification accuracy and time to detect the intrusion [25].

The benchmark water network is composed of one reservoir, seven storage tanks, eleven pumps distributed across five pumping stations (S1–S5), and five valves (only V2 is actuated and instrumented in the SCADA dataset). These physical assets are connected to a cyber layer comprising 9 PLCs and a SCADA system. The PLCs monitor and control pumps, tanks, and valves, and the SCADA system coordinates operations and stores PLC readings. Hence BATADAL is regarded as a simulated dataset.

The datasets were created with epanetCPA, a MATLAB toolbox for designing cyber-physical attacks and simulating the hydraulic consequences with EPANET. The hydraulic simulation was run with a higher internal time resolution, and the SCADA observations made available to participants were sampled at regular 1-hour intervals. The set of SCADA observations released includes 43 variables: 7 water-level variables in the tanks, 12 pressure variables at selected locations in the network, and 24 flow-status variables for pumps and the actuated valve.

Three datasets were released. Training Dataset 1 contained 8,761 hourly samples (~365 days) of normal operation without any cyber-attack. Training Dataset 2 contained 4,177 hourly samples (~6 months) and included seven attacks spanning approximately 492 attack-labeled hours; attack labels were provided for training and validation, but not all attacks were fully disclosed during the competition. The Test Dataset contained 2,089 hourly samples (~3 months) with seven additional attacks spanning approximately 407 attack-labeled hours, where attack information was withheld from participants during evaluation. In total, the corpus comprises 14 attack scenarios across the three datasets.

The attack scenarios include actuator activation and changes to actuator settings, and deception attacks affecting sensor, PLC, and SCADA communications. Examples include changing pump-control thresholds, reducing pump speed, maliciously activating pumps, manipulating tank-level signals, and concealing the effects of attacks. The original BATADAL dataset contains time-series sensor/actuator observations and does not contain network-level packet captures or PCAP files. This sets it apart from the later BATADAL 2.0 work that was created to provide a more detailed representation of industrial communication/network processes.

#### 4.3.5. S317 Dataset

The S317 dataset is from the SWaT Security Showdown (S3), an Industrial Control System security event. It was a gamified ICS education and research event where independent attacker teams from academia and industry attacked SWaT/WADI, while academic teams attempted real-time detection. According to Antonioli et al., a Capture-the-Flag event is an ICS-focused event composed of attacker teams, a live attack-defense phase, and data collection for later analysis [26]. The dataset includes three types of evidence gathered during the event: PCAPs of the network packets captured, Historian data from the process level, and documentation of the attack scenarios conducted by participants. The PCAP files enable researchers to analyze communication-level data such as protocol behavior, packet timing, and network anomalies, while the Historian data enables analysis of physical-process data and actuator states from the SWaT plant. Because of these features, the dataset can be used to assess intrusion detection techniques that combine cyber-network indicators with process-state changes [26].

 doi:10.20944/preprints202606.0145.v1



#### 4.3.6. MSU-GP Dataset

The MSU Gas Pipeline dataset was released by Morris and Gao (2014) at Mississippi State University as one of two companion datasets derived from laboratory-scale physical ICS testbeds that use the Modbus application-layer protocol [27]. The gas pipeline testbed implements a supervisory control system for a small-scale gas-flow process, with a SCADA master station (MTU) communicating over serial Modbus with a remote terminal unit (RTU) that monitors and actuates process variables such as pressure, flow, and pump/solenoid states. While the testbed is not an operational natural-gas pipeline, it physically reproduces the cyber-physical interactions characteristic of gas pipeline SCADA communication and control [27].

The testbed adopts a typical master–slave architecture: the master station (MTU) periodically polls a remote terminal unit (RTU) controlling field devices, communicating over Modbus RTU (serial Modbus). The Modbus protocol provides no inherent authentication, integrity, or availability protections, leaving it vulnerable to attacks that alter commands, responses, measurements, or service availability. Morris and Gao's work documents datasets of process variables and network-traffic features captured during the execution of 28 attacks across the gas pipeline and water storage tank testbeds [27].

The dataset is distributed in ARFF format (a CSV-compatible format), with both a preprocessed feature dataset and a raw Modbus-frame dataset that enables independent feature engineering. Each record combines network-communication attributes, process-related measurements, and a ground-truth attack label, making the corpus suitable for supervised intrusion detection, both binary (attack vs. normal) and multiclass (attack-category) classification. Attack scenarios cover four high-level categories: reconnaissance, command injection, response or measurement injection, and denial-of-service. These threats either target the cyber layer by manipulating Modbus communication or affect the cyber-physical layer by altering the apparent or actual state of the gas pipeline process [27]. The widely used gas-pipeline subset evaluated in the IDS literature is the Turnipseed (2015) variant built on the same MSU testbed, which contains 274,627 labeled instances with refined randomness compared to the original release [28].

#### 4.3.7. MSU-PWR Dataset

This dataset was generated at the MSU Power and Energy Research Lab electric-power cyber-physical testbed by Adhikari, Pan, and Morris, with the raw logs subsequently formatted into ARFF/CSV releases through a collaboration with Borges and Beaver at Oak Ridge National Laboratory (ORNL) [29,31].

It simulates a simple transmission network comprising generators, transmission lines, switching/protection devices, and load elements, allowing researchers to observe the impact of cyber/operational events on both power-system measurements and operational/security logs [29].

The released dataset has been utilized in supervised and anomaly-based IDS studies because it offers labeled examples of normal, disturbance, and attack behavior in a power-system context. The dataset is composed of Synchrophasor/PMU measurements, relay logs, network-event-monitor logs, control-panel logs, and Snort alert indicators, all of which are time-synchronized but heterogeneous. It consists of 15 sets of power-system scenarios, with 37 event scenarios in each set. The attack scenarios are remote tripping command, data injection, and changes in relay setpoints [29,59]. This structure is important for cyber-physical IDS research as it helps to differentiate between malicious cyberattacks and non-malicious power-system disturbances, which is essential for minimizing false alarms in real power-system monitoring.

The dataset is released in three sub-versions (binary, three-class, and multiclass). They are built from fifteen sets of 37 scenarios with 1% random sampling [29,30], and the figures reported in Table 5 correspond to the binary sub-dataset. This dataset has been used as a benchmark for hybrid IDS [60] and disturbance-versus-attack classification [59].

 doi:10.20944/preprints202606.0145.v1



#### 4.3.8. ICS-Flow Dataset

It was introduced by Dehlaghi-Ghadim et al. (2023) and created in a simulated ICS environment of a manufacturing bottle-filling industrial process. In such applications, the sensors, actuators, valves, tanks, and conveyor are monitored and controlled. This dataset includes process data and networking traffic. The released information includes raw packet captures, network flow data, and logs of process variables, which allow researchers to examine the behavior of both communications and the physical process under normal and abnormal conditions. The authors also developed the open-source tool ICSFlowGenerator, which allows the extraction of ICS-oriented flow features from raw network packets. In total, the final dataset includes over 25 million raw network packets, flow records, and logs of process-state variables [12].

Two categories of attacks are performed: reconnaissance and disruption (IP/port scanning, DoS/DDoS), and Modbus communication manipulation (replay of previously captured valid packets, alteration of Modbus register values) [12]. Because the dataset is released simultaneously at packet, flow, and process-variable granularity, it supports both supervised and unsupervised IDS evaluation across abstraction levels.

#### 4.3.9. Lemay Dataset

The Lemay/Fernandez dataset provides SCADA network traffic and was generated in a SCADA sandbox rather than collected from a real plant or testbed. The Modbus/TCP protocol is used for communication between Master Terminal Units (MTUs) deployed via ScadaBR and controller/RTU emulators deployed via Modbus_tk. To give the process values realism, the authors added an electrical-network simulation: each controller represents a small electrical network including an electrical source, breakers, and voltage measurements. We can consider this a software-emulated SCADA/Modbus dataset [32].

The captured data includes both benign and malicious Modbus packets in full capture, along with CSV label files for supervised machine-learning experiments. The normal traffic is primarily the deterministic polling of typical Modbus SCADA traffic, and some of the captures include simulated manual operations (Modbus write commands). The attack scenarios were executed live against the sandboxed SCADA network, rather than simply being injected into an already finished capture. The authors detail attacks that include targeting another RTU with a Metasploit MS08-netapi exploit, transferring files via a Meterpreter channel, Modbus fingerprinting using multiple read commands, and an unauthorized Modbus WRITE_COIL command sent from an attacked RTU. There is also a separate covert-channel dataset for Modbus in the dataset [32].

The dataset does not fully represent a production SCADA network. The simulated RTUs share near-identical configurations, polling intervals are fixed within each capture, and the manual write operations are scripted rather than driven by a human operator. Only MTU-to-RTU traffic is captured; the link between RTUs and field devices lies outside the dataset.

#### 4.3.10. Electra Dataset

It is a relatively realistic ICS dataset compared to many laboratory-only or purely simulated datasets, because it was generated from actual network traffic of an electric traction substation in the railway sector, as introduced by Perales Gómez et al. (2019). The system is not a typical railway signaling or rolling-stock system, but a railway traction-power system. It comprises five PLCs (one master and four slaves), a SCADA system, a switch, a firewall, and an additional attacker node for deploying a man-in-the-middle threat. The system architecture supports Modbus TCP, S7Comm, and OPC, while the released dataset is predominantly split into two protocol-specific subsets (Electra Modbus and Electra S7Comm) [33].

All normal and malicious network traffic was captured during data collection. Multiple PCAP files (about 20 GB) were generated by capturing mirrored network traffic using Wireshark, with the buffer size set to 200 MB. Once captured, the traffic was split into two sets: Modbus packets (TCP port

 doi:10.20944/preprints202606.0145.v1



502), and S7Comm (TCP port 102). The authors then used Python and Scapy to parse the packet captures and extract packet-level features into CSV files. A variety of different features were selected, including timestamp, src/dest MAC and IP address, request/response indicator, function code, error indicator, memory address, data value, and label. Multi-operation packets were split into one read or write operation per dataset instance [33].

Reconnaissance, false data injection, and replay attacks are represented in Electra. Specifically, the attacks include function-code recognition, read-attack traffic, write-attack traffic, modification of responses and commands, force-error attacks, and replay of previously observed valid packets. These attacks do not stem from truly uncontrolled adversary activity; they are actively executed in the substation environment by an attacker node. The dataset is not real incident traffic, but rather realistic industrial testbed traffic with executed/injected attacks.

A subsequent study by Pallakonda et al. (2025) tested a secure anomaly detection system that integrates machine-learning classifiers with hybrid AES/RSA encryption, protocol hardening, and threat-hunting techniques, employing the Electra Modbus subset. The model achieved good performance for binary and multi-class classification [61].

#### 4.3.11. Rodofile Dataset

The Rodofile dataset was produced at the Queensland University of Technology using a physical SCADA process-control testbed that simulates a simplified mining refinery plant. The plant comprises three processes (conveyor, wash tank, and pipeline reactor). The testbed architecture (Figure 5) includes a Siemens SIMATIC S7-300 Master and three S7-1200 slave PLCs, where each slave PLC is responsible for one process. The HMI is used by the operator to monitor and control the simulated process and store historian logs that record the state of the control process during the experiment [34]. Additionally, there is an attacker workstation, a GPS clock, managed switches, and two extra hubs to capture traffic from the Master PLC and HMI viewpoints, as well as for dataset labeling.

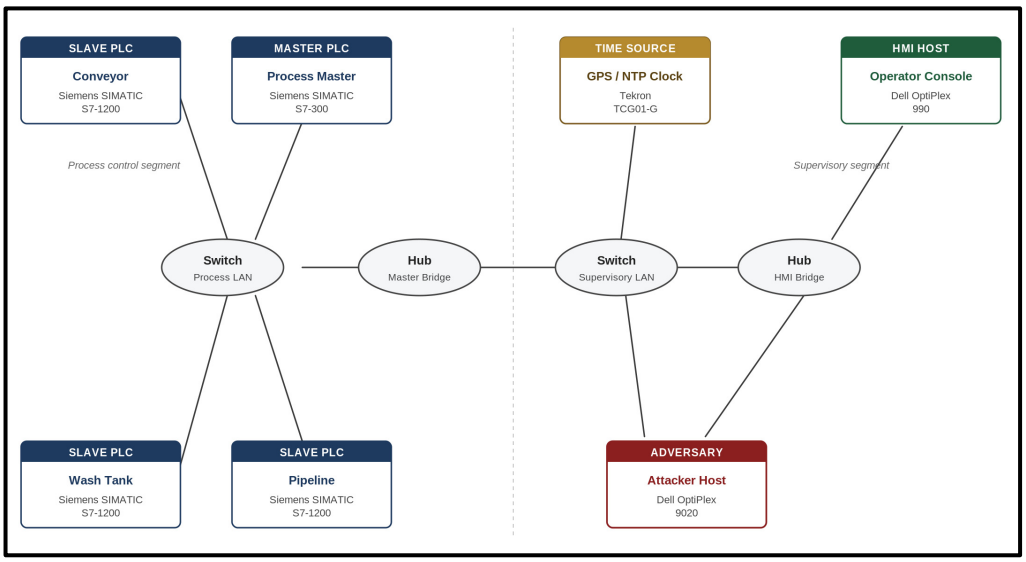


**Figure 5.** Testbed Interfacing Diagram Described by Rodofile et al. [34].

The attacks do not exploit vulnerabilities in the S7Comm parser or its implementation. They leverage the lack of authentication and integrity in the protocol to write arbitrary values to PLC registers. This disrupts the timed sequencing and safe operation of the industrial process.

The released dataset includes network traffic and process logs. The attack dataset was recorded for about 9 hours and consists of 30 process runs, whereas the control dataset consists of about 8.5 hours of data during normal operation with 32 process runs. The attack data consist of 64 attack instances. There are packet-capture files and four process logs included for each dataset (Tank log,

 doi:10.20944/preprints202606.0145.v1



Conveyor log, Reactor log, and Master log). The dataset is thus useful for IDS research, as both S7Comm network traffic and process-state behavior can be analyzed. A major drawback, however, is that the process is small-scale, scripted, and modeled in a laboratory setting rather than a complete, production-scale mining or mineral-processing plant.

#### 4.3.12. WUSTL-IIoT Dataset

This dataset was produced by Zolanvari et al. from Washington University in St. Louis (WUSTL) for cybersecurity research in the Industrial Internet of Things. It was generated from an IIoT testbed designed to simulate an industrial plant and enable real cyberattacks against the system. The physical process consists of a water tank, water-level sensors, a turbidity sensor, pumps, a valve, alarm indicators, an HMI, historian/logging components, and a PLC. The testbed communicates with the control components via Modbus/TCP, and the PLC is based on a Schneider Electric controller [36]. As a consequence, the dataset is referred to as a laboratory IIoT/ICS water-process testbed.

The released dataset consists of network flow records extracted from traffic collected during both normal operation and attack scenarios. The authors report collecting 2.7 GB of data in approximately 53 hours, followed by preprocessing. The released dataset has 1,194,464 instances and 41 selected network-flow features, of which 1,107,448 are normal and 87,016 are attack. The authors recommend removing certain features (such as start/end time or IP identifier/address) prior to modeling, as they can provide attack-specific information and reduce the ability to generalize to unseen data.

The dataset labels support both multi-class and binary intrusion-detection formulations. The original Traffic label identifies the traffic type, while binary classification can be obtained by mapping all attack traffic to one attack class and normal traffic to another. The main attack categories reported are command injection, denial-of-service, reconnaissance, and backdoor traffic. The overall attack proportion is less than 8%, with normal traffic representing about 92.72% of the dataset and total attack traffic about 7.28%. Among the attack samples, DoS dominates the distribution, while command injection and backdoor samples are comparatively rare. This deliberate imbalance reflects the authors' motivation to approximate a more realistic IIoT security setting, where attack events are much less frequent than normal traffic [36].

WUSTL-IIoT-2021 is therefore useful for evaluating IDS models in an IIoT/ICS context, especially under class-imbalance conditions. Its limitations are that it is still a controlled laboratory dataset rather than production traffic. The released data are primarily preprocessed flow features rather than full raw packet/process datasets, and some features must be carefully removed to avoid leakage during model training.

#### 4.3.13. Edge-IIoTset Dataset

Ferrag et al. (2022) presented this dataset for IoT/IIoT intrusion detection research that can be evaluated using either centralized or federated machine learning. The dataset is captured through a specially designed IoT/IIoT testbed. It is structured into seven layers covering cloud computing, network function virtualization, blockchain, fog computing, software-defined networking, edge computing, and IoT/IIoT perception. This layered architecture was intended to depict a modern edge-cloud IoT/IIoT deployment and not a traditional SCADA-only architecture [39].

It has 2,219,201 labeled records: 1,246,500 normal and 972,701 attack. The authors first identified 1,176 raw features from various sources, including network traffic, logs, system resources, and alerts. They then selected 61 high-correlation features for machine learning. The generated files consist of normal and attack traffic in PCAP and CSV formats, along with selected CSV files ready for traditional ML and deep-learning experiments [39]. It may not be a classical OT/SCADA dataset, since it primarily focuses on heterogeneous IoT/IIoT edge and cloud traffic.

 doi:10.20944/preprints202606.0145.v1



#### 4.3.14. HIL-WDT Dataset

The Hardware-in-the-Loop Water Distribution Testbed dataset was introduced by Faramondi et al. (2021) at the University Campus Bio-Medico of Rome, Italy. The testbed is a water distribution process within a hardware-in-loop architecture, in which a real physical water distribution process is virtually interconnected with a simulated unit. This design allows the dataset to capture both the physical consequences of attacks and their corresponding network-level behavior.

The real subsystem comprises five physical water tanks, 20 solenoid valves, four pumps, and five pressure sensors. The simulated subsystem, implemented using the MiniCPS tool, extends the system with three simulated tanks, two simulated pumps, four flow sensors, two solenoid valves, and three pressure sensors. Thus, the entire HIL-WDT process consists of eight tanks, five of which are actually implemented. The process is divided into four stages; the first stage is controlled by a real PLC, and the other stages are controlled by simulated PLCs. The PLC used in the testbed is a Modicon M340, and the communication protocol is Modbus TCP/IP.

Physical process data and network traffic data are included in the dataset. The historian records the physical measurements every second and saves them in CSV files. While the network traffic is captured in PCAP files using Wireshark, the features were extracted and saved in CSV files. There are 41 features in the physical data, including tank levels, pump states, flow-sensor readings, and valve states. The network dataset includes 14 features such as src/dest IP and MAC addresses, ports, protocol, TCP flags, payload size, Modbus function code/value, and short-window packet counts.

The data are broken into four acquisitions of approximately 2 hours total. One acquisition is normal operation, and the others are attack scenarios. The authors specify 28 attack scenarios, including both cyber and physical attacks. Cyberattacks include man-in-the-middle attacks based on ARP poisoning, denial-of-service attacks (TCP flood, ICMP flood, LAND), and scanning attacks (SYN, FIN, NULL, XMAS). Physical attacks involve water leaks from manual valves, sensor and pump failures. Importantly, some attacks impact only network traffic, some only the physical process, and some both. This renders HIL-WDT suitable for analyzing the correlation between cyber events and physical process anomalies.

#### 4.3.15. X-IIoTID Dataset

Al-Hawawreh et al. (2022) introduced the X-IIoTID dataset for IIoT. It is primarily designed to overcome the heterogeneity of IIoT systems, where different types of devices, communication patterns, and protocol families are present. The dataset was created using the Brown-IIoTbed testbed at UNSW Canberra, which combines legacy industrial devices and newer IoT, edge, cloud, and enterprise devices. The testbed consists of three tiers: edge, platform, and enterprise. The edge tier contains field devices, PLC-related components, sensors, actuators, an edge gateway, and local clients. The platform tier comprises a cloud application, an MQTT broker, and cloud storage. Finally, the enterprise tier encompasses the Web-SCADA/API, remote maintenance, and attacker machines [41].

The dataset is a collection of IIoT communication and system behavior from various sources. The authors gathered end-to-end network traffic, host-device logs, host-resource measurements, device-operational logs, and alert logs from OSSEC and Zeek. Zeek was used to generate connection logs for network feature extraction, and a combination of Python and batch scripts was used to parse, correlate, enrich, and label the data collected. The dataset comprises traffic and attack behavior with legacy and recent IIoT protocols, including Modbus, MQTT, CoAP, WebSocket, HTTP, SSH, DNS, ICMP, SMTP, TCP, UDP, and ARP. Thus, X-IIoTID can be considered a heterogeneous IIoT intrusion dataset rather than a traditional single-protocol ICS dataset.

X-IIoTID offers a structured taxonomy of attacks comprising nine high-level attack classes: reconnaissance, weaponization, exploitation, lateral movement, command and control, exfiltration, tampering, crypto-ransomware, and ransom denial-of-service. These are further broken down into 18 attack sub-types, including generic scanning, vulnerability scanning, CoAP resource discovery, fuzzing, brute force, dictionary attack, malicious insider activity, reverse shell, man-in-the-middle,

 doi:10.20944/preprints202606.0145.v1



Modbus register reading, MQTT cloud-broker subscription, TCP relay, false-data injection, fake notification, crypto-ransomware, and RDoS. The final dataset totals 820,834 records, of which 421,417 are benign and 399,417 are malicious, and supports labeling at three levels of granularity: binary, nine-category, and eighteen-subtype classification.

One of the main advantages of X-IIoTID is that it facilitates more generalizable IIoT intrusion detection models. Its features are described as connectivity- and device-agnostic, meaning they are not restricted to a single vendor, device family, or application protocol. This makes the dataset more representative of the evaluation of IDS methods in heterogeneous IIoT environments than datasets restricted to a particular protocol, such as Modbus, S7Comm, or DNP3.

The current version, however, does not address direct attacks such as false commands or false process data at the field-device level. It has very imbalanced minority attack classes, such as fake notifications and MitM. Therefore, X-IIoTID is strong for multi-protocol IIoT network/host intrusion detection, but it should not be presented as a complete replacement for physical-process ICS datasets.

#### 4.3.16. TEP Dataset

The Tennessee Eastman Process (TEP) dataset, released by Rieth et al. (2017), is a simulation-based industrial process dataset designed for anomaly detection evaluation. It is based on the Tennessee Eastman Process benchmark, a chemical process simulation widely used in fault detection and process monitoring research. The dataset is provided through Harvard Dataverse as "Additional Tennessee Eastman Process Simulation Data for Anomaly Detection Evaluation" [42] and is intended to support more reliable evaluation of anomaly-detection methods by providing multiple simulation replications with non-overlapping random-number-generator seeds, rather than the single training/testing run available in earlier TEP releases.

The dataset is organized into four R data files (fault_free_training, fault_free_testing, faulty_training, faulty_testing). Each file is loaded as an R dataframe with 55 columns, where 3 identifier columns (faultNumber, simulationRun, sample) and 52 TEP process variables constitute the features reported in Table 5. The faultNumber column identifies the process condition: fault-free data carry fault 0, and faulty data carry fault types numbered 1–20, for a total of 21 process conditions. The simulationRun column ranges from 1 to 500, identifying the replicate from which the row was generated. The sample column is the time index. Sampling follows the standard TEP rate of one observation every 3 minutes. Each training simulation has 500 samples over 25 hours of operation; each testing simulation has 960 samples over 48 hours. In the faulty datasets, faults are introduced at one hour into the training simulations and at eight hours into the testing simulations, so the datasets support evaluation of both detection accuracy and detection delay after fault introduction [42].

This dataset is best characterized as a simulated chemical-process time series that includes process measurements under both normal and faulty operating conditions. It is highly suitable for process-level anomaly and fault-detection studies, but less suitable for evaluating network-based IDS models that require packet, flow, or protocol features.

#### 4.3.17. TLIGHT Dataset

Yau and Chow (2015) proposed the TLIGHT dataset to study the detection of anomalous events in PLCs using machine-learning techniques. Rather than a conventional network-traffic IDS dataset, TLIGHT is better understood as a PLC memory-state/event-log dataset generated from a simulated traffic-light control system implemented on a Siemens S7 PLC. The PLC I/O configuration maps pedestrian-request switches to PLC inputs and pedestrian/vehicle traffic lights to PLC outputs, with a memory bit and several on-delay timers in the ladder logic governing the traffic-light sequence [43].

Data was collected by polling the relevant PLC memory addresses through a custom program built on libnodave (an open-source communications library for Siemens S7 PLCs). The values of the monitored inputs, outputs, timers, and memory bits were logged with timestamps, and non-binary values such as timer counts were converted into binary features for machine-learning processing. The

 doi:10.20944/preprints202606.0145.v1



paper identifies seven normal traffic-light operating states, each defined by a specific combination of timer and output values, which serve as the ground truth for the normal class.

Two experimental datasets were constructed using 10 features per record. Dataset 1 contains 2,800 records (560 training, 2,240 testing); Dataset 2 contains 8,000 records (1,600 training, 6,400 testing). Anomalous observations were generated by manipulating selected PLC address values with Snap7, an open-source suite of Ethernet communication tools for Siemens S7 PLCs. Yau and Chow report high classification accuracy with Decision Tree and Support Vector Machine classifiers from scikit-learn. The dataset has several documented limitations: the scenario is a small simulated traffic-light controller; the anomalies are scripted PLC-memory writes rather than realistic adversary behavior; the learning approach treats each observation independently and does not model temporal dependencies between states; and no publicly downloadable dataset repository is referenced, so subsequent works regenerate the data from the published system specification.

#### 4.3.18. IUNO Dataset

Duque Antón et al. (2019) proposed the IUNO/DFKI dataset to provide labeled training data for industrial intrusion detection. It was designed and built at the German Research Center for Artificial Intelligence (DFKI) in Kaiserslautern as part of the IUNO Insec project, funded by the German Federal Ministry of Education and Research (BMBF). The dataset is derived from a SCADA scenario with real industrial hardware in which Siemens S7 PLCs control a Festo Didactic MPS PA Compact Workstation (a laboratory training rig that emulates a water-tank filling and emptying process). The physical setup consists of two water containers, a centrifugal pump, a solenoid-controlled ball valve, and a set of process sensors covering level, flow, pressure, and temperature. Although the controllers are Siemens S7 PLCs, the captured communication is OPC UA rather than S7Comm or Profibus, so the dataset is properly characterized as an OPC UA-based laboratory industrial-process dataset [44].

Three datasets were created, each consisting of normal process operation with introduced attacks. The first is the captured baseline traffic with synthetic malicious OPC UA messages added offline, and sensor/actuator values are manipulated to simulate false-data injection. The second introduces malicious behavior at the PLC/application layer: a malicious OPC UA client performs reconnaissance, and the PLC returns falsified process values in response. The third captures apparently normal PLC-reported data while the physical process is in fact abnormal, and uses side-channel measurements (sound, flow, and temperature) to expose the deviation. Every attack is labeled across the three datasets, and the ground truth is known, so the corpus can be used to train and test anomaly-based IDS techniques.

The dataset has notable limitations: it is small, restricted to a single laboratory training rig and a single industrial protocol (OPC UA), and the released artifacts do not include a documented train/test split or per-class instance counts. Its scenarios are valuable because they target OPC UA and combine network/process behavior with covert physical deviations, but the experimental scope is limited to a laboratory-scale water process with a small attack inventory.

#### 4.3.19. ICS-NAD Dataset

In 2026, Zhou et al. released ICS-NAD, a large-scale ICS networking dataset captured from real production environments at Zhejiang University. Three PLCs are deployed (ABB, Siemens, and Schneider) covering two industrial applications. The thermal-power process is controlled by ABB equipment communicating over a private TCP-based protocol, while the sewage-treatment process is implemented twice in parallel, with Siemens controllers (using S7Comm) and Schneider controllers (using Modbus) operating as alternative configurations for the same plant [45]. The three setups are wired independently, which allows the release to provide separate collections without cross-contamination of traffic. ICS-NAD therefore qualifies as a real-world ICS scenario dataset spanning multiple vendors and protocols, distinct from the laboratory-testbed releases that dominate the rest of this corpus.

 doi:10.20944/preprints202606.0145.v1



The dataset includes 20 common ICS attack types, falling into four categories: reconnaissance, DoS/DDoS, false data injection (FDI), and man-in-the-middle (MitM). Two attack traffic patterns are defined. The first pattern alternates one minute of normal traffic with four minutes of attack traffic carrying a specific ICS attack. The second pattern is designed to simulate denial-of-service behavior and consists of repeated three-minute cycles with random attack bursts in the second minute, capturing both attack onset and recovery dynamics.

ICS-NAD provides raw packet captures alongside extracted feature sets. Raw traffic is collected via switch-port mirroring and saved in PCAP format; 60 packet- and flow-based features are then extracted using a modified version of the ICSFlowGenerator workflow originally developed for ICS-Flow [12]. The final release totals 245.96 GB across 272 files, comprising raw PCAPs, labeled feature tables, and timestamp logs used during the labeling pipeline.

Overall, ICS-NAD is a strong dataset covering real-world deployment environments, multiple vendor platforms, multiple ICS protocols, broad attack coverage, and labels released alongside both raw and feature-level data.

#### 4.3.20. ICS-ADD Dataset

Gaggero et al. (2024) presented the ICS Anomaly Detection Dataset, collected from a laboratory smart-industry ICS testbed assembled to reproduce the architecture of a small operational SCADA deployment. The release combines raw network traffic with the runtime outputs of open-source security-monitoring tools, enabling analysis of both attack behavior and the responses of operational security tools. The control loop is realized on virtual machines hosted on a rack server: ScadaBR provides the SCADA/HMI, and OpenPLC Runtime provides the PLC. The simulated physical process is a water-treatment-like scenario involving two tanks that are filled and emptied by pumps. Communication between SCADA (Modbus master) and PLC (Modbus slave) is over Modbus/TCP. The network and security plane comprises a pfSense firewall, a managed switch with a SPAN port for traffic mirroring, OSSIM as the SIEM, and Suricata as the NIDS engine running on OSSIM [46].

The dataset captures a single multi-stage attack scenario structured along the Cyber Kill Chain concept and driven by a compromised internal endpoint that serves as the attacker's pivot. The seven attack stages comprise: DNS tunneling for command-and-control communication; port scanning and Modbus scanning for reconnaissance; SCADA password brute-forcing for credential access; ARP spoofing to establish man-in-the-middle positioning; Modbus false-data injection that causes the pump to activate without operator intervention; and denial of service against the SCADA web service. The released dataset has three components: a raw packet-capture file (traffic_capture_span.pcap), ScadaBR event logs (ScadaBr_events.csv), and OSSIM event logs containing both pfSense firewall records and Suricata alerts [46].

ICS-ADD is one of the few datasets that release security-tool outputs alongside raw network traffic, enabling two evaluation modes simultaneously: training ML-based anomaly detection on the PCAP and comparing ML results against the rule-based decisions produced by the operational IDS/SIEM stack on the same captured traffic.

#### 4.3.21. HiTar Dataset

To address the challenge of intrusion detection in smart-manufacturing IIoT environments, Dhaouadi et al. (2025) introduced the HiTar dataset. The dataset is generated using the AREZZO flexible-manufacturing simulator [62], which provides realistic shop-floor scenarios. The simulated environment models a shop-floor automation network with controllers, workstations, RFID reader/writer devices, shuttle stop-and-go elements, and an Ethernet-based Modbus/TCP communication architecture [47].

HiTar is constructed from the LOG files produced by AREZZO and labeled by a script (attack_labelling.sh) released alongside the dataset. The resulting release is a labeled tabular dataset of 15,842 instances with 39 extracted features per record, including timestamp, source and destination

 doi:10.20944/preprints202606.0145.v1



IP, source and destination port, protocol type, TCP-flag indicators, service indicators, and the attack-type label [47].

The dataset contains five labeled classes: Normal, Probing, Remote-to-Local (R2L), User-to-Root (U2R), and Denial of Service (DoS). This taxonomy is analogous to the historical IT-style intrusion-detection datasets such as KDD/NSL-KDD and is not based on OT-specific attack mechanisms such as PLC-logic modification, setpoint alteration, or process-variable falsification. HiTar is therefore suitable for supervised ML-based IDS studies in an IIoT manufacturing scenario, but it is not appropriate for research that requires fine-grained OT-protocol semantics or process-aware attack labels.

#### 4.3.22. EDS Dataset

The EDS dataset was created by Xue et al. (2024) as part of a real-time intrusion-detection study based on decision fusion. To address the well-documented simulation-to-reality gap in ICS cybersecurity datasets, the authors built a full-hardware, high-fidelity Ethanol Distillation System (EDS) testbed. The platform realizes a scaled-down but fully functional ethanol-water distillation process capable of both cold start-up and normal operation. It comprises a PLC-based control system, HMI, supervisory functions, sensors, actuators, and a physical distillation process governed by three control loops: temperature, flow, and liquid level.

Communication is implemented over Ethernet between a Siemens PLC and the supervisory computer using the Siemens S7Comm protocol. Data acquisition is performed by a Python program that uses the snap7 library to periodically poll the PLC, recording both the values of the monitored PLC data points and the associated network communication metadata. The release contains 72,965 data points captured over 10 hours of normal operation; attack-scenario recordings bring the full released corpus to 843,321 instances as reported in Table 5. Each record carries 47 parameters: one timestamp, eight control parameters, 24 digital parameters, and 14 analog parameters [48].

The attack design covers seven ICS-specific threat categories: information leakage, replay attack, command injection, sensor data tampering, control parameter tampering, multi-point attack, and physical attack. The seven attacks are organized along an escalating attacker-capability ladder: from network access alone, to knowledge of the physical process and control logic, to direct access to field devices. This makes EDS richer than purely network-based IDS releases, since it captures cyber and physical-process characteristics in the same experimental environment. The released dataset does not provide complete per-class label distributions for each attack type.

#### 4.3.23. HAI Dataset

Shin et al. (2020, 2021) introduced the HIL-based Augmented ICS (HAI) Security Dataset to address the scarcity of cyber-physical-system datasets with strong cross-process coupling and repeatable attack execution. HAI was produced at the Affiliated Institute of ETRI (South Korea) by combining three independently developed laboratory-scale testbeds (a GE turbine testbed, an Emerson boiler testbed, and a FESTO modular production system water-treatment testbed) via a dSPACE hardware-in-the-loop (HIL) simulator that emulates steam-turbine power generation and pumped-storage hydropower generation. Data are collected through an OPC UA gateway that interfaces with Siemens PLCs and ET200 remote I/O modules [49,50]. The dataset is publicly available at https://github.com/icsdataset/hai with a separate technical manual.

The testbed exposes four process areas: a boiler process (P1) on Emerson Ovation DCS, a turbine process (P2) on GE Mark VIe DCS, a water-treatment process (P3) on a Siemens S7-300 PLC, and an HIL simulation (P4) that synchronises the boiler and turbine processes with a virtual steam-turbine power-generation model and drives the P3 water-treatment pump and valve via a pumped-storage hydropower model. The cross-process coupling supplied by the HIL simulator is HAI's principal differentiator from single-process ICS datasets such as SWaT or BATADAL.

The HAI family has had four releases: HAI 1.0 / 20.07 (February and August 2020; 38 attack scenarios, 59 monitored SCADA tags), HAI 21.03 (2021; 50 attack scenarios, 78 tags), HAI 22.04 (2022;

 doi:10.20944/preprints202606.0145.v1



58 attack scenarios, 86 tags, with detection difficulty approximately four times higher than HAI 21.03 per the official documentation), and HAI 23.05 (2023; 52 attack scenarios, 86 tags) with the companion HAIEnd 23.05 (225 internal boiler-DCS tag values intended for endpoint-threat-detection research).

Each release provides CSV files with a one-second-resolution timestamp column, SCADA tag values, and a final attack-label column. From HAI 22.04 onward, per-process attack labels (attack_P1, attack_P2, attack_P3) were replaced by per-attack target metadata identifying which process each attack manipulated. Attacks are deliberately injected by an automated attack tool that manipulates feedback-control loop components (set-points, controller parameters, sensor readings, and actuator commands) at the SCADA-point level rather than the network-protocol level. The HAI 22.04 release in particular is designed for evaluating attacks with process-level impact but no obvious network-level signatures, which makes HAI complementary to network-IDS-oriented datasets such as Electra and ICS-NAD.

Multi-version documentation, the TaPR evaluation library, and public HAICon competition baselines together make HAI one of the most reusable datasets in the corpus. Its principal limitation is that the main releases provide process-level CSV time-series only and do not include raw PCAP files.

## 5. Datasets Analysis

This section analyzes the 23 selected ICS-specific datasets along three axes: (i) the basic structural aspects (environment, feature type, format, and OT protocol coverage), (ii) the attack types reported and their mapping to MITRE ATT&CK for ICS tactics, and (iii) the feature dimensionality and class imbalance characteristics. The corresponding summary tables are presented sequentially below, with prose commentary highlighting patterns, gaps, and structural limitations across these datasets.

### *5.1. Basic Aspects: Environment, Features, Format, and OT Protocols*

A comparison of the datasets by source, traffic, and format is summarized in Table 3. The "Environment" indicates how data is acquired, such as data collected from a physical testbed (a scaled-down physical process) or data created by a simulation for an industry application. The "Features" column examines the data type, whether it is N (Networking traffic) or P (Process Data), while the "Format" represents how the dataset is delivered for use by researchers, such as a CSV spreadsheet, network data in PCAP format, or others.

**Table 3.** Basic aspects across the studied datasets.

| Sr | Dataset | Environment | Features | Format | OT Protocols in Testbed |
|---|---|---|---|---|---|
| 01 | SWaT | Physical Testbed | P / N | CSV / PCAP | EtherNet/IP |
| 02 | WADI | Physical Testbed | P | CSV | EtherNet/IP, Modbus TCP |
| 03 | EPIC | Physical Testbed | P / N | CSV / PCAP | IEC 61850 (GOOSE, MMS), Modbus TCP |
| 04 | BATADAL | Simulation | P | CSV | No protocols |
| 05 | S317 | Physical Testbed | P / N | CSV / PCAP | EtherNet/IP |
| 06 | MSU-GP | Physical Testbed | P / N | CSV | Modbus RTU |
| 07 | MSU-PWR | Physical Testbed | P / N | CSV | DNP3, IEEE C37.118 |
| 08 | ICS-Flow | Simulation | P / N | CSV / PCAP | Modbus TCP |
| 09 | Lemay | Simulation | N | CSV / PCAP | Modbus TCP |
| 10 | Electra | Real Application | N | CSV, PCAP | Modbus, S7Comm, OPC |
| 11 | Rodofile | Physical Testbed | P / N | CSV / PCAP | S7Comm, DNP3 |
| 12 | WUSTL-IIoT | Physical Testbed | N | CSV | Modbus TCP |
| 13 | Edge-IIoTset | Physical Testbed | P / N | CSV / PCAP | Modbus TCP, MQTT |
| 14 | HIL-WDT | Physical Testbed | P / N | CSV / PCAP | Modbus TCP |
| 15 | X-IIoTID | Physical Testbed | N | CSV | Modbus TCP, MQTT, CoAP |

 doi:10.20944/preprints202606.0145.v1



| | | | | | |
|---|---|---|---|---|---|
| 16 | TEP | Simulation | P | R-Data | N/A |
| 17 | TLIGHT | Simulation | P | Unspecified | S7Comm |
| 18 | IUNO | Physical Testbed | P / N | Unspecified | OPC UA |
| 19 | ICS-NAD | Real Application | N | CSV / PCAP | Modbus, S7Comm, ABB (TCP) |
| 20 | ICS-ADD | Simulation | N | CSV / PCAP | Modbus TCP |
| 21 | HiTar | Simulation | N | Log | Modbus TCP |
| 22 | EDS | Physical Testbed | P / N | CSV | S7Comm |
| 23 | HAI | Physical Testbed + HIL | P | CSV | OPC UA |

*Note: Features and formats in this table are verified by inspecting each dataset whenever available.*

Table 3 summarizes four basic aspects of the 23 ICS-specific datasets: the context in which the data were generated, the types of features captured, the delivery format, and the OT protocols represented. The Environment column identifies data obtained from a physical testbed (a scaled-down version of an industrial process), data synthesized via simulation, and data from a live network in operation. The Features column indicates whether records contain network traffic (N), process (P), or both. The Format column indicates how each dataset is delivered (CSV tables, PCAP captures, raw logs, or others).

The distribution of the environment type is heavily weighted toward testbeds, with 14 out of 23 datasets (60.9%) coming from testbeds (HAI is included here because its HIL augmentation overlays physical-testbed components), 7 from simulation (30.4%), and 2 datasets (8.7%) classified as Real Application: Electra (railway traction substation) and ICS-NAD (multi-vendor real-environment test site). This dominance reflects the actual constraints that limit live-network capture in production ICS environments, and the persistent scarcity of operational data remains a fundamental obstacle to evaluating IDS realism in field conditions.

There are three categories of feature coverage: hybrid network+process (11 datasets, 47.8%), network-only (7 datasets, 30.4%), and process-only (5 datasets, 21.7%, including HAI). Of particular interest to IDS research is the hybrid set (SWaT, EPIC, S317, MSU-GP, MSU-PWR, ICS-Flow, Rodofile, Edge-IIoTset, HIL-WDT, IUNO, and EDS), which enables cross-layer correlation between network anomalies and process-state deviations. The most common delivery format is paired CSV/PCAP release (11 datasets, 47.8%), where raw captures are delivered alongside feature tables already prepared for ML. Other releases are pure CSV (8 datasets, 34.8%), usually process-only or feature-engineered. Less standard formats are found as R-Data (TEP), raw logs (HiTar), and undocumented formats (TLIGHT, IUNO). This collectively makes up only four datasets and raises reproducibility concerns where the underlying schema is not specified.

Protocol coverage is heavily concentrated. Modbus variants (TCP or RTU) are found in 13 of 23 datasets, reflecting the widespread use of Modbus in industrial deployments. S7Comm is found in five datasets: Electra, Rodofile, ICS-NAD, TLIGHT, and EDS; EtherNet/IP is found in three datasets: SWaT, WADI, and S317; and DNP3 is found in two datasets: MSU-PWR and Rodofile. Two datasets (Edge-IIoTset, X-IIoTID) are relevant to IIoT scenarios where MQTT is mentioned. Several operationally important protocols are severely underrepresented: OPC UA is found in IUNO and in HAI (where OPC UA serves as the data-acquisition gateway across heterogeneous DCS and PLC platforms). Recent releases add some new coverage, with ICS-NAD adding three vendors (Modbus, S7Comm, and ABB over TCP), and EDS adding S7Comm-based industrial Ethernet traffic. However, the skew toward Modbus and away from process automation (OPC UA) and Siemens-specific (Profinet) protocols remains a structural limitation for transfer learning and protocol generalization studies. The frequency of each OT protocol across the datasets is shown in Figure 6.

 doi:10.20944/preprints202606.0145.v1



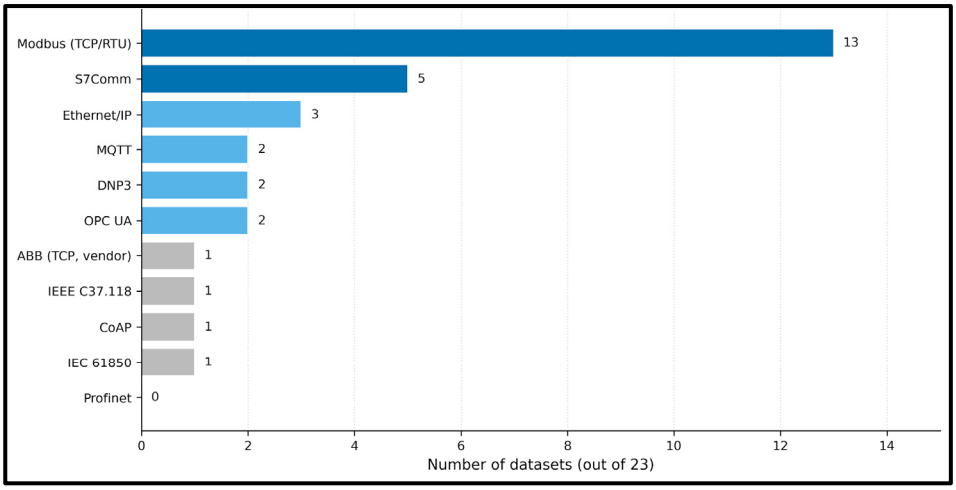


**Figure 6.** OT protocol coverage across the 23 selected datasets.

*5.2. Analysis of Attacks and Mapping to MITRE ATT&CK for ICS*

The selected Industrial Control System (ICS) datasets represent a range of attack scenarios. However, the scope, diversity, and level of detail of the reported attacks vary considerably across datasets. Therefore, examining the attack types included in each dataset is an important step in understanding their relevance for ICS intrusion-detection research.

In this study, the attacks reported in the 23 datasets were mapped to the MITRE ATT&CK for ICS tactic categories [18]. MITRE defines tactics as the adversary tactical objective, or the "why" behind an action, while techniques describe "how" that objective is achieved. Accordingly, the mapping in this study is based primarily on the apparent adversarial objective of each attack, rather than on the technical implementation alone. MITRE ATT&CK for ICS currently organizes ICS adversary behavior into 12 tactic categories: Initial Access, Execution, Persistence, Privilege Escalation, Evasion, Discovery, Lateral Movement, Collection, Command and Control, Inhibit Response Function, Impair Process Control, and Impact.

As shown in Table 4, the studied datasets cover a wide range of adversarial behaviors, from process-level manipulations (such as false data injection and actuator tampering) to network-level attacks (such as scanning, replay, man-in-the-middle, and DoS/DDoS). Most attacks are mapped to Inhibit Response Function and Impair Process Control, which reflects the cyber-physical nature of ICS security datasets. These tactics are especially relevant because many ICS attacks aim to manipulate control logic, disturb physical processes, mislead operators, or degrade system availability.

**Table 4.** Mapping of dataset attack types to MITRE ATT&CK for ICS tactics.

| Dataset | Attacks mentioned / normalized attack classes | MITRE ATT&CK for ICS tactic mapping |
|---|---|---|
| SWaT | False data injection; sensor spoofing; actuator manipulation; false process-state reporting; overflow/underflow; pump stress; wrong dosing | Impair Process Control; Inhibit Response Function; Impact |
| WADI | False data injection; sensor/actuator manipulation in water distribution | Impair Process Control; Inhibit Response Function; Impact |
| EPIC | False data injection; malicious/nuisance tripping; power interruption; physical damage; economic impact; malware-related disturbance | Execution; Impair Process Control; Inhibit Response Function; Impact |
| BATADAL | Pump/valve manipulation; tank-level manipulation; sensor falsification; replay/offset of process values | Impair Process Control; Inhibit Response Function; Impact |
| S317 | Reconnaissance; DoS/SYN flooding; Layer-1 DoS; false data injection/process manipulation | Discovery; Inhibit Response Function; Impair Process Control; Impact |
| MSU-GP | Reconnaissance; command injection; response/measurement injection; DoS | Discovery; Execution; Impair Process Control; Inhibit Response Function; Impact |

 doi:10.20944/preprints202606.0145.v1



| | | |
|---|---|---|
| MSU-PWR | Data injection; remote tripping; relay setting change; control/relay manipulation | Execution; Impair Process Control; Inhibit Response Function; Impact |
| ICS-Flow | Port scan; IP scan; DDoS; MitM-based false data injection; replay | Discovery; Collection; Impair Process Control; Inhibit Response Function; Impact |
| Lemay | Polite/loud reconnaissance; data exfiltration; replay; Modbus command injection/manipulation | Discovery; Collection; Execution; Impair Process Control; Inhibit Response Function |
| Electra | Function-code recognition; Modbus/S7 read and write attacks; response modification; command modification; error response manipulation; replay | Discovery; Collection; Execution; Impair Process Control; Inhibit Response Function |
| Rodofile | PLC memory/register manipulation; turning sub-processes on/off; conveyor-direction change; wash-tank mode manipulation; reactor-threshold manipulation; emergency stop; global reset | Execution; Impair Process Control; Inhibit Response Function; Impact |
| WUSTL-IIoT | Reconnaissance; command injection; DoS; backdoor | Discovery; Execution; Persistence; Command and Control; Impair Process Control; Inhibit Response Function |
| Edge-IIoTset | DoS/DDoS; information gathering; port scanning; OS fingerprinting; vulnerability scan; MitM (DNS/ARP spoofing); XSS; SQL injection; file upload; backdoor; password cracking; ransomware | Discovery; Collection; Initial Access; Execution; Persistence; Command and Control; Inhibit Response Function; Impact |
| X-IIoTID | Reconnaissance; weaponization; exploitation; lateral movement; command and control; exfiltration; tampering; crypto-ransomware; ransom DoS (RDoS) | Discovery; Initial Access; Execution; Persistence; Lateral Movement; Collection; Command and Control; Impair Process Control; Inhibit Response Function; Impact |
| HIL-WDT | MitM (ARP poisoning); DoS (TCP flood, ICMP flood, LAND); scanning (SYN, FIN, NULL, XMAS); physical attacks (water leaks from manual valves; sensor failures; pump failures) | Discovery; Collection; Impair Process Control; Inhibit Response Function; Impact |
| TEP | Process faults/anomalies rather than explicit cyber-attacks | Not directly MITRE-mappable; can only be loosely related to process anomaly/fault-detection evaluation, not cyber-attack tactics |
| TLIGHT | False data injection using Snap7; PLC memory/address modification; traffic-light state manipulation | Execution; Impair Process Control; Impact |
| IUNO | Reconnaissance/scanning; malicious OPC UA response/measurement manipulation; covert PLC/process deviation with apparently normal reported values | Discovery; Impair Process Control; Inhibit Response Function; Impact |
| ICS-NAD | Reconnaissance; DoS/DDoS; false data injection; MitM | Discovery; Collection; Impair Process Control; Inhibit Response Function; Impact |
| ICS-ADD | DNS tunneling C2; port scanning; password brute-forcing; Modbus scanning; ARP spoofing/MitM; Modbus FDI; DoS | Discovery; Initial Access; Execution; Persistence; Command and Control; Collection; Impair Process Control; Inhibit Response Function; Impact |
| HiTar | Probing; R2L; U2R; DoS | Discovery; Initial Access; Privilege Escalation; Execution; Inhibit Response Function; Impact |
| EDS | Information leakage; replay attack; command injection; sensor data tampering; control parameter tampering; multi-point attack; physical attack | Collection; Execution; Impair Process Control; Inhibit Response Function; Impact |
| HAI | False data injection (set-point and sensor manipulation); control-parameter tampering; actuator override; coordinated multi-loop attacks; stealthy attacks (HAI 22.04+) | Impair Process Control; Inhibit Response Function; Impact |

In contrast, datasets with broader IT/IIoT attack coverage, such as Edge-IIoTset, ToN_IoT [63], X-IIoTID, ICS-ADD, and HiTar, include attacks that extend beyond direct process manipulation. These datasets additionally cover tactics such as Initial Access, Execution, Persistence, Lateral Movement, Collection, and Command and Control. This indicates that the 23 datasets are not uniform in their attack representation: some are strongly process-centric, while others represent broader cyber or IIoT intrusion scenarios. Therefore, the choice of dataset should be aligned with the intended IDS research objective, whether the focus is on physical-process anomaly detection, network intrusion detection, or broader cyber-physical attack detection.

Three structural observations follow from Table 4. First, Discovery is the most consistently represented preparatory tactic across datasets that include any form of reconnaissance, scanning, or

 doi:10.20944/preprints202606.0145.v1



probing, including the recently corrected entries for ICS-ADD and IUNO. Second, the Impact tactic is mapped in 19 of the 23 datasets, confirming that current ICS datasets predominantly capture attacks with observable operational consequences rather than purely stealthy or long-dwell behavior. Third, data-theft scenarios are underrepresented across the datasets and remain an open gap for future dataset releases. The per-dataset tactic coverage is visualized as a heatmap in Figure 7, with column sums highlighting Inhibit Response Function, Impair Process Control, and Impact as the most populated tactics, and Privilege Escalation, Evasion, and Lateral Movement as the structural gaps.

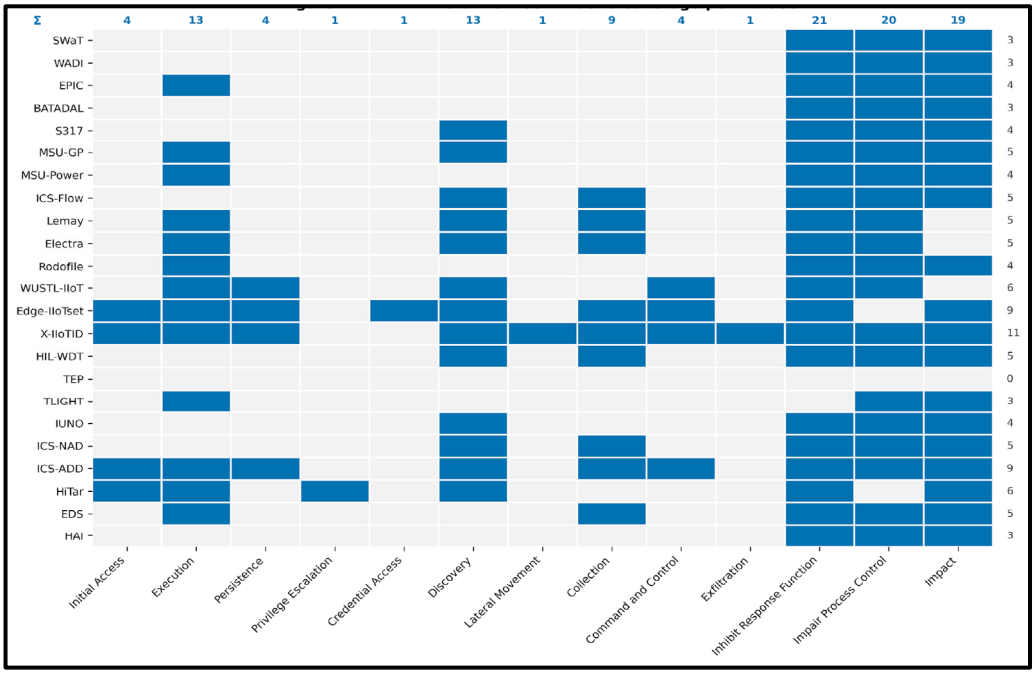


**Figure 7.** MITRE ATT&CK for ICS tactic coverage per dataset.

*5.3. Feature Dimensionality and Class Imbalance*

In machine-learning-based intrusion detection, dataset characteristics such as feature dimensionality, sample size, and class distribution directly influence model performance. A high number of features may increase the model representational capacity, but it may also increase computational cost and the risk of overfitting if the dataset is not sufficiently large. Similarly, class imbalance is a common issue in ICS/OT intrusion-detection datasets, where normal operating data often dominate attack samples. This imbalance can bias learning algorithms toward the majority class and lead to misleadingly high accuracy while reducing the detection performance for minority attack classes. Previous studies have also highlighted that imbalance affects IDS performance and that other metrics can be more informative [64,65]. Table 5 summarizes the number of features, number of instances, anomaly percentage, and the normal-to-anomaly ratio for the selected datasets. Values marked as Not reported indicate that the corresponding information was not explicitly available in the dataset paper or public release documentation.

**Table 5.** Feature counts, instance counts, and class imbalance of the selected datasets.

| Sr | Dataset | Features | Instances | Anomaly % | Imbalance Ratio |
|---|---|---|---|---|---|
| 01 | SWaT | 51 | 946,722 | 12% | 7.3 : 1 |
| 02 | WADI | 123 | 1,221,372 | 5.8% | 16.3 : 1 |
| 03 | EPIC | Not reported | Not reported | Not reported | Not reported |
| 04 | BATADAL | 43 | 15,027 | 14.35% | 7.5 : 1 |
| 05 | S317 | Not reported | Not reported | Not reported | Not reported |

 doi:10.20944/preprints202606.0145.v1



| | | | | | |
|---|---|---|---|---|---|
| 06 | MSU-GP | 26 | 274,627 | 22% | 3.5 : 1 |
| 07 | MSU-PWR | 128 | 78,377 (binary var.) | 68% | 0.5 : 1 |
| 08 | ICS-Flow | 50 | 45,719 | 19.7% | 4.07 : 1 |
| 09 | Lemay | Not reported | 134,690 | Not reported | Not reported |
| 10 | Electra | Not reported | Not reported | 7.87% | 11.7 : 1 |
| 11 | Rodofile | 8 | 16.7 M | Not reported | Not reported |
| 12 | WUSTL-IIoT | 41 | 1,194,464 | 7.3% | 12.7 : 1 |
| 13 | Edge-IIoTset | 61 | 2,219,201 | 44% | 1.3 : 1 |
| 14 | HIL-WDT | 55 | 7,747 | 15.85% | 5.31 : 1 |
| 15 | X-IIoTID | 59 | 820,834 | 49% | 1.06 : 1 |
| 16 | TEP | 52 | 15,330,000 | Not reported | Not reported |
| 17 | TLIGHT | 10 | 2,800 / 8,000 | Not reported | Not reported |
| 18 | IUNO | Not reported | Not reported | Not reported | Not reported |
| 19 | ICS-NAD | 60 | Not reported | Not reported | Not reported |
| 20 | ICS-ADD | Not reported | Not reported | Not reported | Not reported |
| 21 | HiTar | 39 | 15,842 | Not reported | Not reported |
| 22 | EDS | 47 | 843,321 | Not reported | Not reported |
| 23 | HAI | 86 / 225 | 1.18 M | Not reported | Not reported |

- *Normal-to-anomaly ratio = (100 − A%) / A% : 1.*
- *The Lemay 134,690 figure refers to labeled records from the exfiltration scenario PCAP.*
- *The 47 EDS parameters comprise 1 timestamp, 8 control, 24 digital, and 14 analog values; the 22 continuous numeric features (8 control + 14 analog) are the subset typically used for ML modeling.*
- *TEP: the 52 features reported here are the process variables only; the released R dataframes contain three additional identifier columns (faultNumber, simulationRun, sample).*
- *For HAI, 86 (HAI 23.05) / 225 (HAIEnd).*
- *The SWaT (946,722 records) and WADI (1,221,372 records) instance counts in Table 5 reflect the iTrust dataset releases accessed during this study [21] (accessed 14 May 2026). Published iTrust summaries report different totals across release iterations; the values in Table 5 correspond to the files retained for this review.*
- *BATADAL anomaly % (14.35%) in Table 5 is computed over the labeled subsets only (Training 2 + Test = 6,266 hourly samples with ≈ 899 attack-labeled hours), not over all 15,027 samples in the released corpus.*

The results show that most datasets are imbalanced toward normal operation, which is expected in ICS environments where attacks are rare compared with stable process operation. However, some datasets, such as MSU-PWR (0.5 : 1) and X-IIoTID (1.06 : 1), show a more balanced distribution because they were generated with many attack scenarios for security evaluation. Therefore, model evaluation on these datasets should not rely only on accuracy. Instead, recall, F1-score, and macro-F1 should be reported to avoid overestimating model performance on majority classes.

Across Tables 3, 4, and 5, three patterns deserve emphasis before turning to the documentation-completeness and decision-aid rubric analyses in Section 6. First, network-only datasets dominate the recent IIoT-oriented releases (ICS-ADD, HiTar, X-IIoTID), while hybrid network+process coverage remains concentrated in physical-testbed releases (EPIC, MSU-GP, MSU-PWR, HIL-WDT, EDS). Second, the most diverse attack inventories (Edge-IIoTset, X-IIoTID, ICS-ADD) belong to datasets explicitly designed around adversary lifecycles rather than process-specific scenarios, which has direct implications for benchmarking and APT-oriented evaluation. Third, the imbalance picture splits the group into three classes: severely imbalanced ICS datasets (WADI, WUSTL-IIoT, Electra), moderately imbalanced datasets (SWaT, BATADAL, HIL-WDT, ICS-Flow), and near-balanced or attack-dominant datasets (MSU-PWR, Edge-IIoTset, X-IIoTID). The decision-aid scoring rubric in Section 6 makes these patterns explicit by inverse-scoring imbalance as a usability criterion.

 doi:10.20944/preprints202606.0145.v1



## 6. Results and Findings

The studied ICS datasets exhibit several limitations that can hinder IDS performance. These limitations fall into three categories: data representation, dataset management, and attack-scenario issues.

### *6.1. Data Representation Issues*

The use of simulated or testbed environments may not fairly represent real-world ICS behavior, which can be considered a limitation in datasets. Models trained on such data risk overfitting, limiting their transferability to operational deployments. A related concern arises where attack traffic is generated synthetically, potentially failing to capture the timing characteristics and protocol-level responses of live systems.

As shown in Table 6, the majority of datasets exhibit class imbalance, which will result in bias of classifiers toward the majority class and complicate effective model training, particularly for datasets with limited overall size. Multiple mitigation strategies are commonly used in the ML field at the pre-processing stage, including oversampling, undersampling, and cost-sensitive learning [66].

**Table 6.** *Attack generation, realism, imbalance, and diversity of the selected datasets.*

| Sr | Dataset | Generation / Evidence | Realism | Imbalance (1) | Diversity (2) |
|---|---|---|---|---|---|
| 01 | SWaT | Manual documented attack scenarios; tools NR | Injected | Moderate | Moderate |
| 02 | WADI | Manual/scripted process manipulation; tools NR | Injected | Severe | Moderate |
| 03 | EPIC | Smart-grid attack scenarios; tools NR | Injected | NR (3) | Moderate |
| 04 | BATADAL | epanetCPA / EPANET attack simulation | Synthetic | Moderate | Moderate |
| 05 | S317 | S3 attacker-team tools/scripts; not fully enumerated | Injected | NR (3) | Moderate |
| 06 | MSU-GP | Custom Modbus attack scenarios; tools NR | Injected | Mild | Moderate |
| 07 | MSU-PWR | Attack/event scenarios; Snort is logging/IDS | Injected | Mild; attack-heavy | Moderate |
| 08 | ICS-Flow | Scripted scan, DoS/DDoS, replay, MitM-FDI; ICSFlowGenerator for extraction | Synthetic | Moderate | Moderate |
| 09 | Lemay | SCADA sandbox; Modbus traffic and exploit/command scenarios | Synthetic | NR (3) | Diverse |
| 10 | Electra | Wireshark capture; Python/Scapy parsing; attacker node for injected attacks | Real+Injected | Severe | Diverse |
| 11 | Rodofile | Python/Snap7-based PLC memory/register manipulation on physical S7Comm testbed | Injected | NR (3) | Moderate |
| 12 | WUSTL-IIoT | Backdoor, command injection, DoS, reconnaissance; tools not fully specified | Injected | Severe | Moderate |
| 13 | Edge-IIoTset | Reported IoT/IIoT attack tools/scripts for scan, brute force, web, malware, DoS | Injected | Mild | Diverse |
| 14 | HIL-WDT | Cyber and physical scenarios: MitM, flood, scan, leak, failure; tools partly NR | Injected | Moderate | Diverse |
| 15 | X-IIoTID | Lifecycle IIoT scenarios; Zeek/OSSEC are parsing/alert tools | Injected | Mild | Diverse |
| 16 | TEP | MATLAB/Simulink TEP faults; not explicit cyberattacks | Synthetic | Mild (4) | N/A — faults |

 doi:10.20944/preprints202606.0145.v1



| | | | | | |
|---|---|---|---|---|---|
| 17 | TLIGHT | Snap7 alteration of selected Siemens PLC addresses | Synthetic | NR (3) | Mild |
| 18 | IUNO | OPC UA packet/application attacks and covert physical deviation | Injected | NR (3) | Moderate |
| 19 | ICS-NAD | 20 deliberate ICS attacks: recon, DoS/DDoS, FDI, MitM; tools NR | Real+Injected | NR (3) | Diverse |
| 20 | ICS-ADD | Cyber Kill Chain-style pentest sequence; OSSIM/Suricata are monitoring tools | Injected | NR (3) | Diverse |
| 21 | HiTar | AREZZO simulator logs; attack_labelling.sh; no live attack tool | Synthetic | Moderate | Moderate |
| 22 | EDS | snap7/Python acquisition; leakage, replay, command injection, tampering, physical | Injected | NR (3) | Diverse |
| 23 | HAI | Automated attack tool; 14 attack primitives → 38–58 scenarios across HAI releases | Injected | NR (3) | Diverse |

*(1) Imbalance degree follows Table 5: Mild ≤ 4:1; Moderate = 4:1 to 9:1; Severe ≥ 10:1. (2) Attack diversity is based on normalized attack categories identified from the dataset descriptions and the MITRE-mapping table: Mild ≤ 2 categories; Moderate = 3–4 categories; Diverse ≥ 5 categories. For TEP, the value is marked N/A because the dataset is primarily a process-fault dataset. (3) NR: not reported or not reliably derivable from author or dataset documentation. (4) For TEP, Imbalance is marked Mild because the 21 fault classes are balanced by design (500 simulation runs × equal sample counts per fault number); the Anomaly % / Imbalance Ratio in Table 5 is NR because TEP is fault-classification rather than binary normal/anomaly.*

A further representation gap is the absence of raw packet capture (PCAP) files in several datasets. Without raw data, researchers cannot recover low-level information such as inter-packet timing, protocol flags, and flow-level metadata, which is often critical for protocol-aware detection methods.

### *6.2. Dataset Management Issues*

A persistent challenge across these datasets is the absence of comprehensive documentation. Many datasets lack a system architecture description, a clear account of the attacks implemented, and details of the testbed configuration. Such information is essential for reproducibility and practical benchmarking.

Restricted availability compounds this problem. The sensitive nature of ICS environments leads some organizations to withhold detailed datasets to avoid exposing proprietary network configurations or operational procedures, limiting the data accessible to the broader research community.

A further practical issue is dataset fragmentation, where several releases distribute traffic captures, labels, and metadata across multiple files. This increases the required effort for pre-processing and analysis.

### *6.3. Attack Landscape Issues*

Attack realism is evaluated using the dataset descriptions and availability categories already reported in this manuscript. Based on Table 2, sixteen of the 23 selected ICS/IIoT-related datasets are publicly available, four are available on request, and three are restricted or not clearly obtainable. Most available labeled datasets contain attacks executed in controlled physical testbeds, hardware-in-the-loop environments, or simulations. This makes them valuable for IDS benchmarking, but it also means they should not be interpreted as uncontrolled real-incident datasets. Electra and ICS-NAD are the only two datasets classified in Table 2 as Real Application. Electra captures traffic from an electric traction substation in the railway sector, while ICS-NAD captures multi-vendor real-environment traffic with deliberately labelled attacks; neither is an uncontrolled incident capture.

 doi:10.20944/preprints202606.0145.v1



The reviewed datasets also show uneven attack coverage. Traditional ICS datasets are mostly process-centric and emphasize false data injection, replay, actuator/sensor manipulation, command or response manipulation, DoS/DDoS, and reconnaissance. Newer IIoT-oriented datasets, such as Edge-IIoTset, X-IIoTID, ICS-ADD, and HiTar, extend the scope to encompass broader IT/OT intrusion behavior, including scanning, brute-force attacks, backdoors, command-and-control activity, ransomware-type activity, and privilege-oriented attacks. However, this broader coverage does not automatically imply stronger process realism; some of these datasets contain rich cyber labels but weaker closed-loop physical-process semantics.

Protocol dependence remains another limitation. Several attacks are tied to specific industrial communication stacks such as Modbus/TCP, S7Comm, EtherNet/IP, OPC UA, or IEC 61850. As discussed in the protocol analysis, Modbus-based datasets dominate the available datasets, while OPC UA, Profinet, and vendor-specific traffic are much less represented. This skew limits the ability to claim protocol-level generalization from results obtained on a single dataset or protocol family.

Insider threats and multi-stage adversary behavior remain underrepresented. A few datasets provide richer adversarial structure, such as S317 through a live attack-defense exercise, X-IIoTID through lifecycle-oriented IIoT attack classes, and ICS-ADD through a Cyber Kill Chain-style sequence. Nevertheless, most datasets model short, isolated attack episodes rather than campaigns involving reconnaissance, initial access, persistence, lateral movement, process manipulation, operator deception, and physical impact. This constrains the evaluation of IDS methods aimed at detecting long-horizon attack progression.

The documentation of attack generation is inconsistent. Some papers clearly describe the attack scenario and affected process variables, but do not always report the exact tools used. Capture, monitoring, parsing, and IDS/SIEM tools such as Wireshark, Zeek, Snort, OSSIM, Suricata, and OSSEC are treated as instrumentation rather than attack-generation tools unless the source explicitly identifies them as part of the attack execution. This distinction improves reproducibility and prevents overclaiming about dataset realism or adversary tooling [1,10,11,64,65].

The selected datasets support useful IDS benchmarking, but attack realism varies. Physical testbeds provide cyber-physical fidelity at the cost of scale; simulations offer safety and repeatability but lose authenticity; real-network captures remain scarce. When reporting IDS results, accuracy alone is misleading. Attack diversity, class imbalance, protocol coverage, and documentation quality should be reported alongside it.

### *6.4. Quantitative Analysis of Datasets Documentation*

The documentation-completeness analysis evaluates how well each dataset can be understood, reproduced, and reused by other researchers. Following the 0–7 checklist defined in the methodology, one point is assigned for each documented element: (System architecture, protocol/communication stack, attack or anomaly methodology, record or class counts, feature description, labeling or ground-truth method, and availability conditions).

The results show that documentation quality is generally strong across the selected datasets, with an average score of 6.04 / 7.

The scoring also reveals recurring documentation gaps. The most common missing elements are detailed record/class counts and feature descriptions, while architecture, attack methodology, and availability are more frequently reported. This confirms that many ICS datasets are usable for benchmarking, but not always fully reproducible without additional preprocessing effort, manual interpretation, or consultation of repository files. This observation is consistent with prior ICS dataset surveys, which emphasize that dataset selection should consider not only realism and attack coverage, but also documentation, labeling transparency, availability, and reproducibility.

### *6.5. Decision-Aid Scoring Rubric and Per-Direction Recommendations*

The taxonomic axes defined in Section 3 and applied in Tables 2, 3, 6, and 7 describe each dataset from several perspectives. However, practitioners normally select a dataset for a specific task rather

 doi:10.20944/preprints202606.0145.v1



than for a generic comparison. This section can convert the manuscript existing taxonomy into a transparent 0–15 decision aid. The rubric is not intended to identify a universal best dataset; it is intended to expose trade-offs among realism, attack breadth, imbalance, documentation, and reusability.

**Table 7.** Dataset documentation scoring.

| Sr | Dataset | Documentation Score * | Doc. Tier ** | Notes |
|---|---|---|---|---|
| 01 | SWaT | 7/7: Arch, Proto, Atk, Cnt, Feat, Lbl, Avail | Comprehensive | Comprehensive iTrust documentation covering architecture, protocol, attack scenarios, counts, features, labels, and access conditions. |
| 02 | WADI | 7/7: Arch, Proto, Atk, Cnt, Feat, Lbl, Avail | Comprehensive | Comprehensive iTrust documentation covering architecture, protocol, attack scenarios, counts, features, labels, and access conditions. |
| 03 | EPIC | 4/7: Arch, Proto, Atk, Avail | Moderate | Architecture, protocol, attack scope, and access are clear; record counts, feature schema, and label details remain thin. |
| 04 | BATADAL | 6/7: Arch, Atk, Cnt, Feat, Lbl, Avail | Comprehensive | Competition dataset with well-documented C-Town simulation, variables, attacks, labels, counts, and availability. |
| 05 | S317 | 4/7: Arch, Proto, Atk, Avail | Moderate | CTF/event paper documents context, protocol setting, and attack activity; public schema, counts, and label details are limited. |
| 06 | MSU-GP | 6/7: Arch, Proto, Atk, Cnt, Lbl, Avail | Comprehensive | Mississippi State gas-pipeline SCADA dataset; Modbus attack classes, counts, labels, and access are documented, but feature-schema detail is limited. |
| 07 | MSU-PWR | 7/7: Arch, Proto, Atk, Cnt, Feat, Lbl, Avail | Comprehensive | Heterogeneous power-system measurements, logs, events, attack scenarios, counts, and labels are documented. |
| 08 | ICS-Flow | 7/7: Arch, Proto, Atk, Cnt, Feat, Lbl, Avail | Comprehensive | ICSFlowGenerator and the dataset paper document raw packets, flow features, process logs, attack scenarios, labels, and access. |
| 09 | Lemay | 6/7: Arch, Proto, Atk, Cnt, Lbl, Avail | Comprehensive | Modbus/TCP SCADA sandbox documented with PCAPs and labels; feature-extraction/schema detail is limited. |
| 10 | Electra | 6/7: Arch, Proto, Atk, Feat, Lbl, Avail | Comprehensive | Railway traction-substation traffic with Modbus/S7Comm features and labels; detailed per-class/record counts are not fully reported. |
| 11 | Rodofile | 5/7: Arch, Proto, Atk, Lbl, Avail | Moderate | S7Comm process-control testbed is documented and labeled; record-count and feature-schema details are limited. |
| 12 | WUSTL-IIoT | 7/7: Arch, Proto, Atk, Cnt, Feat, Lbl, Avail | Comprehensive | Water-storage IIoT testbed, Modbus/TCP traffic, selected network-flow features, exact instance counts, labels, and availability are documented. |
| 13 | Edge-IIoTset | 7/7: Arch, Proto, Atk, Cnt, Feat, Lbl, Avail | Comprehensive | Layered IoT/IIoT architecture, protocol coverage, 61 selected features from 1,176 raw features, counts, labels, and access are documented. |

 doi:10.20944/preprints202606.0145.v1



| | | | | |
|---|---|---|---|---|
| 14 | HIL-WDT | 6/7: Arch, Proto, Atk, Feat, Lbl, Avail | Comprehensive | HIL water-distribution setup, Modbus/TCP, physical/network features, attack scenarios, labels, and access are documented; exact record counts are less explicit. |
| 15 | X-IIoTID | 7/7: Arch, Proto, Atk, Cnt, Feat, Lbl, Avail | Comprehensive | Brown-IIoTbed architecture, multi-protocol sources, attack taxonomy, counts, feature construction, labels, and access are documented. |
| 16 | TEP | 6/7: Arch, Atk, Cnt, Feat, Lbl, Avail | Comprehensive | Process-fault/anomaly simulation; variables, fault numbers, simulation runs, labels, and Dataverse access are documented; no OT protocol applies. |
| 17 | TLIGHT | 5/7: Arch, Proto, Atk, Cnt, Lbl | Moderate | PLC traffic-light scenario, Snap7 address modification, labels, and dataset sizes are described; feature schema and public availability remain limited. |
| 18 | IUNO | 5/7: Arch, Proto, Atk, Lbl, Avail | Moderate | OPC UA water-process scenario and attack concepts are documented; feature/count details and released-file structure remain limited. |
| 19 | ICS-NAD | 6/7: Arch, Proto, Atk, Feat, Lbl, Avail | Comprehensive | Scientific Data 2026 documentation covers multi-vendor scenarios, protocols, 20 attack types, 60 features, labels, dataset size, and access. |
| 20 | ICS-ADD | 6/7: Arch, Proto, Atk, Feat, Lbl, Avail | Comprehensive | SCADA/PLC/SIEM/NIDS pipeline, Modbus/TCP, attacks, features/log sources, labels, and availability are documented; per-class counts are limited. |
| 21 | HiTar | 6/7: Arch, Proto, Cnt, Feat, Lbl, Avail | Comprehensive | AREZZO simulator/log-based structure, Modbus/TCP service, counts, features, labels, and access are documented; attack-generation methodology is less detailed. |
| 22 | EDS | 6/7: Arch, Proto, Atk, Feat, Lbl, Avail | Comprehensive | Full-hardware ethanol-distillation testbed, S7Comm acquisition, 47 parameters, labels, and seven threat categories are documented; per-class counts are incomplete. |
| 23 | HAI | 7/7: Arch, Proto, Atk, Cnt, Feat, Lbl, Avail | Comprehensive | Multi-version documentation; two USENIX CSET papers; per-version statistics in official README + technical manual; eTaPR evaluation library; public HAICon competition baselines. |

** Documentation completeness is scored 0–7, one point per documented axis: architecture, protocol/communication stack, attack or anomaly methodology, counts, feature description, label/ground-truth method, and availability/access condition.*
*** Documentation tier cutoffs: 0–2 = Mild; 3–5 = Moderate; 6–7 = Comprehensive.*

This interpretation is consistent with prior ICS dataset-survey guidance that dataset selection should consider the testbed or operational setting, protocol coverage, attack realism, documentation, availability, and suitability for the intended IDS task rather than relying on popularity or accuracy results alone [1]. It also reflects the imbalance concern discussed earlier in this manuscript, where precision, recall, F1-score, macro-F1, and PR-AUC are more informative than accuracy for skewed IDS datasets [64,65].

As an illustration, X-IIoTID achieves a near-balanced class distribution (1.06 : 1) that makes it directly convenient for supervised ML experiments, while datasets such as WADI or WUSTL-IIoT exhibit severe imbalance that requires explicit sampling or class-weighting strategies.

 doi:10.20944/preprints202606.0145.v1



6.5.1. Rubric Definition

Each dataset is scored from 0 to 3 on five criteria. The total composite score is the sum of the five criteria, with a nominal maximum of 15. For datasets where the Imbalance ratio is unreported in primary documentation, the Imbalance Usability axis is excluded from the composite and the dataset's total is renormalised to /12 (also reported as a percentage of the achievable maximum). Equal weighting is used to keep the rubric auditable; the final recommendation still depends on the research direction.

- **Realism:** derived from the Realism column in Table 6, which is based on the dataset descriptions and the environment/evidence reported in Table 2. The score jointly reflects the operational realism of the environment and the provenance of the attack traffic. Scoring: 3 = real operational network with naturally-occurring or incident-captured attack traffic that was not deliberately introduced by the dataset authors; 2 = real operational environment in which the attack traffic was deliberately introduced by the dataset authors (e.g., Electra and ICS-NAD); 1 = deliberately injected attacks in a controlled physical or hardware-in-the-loop testbed rather than an operational network; 0 = synthetic or pure-simulation environment with no physical hardware in the loop. Because Realism = 3 requires naturally-occurring or incident-captured traffic, so no dataset reaches the nominal maximum composite (15/15 for datasets with reported Imbalance, 12/12 for datasets where Imbalance is excluded). Edge-IIoTset's 86.7% is the highest score achieved.
- **Attack diversity:** derived from the Diversity column in Table 6. Scoring: 3 = Diverse (five or more normalized attack categories); 2 = Moderate (three to four categories); 1 = Mild (one to two categories); 0 = not applicable or not cyber-attack-oriented, as in pure process-fault datasets.
- **Imbalance usability:** derived from the Imbalance column in Table 6 but inverse-scored. Scoring: 3 = Mild (normal-to-attack ratio ≤ 4:1); 2 = Moderate (4:1 to 9:1); 1 = Severe (> 10:1); NR (not reported) = this axis is excluded from the composite for that dataset, and the total is renormalised to /12 (see introductory paragraph). These cut-offs align with the thresholds applied in Table 6.
- **Documentation:** derived from the documentation-completeness score in Table 7. Scoring: 3 = 7/7; 2 = 5–6/7; 1 = 3–4/7; 0 = 0–2/7.
- **Reproducibility:** derived from the Availability column in Table 2 and the Format column in Table 3. Scoring: 3 = public dataset with raw artifacts available (PCAP for network-oriented datasets, raw time-series for process-oriented datasets, or both); 2 = public feature-only or preprocessed release, or request-based access with raw artifacts; 1 = request-based access feature-only; 0 = restricted, offline, or not clearly obtainable.

Table 8 presents the rubric scores and sorted by Total in descending order. Where multiple datasets share the same Total, the secondary sort is by Realism (higher first), then Documentation, then Reproducibility, then the original serial number.

**Table 8.** Use-case rubric scores per dataset (sorted by score descending).

| Rank | Dataset | Realism | Diversity | Imb. Usg. | Doc. | Repro. | Total |
|---|---|---|---|---|---|---|---|
| 1 | Edge-IIoTset | 1 | 3 | 3 | 3 | 3 | 13 |
| 2 | X-IIoTID | 1 | 3 | 3 | 3 | 2 | 12 |
| 3 | MSU-PWR | 1 | 2 | 3 | 3 | 2 | 11 |
| 4 | HIL-WDT | 1 | 3 | 2 | 2 | 3 | 11 |
| 5 | ICS-NAD | 2 | 3 | — | 2 | 3 | 10 |
| 6 | SWaT | 1 | 2 | 2 | 3 | 2 | 10 |
| 7 | MSU-GP | 1 | 2 | 3 | 2 | 2 | 10 |
| 8 | ICS-Flow | 0 | 2 | 2 | 3 | 3 | 10 |
| 9 | WUSTL-IIoT | 1 | 2 | 1 | 3 | 2 | 9 |
| 10 | HAI | 1 | 3 | — | 3 | 3 | 10 |

 doi:10.20944/preprints202606.0145.v1



| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 11 | ICS-ADD | 0 | 3 | — | 2 | 3 | 8 |
| 12 | BATADAL | 0 | 2 | 2 | 2 | 3 | 9 |
| 13 | Electra | 2 | 3 | 1 | 2 | 0 | 8 |
| 14 | WADI | 1 | 2 | 1 | 3 | 1 | 8 |
| 15 | EDS | 1 | 3 | — | 2 | 2 | 8 |
| 16 | Rodofile | 1 | 2 | — | 2 | 3 | 8 |
| 17 | HiTar | 0 | 2 | 2 | 2 | 2 | 8 |
| 18 | Lemay | 0 | 3 | — | 2 | 3 | 8 |
| 19 | TEP | 0 | 0 | 3 | 2 | 3 | 8 |
| 20 | S317 | 1 | 2 | — | 1 | 2 | 6 |
| 21 | EPIC | 1 | 2 | — | 1 | 2 | 6 |
| 22 | IUNO | 1 | 2 | — | 2 | 0 | 5 |
| 23 | TLIGHT | 0 | 1 | — | 2 | 0 | 3 |

- *TEP is scored 0 for attack diversity because it is primarily a process-fault/anomaly dataset, not an explicit cyber-attack dataset.*
- *The score should be read as a decision-support score, not as an absolute scientific ranking. Domain and protocol fit may override the total score in specialized studies.*

Under the renormalization introduced in §6.5.1 (Imbalance excluded where unreported; total renormalized to /12) and the updated Reproducibility scoring, the effective ranking by percentage of achievable maximum is: 1. Edge-IIoTset (86.7%, 13/15); 2. ICS-NAD (83.3%, 10/12); 3. HAI (83.3%, 10/12); 4. X-IIoTID (80.0%, 12/15); 5. MSU-PWR (73.3%, 11/15); 6. HIL-WDT (73.3%, 11/15); 7. SWaT, Rodofile, MSU-GP, EDS, ICS-Flow, Lemay, ICS-ADD (all 66.7%); 14. WUSTL-IIoT (60.0%, 9/15); 15. BATADAL (60.0%, 9/15); 16. Electra, WADI, TEP, HiTar (all 53.3%); 20. EPIC (50.0%, 6/12); 21. S317 (50.0%, 6/12); 22. IUNO (5/12, 41.7%); 23. TLIGHT (3/12, 25.0%).

### 6.5.2. Per-Use-Case Recommendations

The recommendations below interpret the corrected rubric in terms of five common research and engineering tasks. A dataset is recommended when it meets the primary criteria for that task; where no dataset satisfies all criteria, the limitation is stated explicitly.

**(1) Training and general supervised IDS development.** Priority criteria: documentation and reproducibility. The strongest choices are Edge-IIoTset [39], X-IIoTID [41], MSU-PWR [29,30], SWaT [19], MSU-GP [27,28], ICS-Flow [12], and BATADAL [25]. These datasets combine usable documentation with accessible or requestable artifacts and a non-severe class distribution. WUSTL-IIoT [36] and WADI [22] remain useful, but their severe imbalance requires careful sampling, class weighting, or metric selection. TEP [42] is suitable for process-anomaly or fault-detection studies, but it should not be presented as a full cyber-network IDS benchmark.

**(2) Benchmarking and comparative evaluation across publications.** Priority criteria: realism, documentation, and attack diversity. ICS-NAD [45] is the strongest candidate, scoring 3/3 on attack diversity and reproducibility and 2/3 on realism and documentation (10/12 = 83.3% after Imbalance NR-exclusion; see §6.5.1). Realism is capped at 2/3 because no dataset in this corpus contains naturally occurring incident traffic, and imbalance statistics for ICS-NAD are not fully reported. Electra [33] is also valuable for realistic railway traction-substation traffic with diverse attacks, but its restricted availability reduces reproducibility. Because very few datasets meet all three benchmarking criteria, a practical benchmark suite should pair realism-focused datasets such as ICS-NAD or Electra with high-reproducibility datasets such as Edge-IIoTset [39], X-IIoTID [41], HIL-WDT [40], ICS-Flow [12], or ICS-ADD [46].

**(3) Protocol-specific studies.** Priority criteria: the target protocol, documentation, and reproducibility. For Modbus: ICS-Flow [12], HIL-WDT [40], ICS-NAD [45], ICS-ADD [46], Edge-IIoTset [39], X-IIoTID [41], Lemay [32], MSU-GP [27,28], WUSTL-IIoT [36], and SWaT [19]. For

 doi:10.20944/preprints202606.0145.v1



S7Comm, the strongest accessible choices are Rodofile [34], ICS-NAD [45], and EDS [48], with TLIGHT [43] as a specification-level reference for Siemens S7 PLC scenarios. For EtherNet/IP, SWaT [19], WADI [22], and S317 [26] are the relevant iTrust-linked candidates, subject to request access and artifact availability. For IEC 61850, EPIC [24] is the only selected candidate, but its documentation and the limitations of the released schema should be acknowledged. For OPC UA, IUNO [44] and HAI [49,50] are the most relevant candidates, although availability and file-structure limitations reduce reproducibility. For DNP3, MSU-PWR [29,30] and Rodofile [34] are the relevant selected datasets. For MQTT/CoAP-oriented IIoT studies, X-IIoTID [41] and Edge-IIoTset [39] are the main candidates.

**(4) IIoT and smart-manufacturing contexts.** Priority criteria: IIoT-relevant architecture or protocols, attack diversity, and documentation. X-IIoTID [41] and Edge-IIoTset [39] are the choices because they were designed around heterogeneous IIoT environments and contain broad attack taxonomies. WUSTL-IIoT [36] is appropriate for water-storage/Modbus-based IIoT intrusion detection with clear documentation, but imbalance must be handled carefully. ICS-ADD [46] is useful when the study requires SCADA/PLC traffic together with SIEM/NIDS-style monitoring outputs. HiTar [47] is suitable for log-based smart-manufacturing or transfer-learning studies.

**(5) APT, multi-stage, and kill-chain-oriented evaluation.** Priority criteria: attack diversity and realism, with special attention to whether the dataset captures staged adversary progression rather than isolated attack episodes. No selected dataset fully represents real APT campaigns in operational ICS networks. ICS-NAD [45] is the best realism/diversity candidate, and S317 [26] contributes a live attacker-team exercise context. X-IIoTID [41] and ICS-ADD [46] are recommended as structured alternatives because they include lifecycle or Cyber Kill Chain-style attack organization, even though their attacks are deliberately injected. Electra [33] can support this use case when realistic substation traffic is required and access is available. Edge-IIoTset [39] adds broad attack variety, but it should be used cautiously for APT claims because breadth of attack labels is not the same as a validated multi-stage ICS campaign. HAI [49,50] complements these candidates for studies focused on stealthy process-level attacks rather than network-protocol exploits, particularly from the HAI 22.04 release onward.

Overall, the rubric under renormalisation produces the following composite ranking. Edge-IIoTset obtains the highest score (86.7%, 13/15); ICS-NAD and HAI follow at 83.3% (each 10/12, after Imbalance NR-exclusion); X-IIoTID is at 80% (12/15); MSU-PWR and HIL-WDT at 73.3% (11/15). A broad mid-tier at 66.7% comprises SWaT, MSU-GP, ICS-Flow, ICS-ADD, EDS, Rodofile, and Lemay; followed by WUSTL-IIoT and BATADAL at 60.0%, Electra, WADI, TEP, and HiTar at 53.3%, EPIC and S317 at 50.0%, IUNO at 41.7%, and TLIGHT at 25.0%. The full composite ranking, with each bar segmented into the five rubric criteria, is shown in Figure 8 (the figure regenerates from the updated Table 8 values).

However, the total score should never be used alone. A lower-scoring dataset can still be the most appropriate choice when it uniquely matches the required domain or protocol; for example, EPIC remains important for IEC 61850-oriented smart-grid studies, IUNO for OPC UA-focused studies, and TEP for process-fault/anomaly detection. Conversely, a high total score does not guarantee process realism or suitability for every OT deployment.

 doi:10.20944/preprints202606.0145.v1



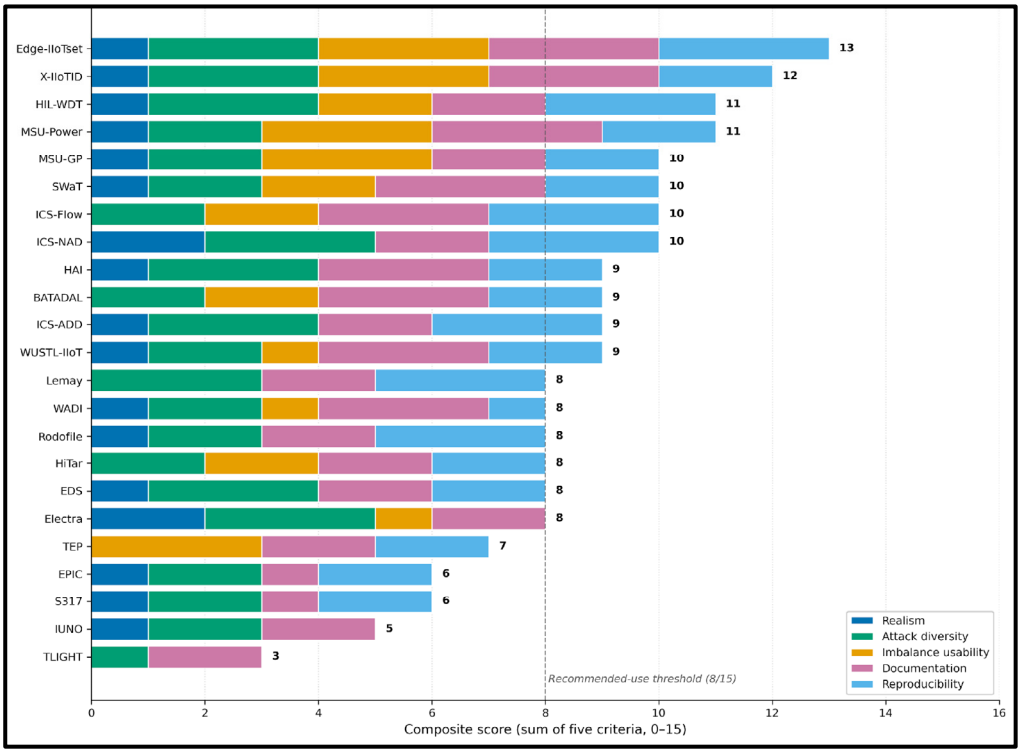


**Figure 8.** Use-case rubric composite scores per dataset.

This avoids overclaiming generalization from narrow, restricted, or weakly documented datasets while still allowing specialized datasets to be used where they provide unique value.

## 7. Discussion and Future Directions

The analysis in Sections 4–6 highlights both the contributions and the structural limitations of the current ICS/IIoT datasets. This section synthesizes those findings into a research-direction roadmap. The discussion is organized around five themes: (i) standardization and reporting, (ii) hybrid and ensemble learning architectures, (iii) federated learning and privacy-preserving collaboration, (iv) cross-domain transfer learning, and (v) explainable AI for operator-facing IDS.

Standardization is the most obvious gap. Across the 23 selected datasets, attack labeling, feature schemas, and protocol coverage are inconsistent, and the documentation-completeness scoring presented in Section 6.4 shows that even datasets with strong adoption omit individual reporting axes. Cross-dataset comparison would benefit from a community-agreed minimal reporting standard. Aligning future dataset releases with established cybersecurity taxonomies such as MITRE ATT&CK for ICS [18] and the IEC 62443 control reference would also improve attack labeling and enable consistent cross-corpus mapping; this alignment is a future-directions item rather than a contribution of the present manuscript.

Hybrid and ensemble learning architectures, in which two or more model families address different signal regimes, are a promising direction for ICS-IDS. Stacked combinations of supervised classifiers with anomaly-based detectors can simultaneously cover the well-labeled attack space and the long-tail unknown space, which is poorly represented in current datasets [16,17].

Federated learning is an important direction for ICS environments where data sharing is constrained by safety, regulation, or competitive concerns. Federated approaches allow multiple plants or operators to jointly train detection models without exchanging raw traffic. The corpus surveyed in this manuscript already contains datasets explicitly designed to support federated evaluation (Edge-IIoTset [39]), which provides a starting point. Open questions include client-drift handling for non-IID OT process distributions, robustness to malicious participants, and the

 doi:10.20944/preprints202606.0145.v1



evaluation of communication-efficient aggregation strategies on realistic ICS feature dimensionalities.

Cross-domain transfer learning addresses a different but related problem: the scarcity of labeled attack traffic in any single ICS environment. Models trained on one dataset rarely transfer cleanly to another because of differences in process dynamics, protocol mix, and adversary tooling. The skew toward Modbus identified in Section 5.1 and the absence of operational-incident captures identified in Section 6.3 reinforce this challenge. Within this corpus, IEC 61850-specific cross-domain transfer learning is particularly underexplored; EPIC [24] is the only IEC 61850-bearing dataset in the selected set, and the broader literature on IEC 61850-to-IEC 61850 or IEC 61850-to-Modbus transfer remains thin. This is an attractive direction for future work because IEC 61850 deployments are operationally important (substation automation, smart-grid protection) but underrepresented in benchmarks.

Explainable AI for IDS is the final research direction worth highlighting. Operator trust and incident-response actions depend on the IDS providing not only a binary alarm, but also an account of the input features, protocol fields, or process variables that drove the alarm. Datasets that release both raw artifacts and feature schemas (Edge-IIoTset [39], X-IIoTID [41], ICS-NAD [45], HIL-WDT [40], ICS-Flow [12]) are best suited for XAI evaluation because they allow attributions to be linked back to interpretable OT signals. Reducing false-alarm rates through XAI-informed feature pruning and operator-feedback loops is closely tied to the imbalance and documentation challenges discussed earlier in this manuscript.

## 8. Conclusions

This study presented a taxonomic review of 23 ICS/OT/IIoT-related datasets used in machine-learning-based intrusion detection research. The datasets were analyzed along seven measurable axes (environment, data type, format, OT protocols, attack representation, feature/class structure, and documentation), and their reported attacks were mapped to MITRE ATT&CK for ICS tactics. Two quantitative instruments were introduced: a 0–7 documentation-completeness checklist (Section 6.4) and a 0–15 decision-aid scoring rubric covering realism, attack diversity, imbalance, documentation, and reproducibility (Section 6.5). Both are fully derived from the public release documentation of each dataset and the analytical tables in this manuscript.

Several structural findings emerged. First, dataset realism is unevenly distributed: only Electra [33] and ICS-NAD [45] are sourced from real operational networks, and even those carry deliberately injected attacks rather than naturally captured incidents. Second, OT protocol coverage is heavily skewed toward Modbus, while operationally important protocols such as OPC UA, IEC 61850, and Profinet remain underrepresented. Third, class imbalance is the dominant statistical feature of ICS datasets, which has direct consequences for the choice of evaluation metric and for the comparability of reported IDS performance. Fourth, multi-stage and APT-style adversary behavior is underrepresented; most datasets capture isolated attack episodes rather than full kill-chain progressions. Fifth, documentation quality is generally strong but not uniformly so, and specific reporting gaps (per-class instance counts, full feature schemas, downloadable artifacts) are persistent across the corpus.

On the basis of these findings, no single dataset achieves the maximum score on all rubric criteria. Edge-IIoTset [39] (86.7%, 13/15) emerges as the strongest general-purpose IIoT release, followed by ICS-NAD [45] (83.3%, 10/12 after Imbalance NR-exclusion): the strongest realism-anchored candidate, and HAI [49,50] (83.3%, 10/12), the strongest process-anomaly-oriented release. X-IIoTID [41] (80.0%, 12/15) rounds out the top general-purpose IIoT releases. For practical benchmarking, the recommended approach is to pair a realism-focused release (ICS-NAD, Electra) with a documentation- and reproducibility-strong release (Edge-IIoTset, X-IIoTID, HIL-WDT, ICS-Flow, ICS-ADD) and to report results against both rather than a single benchmark.

The most pressing open challenges for the ICS dataset community are: (i) generating or releasing operational-incident traffic from production networks, within the legal and safety-disclosure constraints that govern such releases, (ii) broadening OT protocol coverage to include OPC UA, IEC

 doi:10.20944/preprints202606.0145.v1



61850, and Profinet at usable volume, (iii) capturing multi-stage adversary campaigns with kill-chain-aware labels, and (iv) adopting a community-agreed minimal reporting standard that aligns labels with MITRE ATT&CK for ICS and with IEC 62443 reference models. Future research directions identified in Section 7: federated learning across plants, cross-domain transfer learning (especially for IEC 61850), and explainable AI for operator-facing IDS: depend critically on the availability of datasets that satisfy these standardization and realism requirements.

**Author Contributions: Ayman Termanini:** Conceptualization, Data Curation, Formal Analysis, Investigation, Visualization, Writing, Original Draft, Revision. **Hadj Bourdoucen**: Conceptualization, Methodology, Supervision, Validation, Writing, Review, and Editing. **Dawood Al-Abri**: Conceptualization, Methodology, Co-supervision, Validation, Writing, Review, and Editing. **Ahmad Al-Maashri**: Validation, Resources, Writing, Review & Editing.

**Funding:** This research was supported by a PhD scholarship from Sultan Qaboos University (Oman) awarded to the first author. No external funding was received for this study.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data discussed in this study are derived from publicly available and request-access datasets. No new primary data were generated by the authors. The 23 datasets analyzed in this review are listed in Table 2, along with their canonical references and hosting URLs.

**Acknowledgments:** During the preparation of this manuscript, the authors used the Anthropic Claude model for editorial assistance, including only grammar, structuring, and prose polishing. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Conti, M.; Donadel, D.; Turrin, F. A Survey on Industrial Control System Testbeds and Datasets for Security Research. *IEEE Commun. Surv. Tutor.* **2021**, *23*, 2248–2294. https://doi.org/10.1109/COMST.2021.3094360.
2. Falliere, N.; Murchu, L.O.; Chien, E. W32.Stuxnet Dossier, Version 1.4; Technical Report; Symantec Security Response: Cupertino, CA, USA, 2011.
3. Lee, R.M.; Assante, M.J.; Conway, T. Analysis of the Cyber Attack on the Ukrainian Power Grid: Defense Use Case; Technical Report; Electricity Information Sharing and Analysis Center (E-ISAC) and SANS Industrial Control Systems: Washington, DC, USA, 2016. Available online: https://www.nerc.com/pa/CI/ESISAC/Documents/E-ISAC_SANS_Ukraine_DUC_18Mar2016.pdf (accessed on 14 May 2026).
4. Chapelle, O.; Schölkopf, B.; Zien, A. (Eds.) Semi-Supervised Learning; The MIT Press: Cambridge, MA, USA, 2006. https://doi.org/10.7551/mitpress/9780262033589.001.0001.
5. Koay, A.M.Y.; Ko, R.K.L.; Hettema, H.; Radke, K. Machine learning in industrial control system (ICS) security: Current landscape, opportunities and challenges. *J. Intell. Inf. Syst.* **2023**, *60*, 377–405. https://doi.org/10.1007/s10844-022-00753-1.
6. Pinto, A.; Herrera, L.-C.; Donoso, Y.; Gutierrez, J.A. Survey on Intrusion Detection Systems Based on Machine Learning Techniques for the Protection of Critical Infrastructure. *Sensors* **2023**, *23*, 2415. https://doi.org/10.3390/s23052415.
7. García-Teodoro, P.; Díaz-Verdejo, J.; Maciá-Fernández, G.; Vázquez, E. Anomaly-based network intrusion detection: Techniques, systems and challenges. *Comput. Secur.* **2009**, *28*, 18–28. https://doi.org/10.1016/j.cose.2008.08.003.
8. Choi, S.; Yun, J.-H.; Kim, S.-K. A Comparison of ICS Datasets for Security Research Based on Attack Paths. In *Critical Information Infrastructures Security*; Luiijf, E., Žutautaitė, I., Hämmerli, B.M., Eds.; Springer

 doi:10.20944/preprints202606.0145.v1



International Publishing: Cham, Switzerland, 2019; pp. 154–166. https://doi.org/10.1007/978-3-030-05849-4_12.

9. Mubarak, S.; Habaebi, M.H.; Islam, M.R.; Rahman, F.D.A.; Tahir, M. Anomaly Detection in ICS Datasets with Machine Learning Algorithms. *Comput. Syst. Sci. Eng.* **2021**, *37*, 33–46. https://doi.org/10.32604/csse.2021.014384.
10. Martins, N.; Cruz, J.M.; Cruz, T.; Henriques Abreu, P. Adversarial Machine Learning Applied to Intrusion and Malware Scenarios: A Systematic Review. *IEEE Access* **2020**, *8*, 35403–35419. https://doi.org/10.1109/ACCESS.2020.2974752.
11. Mitseva, A.; Thierse, P.; Hoffmann, H.; Er, D.; Panchenko, A. Challenges and Pitfalls in Generating Representative ICS Datasets in Cyber Security Research. In *Computer Security. ESORICS 2022 International Workshops*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2023; Volume 13785, pp. 379–397. https://doi.org/10.1007/978-3-031-25460-4_22.
12. Dehlaghi-Ghadim, A.; Moghadam, M.H.; Balador, A.; Hansson, H. Anomaly Detection Dataset for Industrial Control Systems. *IEEE Access* **2023**, *11*, 107982–107996. https://doi.org/10.1109/ACCESS.2023.3320928.
13. Hu, Y.; Yang, A.; Li, H.; Sun, Y.; Sun, L. A survey of intrusion detection on industrial control systems. *Int. J. Distrib. Sens. Netw.* **2018**, *14*, 1–14. https://doi.org/10.1177/1550147718794615.
14. Anwar, M.; Lundberg, L.; Borg, A. Improving anomaly detection in SCADA network communication with attribute extension. *Energy Inform.* **2022**, *5*, 69. https://doi.org/10.1186/s42162-022-00252-1.
15. Sun, D.; Zhang, L.; Jin, K.; Ling, J.; Zheng, X. An Intrusion Detection Method Based on Hybrid Machine Learning and Neural Network in the Industrial Control Field. Appl. Sci. 2023, 13, 10455. https://doi.org/10.3390/app131810455
16. Yang, Z.; Liu, X.; Li, T.; Wu, D.; Wang, J.; Zhao, Y.; Han, H. A systematic literature review of methods and datasets for anomaly-based network intrusion detection. *Comput. Secur.* **2022**, *116*, 102675. https://doi.org/10.1016/j.cose.2022.102675.
17. Kheddar, H.; Himeur, Y.; Awad, A.I. Deep transfer learning for intrusion detection in industrial control networks: A comprehensive review. *J. Netw. Comput. Appl.* **2023**, *220*, 103760. https://doi.org/10.1016/j.jnca.2023.103760.
18. MITRE Corporation. ATT&CK for ICS. 2024. Available online: https://attack.mitre.org/matrices/ics/ (accessed on 14 May 2026).
19. Mathur, A.P.; Tippenhauer, N.O. SWaT: A water treatment testbed for research and training on the design of industrial control systems. In *Proceedings of the 2016 International Workshop on Cyber-Physical Systems for Smart Water Networks (CySWater)*, Vienna, Austria, 11 April 2016; pp. 31–36. https://doi.org/10.1109/CySWater.2016.7469060.
20. Goh, J.; Adepu, S.; Junejo, K.N.; Mathur, A. A Dataset to Support Research in the Design of Secure Water Treatment Systems. In *Critical Information Infrastructures Security*; Havarneanu, G., Setola, R., Nassopoulos, H., Wolthusen, S., Eds.; Springer International Publishing: Cham, Switzerland, 2017; pp. 88–99. https://doi.org/10.1007/978-3-319-71368-7_8.
21. iTrust Labs. iTrust Labs Dataset Information. Singapore University of Technology and Design, 2025. Available online: https://itrust.sutd.edu.sg/itrust-labs_datasets/ (accessed on 14 May 2026).
22. Ahmed, C.M.; Palleti, V.R.; Mathur, A.P. WADI: A water distribution testbed for research in the design of secure cyber physical systems. In *Proceedings of the 3rd International Workshop on Cyber-Physical Systems for Smart Water Networks*, Pittsburgh, PA, USA, 21 April 2017; pp. 25–28. https://doi.org/10.1145/3055366.3055375.
23. Ahmed, C.M.; Mathur, A.P.; Ochoa, M. NoiSense Print: Detecting Data Integrity Attacks on Sensor Measurements Using Hardware Based Fingerprints. *ACM Trans. Priv. Secur.* **2020**, *23*, 1–24. https://doi.org/10.1145/3410447.
24. Adepu, S.; Kandasamy, N.K.; Mathur, A. EPIC: An Electric Power testbed for Research and Training in Cyber Physical Systems Security. In Computer Security; Katsikas, S.K., Cuppens, F., Cuppens, N., Lambrinoudakis, C., Antón, A., Gritzalis, S., Mylopoulos, J., Kalloniatis, C., Eds.; Springer International

 doi:10.20944/preprints202606.0145.v1



Publishing: Cham, Switzerland, 2018; Volume 11387, pp. 37–52. https://doi.org/10.1007/978-3-030-12786-2_3.

25. Taormina, R.; Galelli, S.; Tippenhauer, N.O.; Salomons, E.; Ostfeld, A.; Eliades, D.G.; Aghashahi, M.; Sundararajan, R.; Pourahmadi, M.; Banks, M.K.; et al. Battle of the Attack Detection Algorithms: Disclosing cyber attacks on water distribution networks. *J. Water Resour. Plan. Manag.* **2018**, *144*, 04018048. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000969.
26. Antonioli, D.; Ghaeini, H.R.; Adepu, S.; Ochoa, M.; Tippenhauer, N.O. Gamifying ICS Security Training and Research: Design, Implementation, and Results of S3. In *Proceedings of the 2017 Workshop on Cyber-Physical Systems Security and PrivaCy*, Dallas, TX, USA, 3 November 2017; pp. 93–102. https://doi.org/10.1145/3140241.3140253.
27. Morris, T.; Gao, W. Industrial Control System Traffic Data Sets for Intrusion Detection Research. In *Critical Infrastructure Protection VIII*; Butts, J., Shenoi, S., Eds.; Springer: Berlin/Heidelberg, Germany, 2014; pp. 65–78. https://doi.org/10.1007/978-3-662-45355-1_5.
28. Turnipseed, I.P. A New SCADA Dataset for Intrusion Detection Research; M.Sc. Thesis, Mississippi State University: Starkville, MS, USA, 2015.
29. Pan, S.; Morris, T.; Adhikari, U. Developing a Hybrid Intrusion Detection System Using Data Mining for Power Systems. *IEEE Trans. Smart Grid* **2015**, *6*, 3104–3113. https://doi.org/10.1109/TSG.2015.2409775.
30. Hink, R.C.B.; Beaver, J.M.; Buckner, M.A.; Morris, T.; Adhikari, U.; Pan, S. Machine learning for power system disturbance and cyber-attack discrimination. In *Proceedings of the 2014 7th International Symposium on Resilient Control Systems (ISRCS)*, Denver, CO, USA, 19–21 August 2014; pp. 1–8. https://doi.org/10.1109/ISRCS.2014.6900095.
31. Adhikari, U.; Morris, T.; Pan, S. Industrial Control System (ICS) Cyber Attack Datasets; Datasets; Mississippi State University: Starkville, MS, USA, 2014.
32. Lemay, A.; Fernandez, J.M. Providing SCADA Network Data Sets for Intrusion Detection Research. In *Proceedings of the 9th Workshop on Cyber Security Experimentation and Test (CSET'16)*, Austin, TX, USA, 8 August 2016. USENIX Association.
33. Perales Gómez, Á.L.; Fernández Maimó, L.; Huertas Celdrán, A.; García Clemente, F.J.; Cadenas Sarmiento, C.; Del Canto Masa, C.J.; Méndez Nistal, R. On the Generation of Anomaly Detection Datasets in Industrial Control Systems. *IEEE Access* **2019**, *7*, 177460–177473. https://doi.org/10.1109/ACCESS.2019.2958284.
34. Rodofile, N.R.; Schmidt, T.; Sherry, S.T.; Djamaludin, C.; Radke, K.; Foo, E. Process Control Cyber-Attacks and Labelled Datasets on S7Comm Critical Infrastructure. In *Information Security and Privacy*; Pieprzyk, J., Suriadi, S., Eds.; Springer International Publishing: Cham, Switzerland, 2017; Volume 10343, pp. 452–459. https://doi.org/10.1007/978-3-319-59870-3_30.
35. Rodofile, N.R. Generating Attacks and Labelling Attack Datasets for Industrial Control Intrusion Detection Systems; Ph.D. Thesis, Queensland University of Technology, Brisbane, Australia, 2018.
36. 36. Zolanvari, M.; Teixeira, M.A.; Gupta, L.; Khan, K.M.; Jain, R. WUSTL-IIoT-2021 Dataset for IIoT Cybersecurity Research. Washington University in St. Louis, USA, 2021. Available online: https://www.cse.wustl.edu/~jain/iiot2/index.html (accessed on 14 May 2026).
37. Teixeira, M.A.; Salman, T.; Zolanvari, M.; Jain, R.; Meskin, N.; Samaka, M. SCADA System Testbed for Cybersecurity Research Using Machine Learning Approach. *Future Internet* **2018**, *10*, 76. https://doi.org/10.3390/fi10080076.
38. Zolanvari, M.; Teixeira, M.A.; Gupta, L.; Khan, K.M.; Jain, R. Machine Learning-Based Network Vulnerability Analysis of Industrial Internet of Things. *IEEE Internet Things J.* **2019**, *6*, 6822–6834. https://doi.org/10.1109/JIOT.2019.2912022.
39. Ferrag, M.A.; Friha, O.; Hamouda, D.; Maglaras, L.; Janicke, H. Edge-IIoTset: A New Comprehensive Realistic Cyber Security Dataset of IoT and IIoT Applications for Centralized and Federated Learning. *IEEE Access* **2022**, *10*, 40281–40306. https://doi.org/10.1109/ACCESS.2022.3165809.
40. Faramondi, L.; Flammini, F.; Guarino, S.; Setola, R. A Hardware-in-the-Loop Water Distribution Testbed Dataset for Cyber-Physical Security Testing. *IEEE Access* **2021**, *9*, 122385–122396. https://doi.org/10.1109/ACCESS.2021.3109465.

 doi:10.20944/preprints202606.0145.v1




41. Al-Hawawreh, M.; Sitnikova, E.; Aboutorab, N. X-IIoTID: A Connectivity- and Device-agnostic Intrusion Dataset for Industrial Internet of Things. *IEEE Internet Things J.* **2022**, *9*, 3962–3977. https://doi.org/10.1109/JIOT.2021.3102056.
42. Rieth, C.A.; Amsel, B.D.; Tran, R.; Cook, M.B. Additional Tennessee Eastman Process Simulation Data for Anomaly Detection Evaluation. Harvard Dataverse, V1, 2017. https://doi.org/10.7910/DVN/6C3JR1.
43. Yau, K., Chow, KP. (2017). Detecting Anomalous Programmable Logic Controller Events Using Machine Learning. In: Peterson, G., Shenoi, S. (eds) Advances in Digital Forensics XIII. DigitalForensics 2017. IFIP Advances in Information and Communication Technology, vol 511. Springer, Cham. https://doi.org/10.1007/978-3-319-67208-3_5
44. Duque Antón, S.; Sinha, S.; Schotten, H.D. Anomaly-based Intrusion Detection in Industrial Data with SVM and Random Forests. In *Proceedings of the International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, Split, Croatia, 19–21 September 2019; pp. 1–6. https://doi.org/10.23919/SOFTCOM.2019.8903672.
45. Zhou, X., Cheng, Z., Wang, C. et al. A dataset collected in real-world industrial control systems for network attack detection. Sci Data 13, 399 (2026). https://doi.org/10.1038/s41597-026-06738-x
46. Gaggero, G.B.; Caviglia, R.; Sanguineti, A.; Marchese, M. Industrial Control System Anomaly Detection Dataset. *Data* **2024**, *9*, 138. https://doi.org/10.1109/ACCESS.2024.3395991.
47. T. Dhaouadi, H. Mrabet, A. Alhomoud, and A. Jemai, "An Intrusion Detection System Based on HiTar-2024 Dataset Generation from LOG Files for Smart Industrial Internet-of-Things Environment," Comput. Mater. Contin., vol. 82, no. 3, pp. 4535–4554, 2025. https://doi.org/10.32604/cmc.2025.060935
48. Xue, Y.; Pan, J.; Geng, Y.; Yang, Z.; Liu, M.; Deng, R. Real-Time Intrusion Detection Based on Decision Fusion in Industrial Control Systems. IEEE Trans. Ind. Cyber-Phys. Syst. 2024, 2, 143–153. DOI 10.1109/TICPS.2024.3406505
49. Shin, H.-K.; Lee, W.; Yun, J.-H.; Kim, H. HAI 1.0: HIL-based Augmented ICS Security Dataset. In *Proceedings of the 13th USENIX Workshop on Cyber Security Experimentation and Test (CSET'20)*, online, 10 August 2020. USENIX Association. https://www.usenix.org/conference/cset20/presentation/shin
50. Shin, H.-K.; Lee, W.; Yun, J.-H.; Min, B.-G. Two ICS Security Datasets and Anomaly Detection Contest on the HIL-based Augmented ICS Testbed. In *Proceedings of the 14th USENIX Workshop on Cyber Security Experimentation and Test (CSET'21)*, online, 9 August 2021. USENIX Association. https://doi.org/10.1145/3474718.3474719
51. Stolfo, S.J.; Fan, W.; Lee, W.; Prodromidis, A.; Chan, P.K. KDD Cup 1999 Data. UCI KDD Archive, 1999. Available online: https://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html (accessed on 14 May 2026).
52. Tavallaee, M.; Bagheri, E.; Lu, W.; Ghorbani, A.A. A detailed analysis of the KDD CUP 99 data set. In *Proceedings of the 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, Ottawa, ON, Canada, 8–10 July 2009; pp. 1–6. https://doi.org/10.1109/CISDA.2009.5356528.
53. Moustafa, N.; Slay, J. UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In *Proceedings of the 2015 Military Communications and Information Systems Conference (MilCIS)*, Canberra, ACT, Australia, 10–12 November 2015; pp. 1–6. https://doi.org/10.1109/MilCIS.2015.7348942.
54. Sharafaldin, I.; Lashkari, A.H.; Ghorbani, A.A. Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization. In *Proceedings of the 4th International Conference on Information Systems Security and Privacy*, Funchal, Madeira, Portugal, 22–24 January 2018; pp. 108–116. https://doi.org/10.5220/0006639801080116.
55. Shiravi, A.; Shiravi, H.; Tavallaee, M.; Ghorbani, A.A. Toward developing a systematic approach to generate benchmark datasets for intrusion detection. *Comput. Secur.* **2012**, *31*, 357–374. https://doi.org/10.1016/j.cose.2011.12.012.
56. Koroniotis, N.; Moustafa, N.; Sitnikova, E.; Turnbull, B. Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics: Bot-IoT dataset. *Future Gener. Comput. Syst.* **2019**, *100*, 779–796. https://doi.org/10.1016/j.future.2019.05.041.
57. iTrust Centre for Research in Cyber Security. iTrust. Singapore University of Technology and Design. Available online: https://itrust.sutd.edu.sg/ (accessed on 14 May 2026).

 doi:10.20944/preprints202606.0145.v1




58. IEC 61850-8-1; Communication Networks and Systems for Power Utility Automation—Part 8-1: Specific Communication Service Mapping (SCSM)—Mappings to MMS (ISO 9506-1 and ISO 9506-2) and to ISO/IEC 8802-3. International Electrotechnical Commission: Geneva, Switzerland, 2011.
59. Pan, S.; Morris, T.; Adhikari, U. A Specification-Based Intrusion Detection Framework for Cyber-Physical Environment in Electric Power System. *Int. J. Netw. Secur.* **2015**, *17*, 174–188.
60. Pan, S.; Morris, T.; Adhikari, U. Classification of Disturbances and Cyber-Attacks in Power Systems Using Heterogeneous Time-Synchronized Data. *IEEE Trans. Ind. Inform.* **2015**, *11*, 650–662. https://doi.org/10.1109/TII.2015.2420951.
61. A. Pallakonda, S. Sanjay Kumar, R. David Amar Raj, R. Muni Reddy Yanamala and K. K. Prakasha, "Secure and Resilient Cyberattack Detection in ICS Networks: Hybrid Encryption, Protocol Hardening, and Threat Hunting on ELECTRA Modbus Traffic," in IEEE Access, vol. 13, pp. 177227-177245, 2025, https://doi.org/10.1109/ACCESS.2025.3619487.
62. Thierry Berger, Dominique Deneux, Thérèse Bonte, Etienne Cocquebert, Damien Trentesaux. Arezzo-flexible manufacturing system: A generic flexible manufacturing system shop floor emulator approach for high-level control virtual commissioning. Concurrent Engineering: Research and Applications, 2015, 23 (4), pp.333-342. ⟨10.1177/1063293X15591609⟩. ⟨hal-03430062⟩
63. Booij, T.M.; Chiscop, I.; Meeuwissen, E.; Moustafa, N.; den Hartog, F.T.H. ToN_IoT: The Role of Heterogeneity and the Need for Standardization of Features and Attack Types in IoT Network Intrusion Data Sets. *IEEE Internet Things J.* **2022**, *9*, 485–496. https://doi.org/10.1109/JIOT.2021.3085194.
64. Saito, T.; Rehmsmeier, M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE* **2015**, *10*, e0118432. https://doi.org/10.1371/journal.pone.0118432.
65. Akosa, J. Predictive Accuracy: A Misleading Performance Measure for Highly Imbalanced Data. In *Proceedings of the SAS Global Forum 2017*, Orlando, FL, USA, 2–5 April 2017; Paper 942.
66. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. https://doi.org/10.1613/jair.953.