Sherlock: A Dataset for Process-aware Intrusion Detection Research on Power Grid Networks

Dataset Paper

Eric Wagner*
eric.wagner@fkie.fraunhofer.de
Fraunhofer FKIE
RWTH Aachen University
Aachen, Germany

Lennart Bader*
lennart.bader@fkie.fraunhofer.de
Fraunhofer FKIE
RWTH Aachen University
Aachen, Germany

Konrad Wolsing
konrad.wolsing@fkie.fraunhofer.de
Fraunhofer FKIE
RWTH Aachen University
Aachen, Germany

Martin Serror
martin.serror@fkie.fraunhofer.de
Fraunhofer FKIE
Aachen, Germany

Abstract

Physically distributed components and legacy protocols make the protection of power grids against increasing cyberattack threats challenging. Infamously, the 2015 and 2016 blackouts in Ukraine were caused by cyberattacks, and the German Federal Office for Information Security (BSI) recorded over 200 cyber incidents against the German energy sector between 2023 and 2024. Intrusion detection promises to quickly detect such attacks and mitigate the worst consequences. However, public datasets of realistic scenarios are vital to evaluate these systems. This paper introduces Sherlock, a dataset generated with the co-simulator Wattson. In total, Sherlock covers three scenarios with various attacks manipulating the process state by injecting malicious commands or manipulating measurement values. We additionally test five recently-published intrusion detection systems on Sherlock, highlighting specific challenges for intrusion detection in power grids. Dataset and documentation are available at https://sherlock.wattson.it/.

CCS Concepts

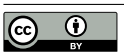
• **Security and privacy → Intrusion detection systems**.

Keywords

dataset, critical infrastructure, power grid, IEC 60870-5-104

ACM Reference Format:
Eric Wagner, Lennart Bader, Konrad Wolsing, and Martin Serror. 2025. Sherlock: A Dataset for Process-aware Intrusion Detection Research on Power Grid Networks: Dataset Paper. In *Proceedings of the Fifteenth ACM Conference on Data and Application Security and Privacy (CODASPY '25), June 4–6, 2025, Pittsburgh, PA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3714393.3726006

*Both authors contributed equally to this work



CODASPY '25, June 4–6, 2025, Pittsburgh, PA, USA

ACM ISBN 979-8-4007-1476-4/2025/06
https://doi.org/10.1145/3714393.3726006

1 Introduction

Cyberattacks against critical infrastructures, such as power grids, are on the rise [3, 22]. These attacks typically exploit vulnerabilities in the underlying Industrial Control System (ICS) networks, which are known to rely on insecure legacy communication protocols [15, 23]. To make matters worse, such legacy protocols are difficult to replace due to long lifecycles of industrial hardware, stringent availability requirements, and limited update capabilities. Under these circumstances, feasible preventive security measures, such as network segmentation and firewalls, do not suffice for protection [7]. Effective intrusion detection promises to identify cyberattacks in their early stages and thus enable timely countermeasures that prevent severe damage. Industrial Intrusion Detection Systems (IIDSs) are thus widely recognized as a retrofittable, non-disruptive, and promising security solution, serving as the last line of defense for critical infrastructure [20]. However, such IIDSs must deal with the unique characteristics of power grid networks.

Traditional network-based IIDSs detect suspicious activities by scanning traffic for known attack patterns [19]. However, ICS networks are often exposed to unknown attacks due to their distinct characteristics and the use of a wide range of different protocols. Meanwhile, the predictable nature of control and sensor traffic, closely tied to underlying physical processes, creates opportunities for process-aware intrusion detection [7]. Here, the core idea is to examine a system's physical state using data transmitted over the network to detect anomalies. However, this approach necessitates domain-specific training data as a prerequisite for accurately modeling the expected system behavior.

This relatively new research area is experiencing significant growth, with at least 130 new process-aware intrusion detection mechanisms proposed in 2021 alone [20]. However, progress is hindered by a lack of high-quality datasets: fewer than half of publications use public datasets, and only 16.4 % utilize more than one. Within critical infrastructures, the energy sector faces a notable dataset gap, as most existing datasets primarily represent small-scale networks or individual components [2, 8, 32]. Consequently, no comprehensive dataset is currently available for evaluating process-aware intrusion detection in power grid networks.

At the same time, the successful attacks on the Ukrainian power grid in 2015 and 2016, as well as the attempted attack in 2022, underscore the devastating impact of such incidents and highlight the interest of powerful adversaries exploiting these weaknesses [28]. Meanwhile, the German Federal Office for Information Security (BSI) reported over 200 cyber incidents against the Germany energy sector between 2023 and 2024 [13, 14].

With this paper, we introduce Sherlock, a comprehensive dataset specifically designed for process-aware, as well as network-based, intrusion detection. Sherlock was recorded using the co-simulator Wattson [6], which can simulate power grids while concurrently emulating their corresponding Information and Communication Technologies (ICT) networks. Wattson has been validated against a physical power grid, ensuring its accuracy, and supports both emulated network components and hardware-in-the-loop integration. Furthermore, it provides a safe research environment for replicating both routine operations and real-world cyberattacks. These features make it an ideal tool for collecting representative, diverse, and reproducible datasets for intrusion detection research.

We passively capture network traffic from critical vantage points across three scenarios (denoted as `01-Basic`, `02-Semiurban`, and `03-Rural`) during simulations spanning 35 days. For two networks, we provide labeled attack-free and attack datasets, while for the third, only attack data is available. This approach promotes research into the generalizability and transferability of detection methods, a critical goal given the expense and complexity of retraining intrusion detection systems with clean training data for each network.

Captured traffic is post-processed using the IPAL toolset [30] to generate time-series data representing the system's physical state. This abstraction layer enhances dataset accessibility and decouples it from region-specific communication protocols.

The IPAL representation enables us to evaluate the detection performance of five IIDSs [4, 12, 18, 29, 31], which claim domain generalizability. This evaluation reveals six challenges in power grid networks that current process-aware intrusion detection mechanisms struggle to address, ranging from handling thousands of data points to accommodating long-term process variations.

2 Related Work

Conti *et al.* [9] identifies 23 public datasets for cybersecurity research in ICS networks. However, most lack process data and some rely on IIDSs explicitly learning from captured attack samples, which limits their ability to detect novel attacks or variations [19]. Among these, only five datasets of decent quality incorporate process data, yet none represent power grid scenarios at realistic scales.

The SWaT dataset [16] captures process data from a scaled-down water treatment plant, featuring 36 physical attacks executed as machine-in-the-middle (MitM) attacks over 11 days. Its single-execution attacks and focus on physical modifications limit its generalizability. The WADI dataset [1] focuses on water distribution, incorporating 14 attacks over 16 days with more physical measurement points but fewer process stages than SWaT. Similarly, the BATADAL dataset [26], based on a simulated water distribution network, samples data at hourly intervals, spanning a year with 14 attacks but offering coarser granularity. The HAI dataset [24] covers different stages of power generation processes. It features 50 diverse attacks of varying complexity. Finally, the EPIC dataset [2] focuses on a small power grid scenario with data collected from Intelligent Electronic Devices (IEDs) monitoring electrical parameters. While it includes both network and process data, its attack scenarios are limited to malicious reconfigurations of devices.

Current datasets in the power grid domain [2, 8, 32] focus on network-based attacks and do either not include attacks against the physical process or do not even accurately reflect the physical process. Thus, large-scale power grid scenarios are not covered by existing datasets. Moreover, there remains a general lack of datasets that combine both network and process data, feature complex, multi-stage attacks, and are grounded in realistic, scalable scenarios. Sherlock should fill this gap.

3 The Sherlock Dataset

The Sherlock dataset aims to address gaps left by existing datasets, providing a valuable resource for assessing intrusion detection methods in power grids. Beyond this goal, Sherlock is designed to support broader research into power grid cybersecurity and the practical deployment of IIDSs in realistic network environments. The following sections detail our testbed setup and provide an in-depth overview of the Sherlock dataset. Additional information is available on the Sherlock website at https://sherlock.wattson.it/.

3.1 Testbed Setup

The backbone of our testbed is the Wattson simulator [6]. Wattson is an open-source power grid co-simulator, *i.e.*, it emulates realistic network traffic among power grid devices while simulating the power grid. Wattson uses PowerOwl [5] to model the power grid, which offers steady-state power flow calculations based on pandapower [27] and emulates its communication network supporting switches, routers, and hosts with lightweight namespaces, based on Docker containers, and with virtual machines in Linux. Wattson uses `tc` for traffic control to configure delays, jitter, bandwidth, and packet-loss for each individual link. The communication between the control center and substations is performed using the IEC 60870-5-104 (IEC 104) protocol.

For our scenarios, we focus on future-oriented settings with a significant fraction of substations being digitized, *i.e.*, they digitally transmit measurements and—if applicable—support the remote execution of control commands. For each scenario, we include load and optional generation profiles to control the behavior of these assets during the evaluation. Wattson performs a real-time co-simulation with a 14x accelerated power profile, *i.e.*, evaluating 12 h of network traffic reflects the power generation and usage patterns over an entire week. To reduce complexity, we abstract from protective relays as intrusion detection should alert before they trigger. The Sherlock dataset is composed of network captures from mirror ports at switches that were identified as key vantage points, enriched with logs from individual hosts and services, additional context information, process ground-truth information, control center events, and documentation. Our online documentation presents the details of the different scenarios, vantage points, and grid values.

For Sherlock, we extract all relevant information passively, aligning with a non-invasive deployment strategy well-suited for real-world power grid networks. The alternative of active polling

consumes substantial bandwidth and exposes the IIDS to manipulated data. Given that the centralized control center typically serves as a data sink, it provides an ideal vantage point, offering a comprehensive overview of the network. Sherlock also provides data from alternative vantage points for further insights when desired.

In total, we simulate 35 days of power grid behavior, split into training and test sets of three different scenarios for Sherlock.

3.2 Scenarios

Sherlock contains three different scenarios of different size and complexity, each consisting of a power grid topology, an ICT network topology, and configurations regarding the coupling of both domains, *i.e.*, responsibilities of remote terminal units (RTUs) along with communication protocol information. For each power grid topology, we use PowerOwl [5] to automatically detect facilities and derive a realistic ICT network, resulting in a simulation scenario compatible with Wattson. The power grid itself comprises multiple stations, *i.e.*, several *distribution substations (DSSs)* and one *transforming substation (TSS)*. Each station contains one or multiple buses that are connected by lines and transformers and further link with assets such as storages (batteries), generators, and loads.

The ICT network comprises multiple subnets, with each scenario including at least two OT subnets for RTUs: one for the TSS and one or more for the DSSs. Additionally, there is a Control Center subnet hosting the master terminal unit (MTU) and multiple office subnets for servers and workstations. The subnets are interconnected via routers using the OSPF protocol and further include switches to link individual facilities and multiple hosts within these facilities. The topologies are part of Sherlock's documentation and are explained further on the dataset website. Beyond the topologies, each scenario specifies the communication behavior of key assets, such as the MTU and RTUs, as well as the operational behavior of power grid components like loads, storage systems, and generators.

For the IEC 104-based communication between the control center and the RTUs, we define an interval of 10 s for periodic measurement transmissions such that each RTU transmits measured voltages, currents, and power values unsolicitedly. Discrete values, such as binary states of circuit breakers, are only transmitted when explicitly requested and every time the value changes.

The power grid behavior is determined by the pre-defined behavior of loads, generators, and storages, further influenced by control operations executed by the grid operator. These operations involve sending control commands to RTUs. Depending on the individual scenario, the respective power profiles target all assets and vary across asset types. For instance, a load representing a household exhibits different behavior compared to that of a supermarket. Whenever possible, we utilize profiles provided by the power grid scenarios; otherwise, we rely on a generic load curve as a fallback.

Next, we briefly introduce the three different scenarios featured in the Sherlock dataset.

3.2.1 `01-Basic`*: The Cigre MV Reference Grid.* This scenario comprises 12 medium voltage (MV) DSSs connected to a high voltage to medium voltage (HV/MV) TSS. It includes 13 generators, 2 storages, 18 loads, and 2 HV/MV transformers. With 32 RTUs distributed across two operational technology (OT) subnets, each substation supports remote monitoring and control via a single MTU. We apply a generic load profile to all 18 loads, while storages and generators operate with static power infeed or consumption. The scenario adopts the *Cigre MV* power grid topology provided by pandapower [27], based on the CIGRE Task Force C6.04.02 paper [25].

3.2.2 `02-Semiurban`*: Simbench MV Semi-urban.* Complementing the `01-Basic` scenario, the Sherlock dataset incorporates two larger, more realistic scenarios derived from Simbench [21]. The *Simbench MV Semi-urban* models an HV/MV distribution grid supplying a semi-urban city area. It features a central TSS with two transformers connecting to two double-busbars. Its 118 DSSs follow a multi-ring topology and, like the Cigre MV scenario, connect a future-oriented number of renewable generators. The OT network includes 9 of 17 subnets, with all 72 RTUs linked to a single control center. Unlike the Cigre MV scenario, this setup applies scenario-specific load and generation profiles to all relevant assets. The power grid topology is based on the simbench key `1-MV-semiurb-2-sw`.

3.2.3 `03-Rural`*: Simbench MV Rural.* Transferability of IIDSs to similar yet different scenarios is a crucial research objective. To support this, Sherlock includes a third scenario that shares similarities with the `02-Semiurban` scenario's topology but differs in size and asset count. The power grid topology is based on the simbench key `1-MV-rural-2-sw` and compromises 95 DSSs and a single TSS, representing a rural distribution grid. The combined nominal power of all loads exceeds 30 MVA, while all generators provide 47 MVA. With 12 OT subnets, 16 subnets in total, and 60 RTUs, the ICT network is smaller compared to `02-Semiurban`. Providing no training data, the Sherlock dataset encourages researchers to enhance transferability by training their IIDSs with the `02-Semiurban` scenario and testing them against the `03-Rural` scenario.

3.3 Commands and Measurements

In all scenarios, RTUs monitor and control power grid assets, including buses, lines, transformers, circuit breakers, loads, generators, and batteries. Most floating-point measurements that are expected to change gradually, such as voltages and currents, are periodically transmitted to the control center using the IEC 104 protocol (`Type ID=13`, `Cause of Transmission (CoT)=1`). Other data points, such as booleans (`Type ID=1`), are configured to be transmitted spontaneously (`CoT=3`), *i.e.*, when they are changed. This includes tap positions on transformers, circuit breaker states, and the connectivity of loads, generators, and storages.

For control commands, the MTU issues target values to the RTUs with desired states, *e.g.*, for circuit breaker positions or power infeed set points. These commands are verified by the responsible RTU, executed, and an acknowledgment is sent back to the MTU. In case of invalid or unrealizable commands, a negative confirmation is sent. During normal operation, the control center can reduce and increase the power infeed of generators or change the topology—either to reduce the load on transformers and lines or to allow maintenance work in distribution substations or on power lines. For the dataset, benign commands issued by the power grid operator will not impede the power supply for customers.

Scenario	Type	Vantage Points	Duration	Attacks	Benign Events
01-Basic	train	4	12 h	-	7
	test	4	12 h	17	10
02-Semiurban	train	6	12 h	-	9
	test	6	12 h	29	10
03-Rural	test	8	12 h	28	8

Table 1: Metadata of Sherlock's scenarios.

3.4 Attacks

There exist many paths for an attacker to get to a state where they have full control over one or multiple devices in a network. These approaches range from supply chain attacks over physical intrusions, such as breaching a substation to connect unauthorized devices, to the classic exploitation of vulnerabilities in existing devices. Some of these steps are not detectable by IIDSs (*e.g.*, supply chain attacks) and others are device-specific (*e.g.*, exploitation of devices). Therefore, we focus on the final phase of attacks that actively impact the state of the power grid, either by injecting control commands, suppressing messaging, or manipulating measurements to provoke damaging reactions or hide critical conditions.

Table 1 provides an overview of the three scenarios and their respective metadata. The attacks are executed consecutively within a single run, with clearly defined start and end points. Sufficient time is allocated between attacks to allow the system to recover to a stable state. Additionally, we include multiple extended attack-free periods in the test set to help minimize false alarms.

Table 2 presents simple examples from the `01-Basic` scenario of the four attack types covered by Sherlock. These attack types include denial-of-service (DoS), control command insertion inspired by the Industroyer attack that caused widespread blackouts in the Ukrainian power grid [10], and advanced false data injection attacks that distort the grid operator's view of the grid state. The *Control & Freeze* attack further manipulates the grid in real-time. Detailed descriptions of all attacks, as well as maintenance events that may be mistakenly identified as attacks, are provided in the documentation.

3.5 Recommended Evaluation Metrics

We recommend that researchers utilizing Sherlock for evaluating their IIDS primarily report three metrics: Detected Attacks, False Alarms, and Average Time to Detection (TTD). To calculate these metrics, we define an alarm as a continuous signalization of an attack by an IIDS. These metrics are then defined as follows:

Detected Attacks. The absolute number of attacks during which an alarm starts. Alarms starting before the attack are considered false alarms. Alarms may start during a recovery phase while the system returns to normal operation. Such alarms should be ignored and neither count as detected attack nor as false alarm.

False Alarms. The number of alarms that start during normal operations without an attack. Maintenance events should also be considered normal operation. One may indicate how many of the false alarms are triggered by such events.

Average TTD. The average time in seconds from the start of an attack until the first alarm starts.

Incorporating additional time-based metrics, such as eTa [17], and performance indicators, such as training and classification time, provide deeper insights into system capabilities. Note that point-based metrics, such as F1-scores, are generally unsuitable for assessing time-aware IDSs [17] since such metrics insufficiently penalize false alarms and fail to account for scenarios where a brief alarm at the start of an attack yields artificially high scores.

Attack Type	Start End	Description
DoS	04:11:23 04:13:30	ARP spoofing attack against RTUs 127 and 128, interrupting the MTU connection.
Industroyer	04:55:28 04:58:34	A secondary IEC 104 client connects to RTU 123 from compromized RTU 121 and issues control commands every 3 s to disconnect circuit breakers 15 and 16, inducing a blackout at DSS 8 (Bus 5).
Drift Off	07:56:27 08:06:26	As MitM between the MTU and RTU 118, the attacker manipulates the voltage measurements regarding Bus 2 at DSS 5 to gradually increase to 1.37 (27 kV).
Control & Freeze	10:04:27 10:13:45	As MitM, the attacker learns measurement trends from multiple RTUs and continues to manipulate future measurements to match this trend after injecting control commands that gradually reduce the power infeed of Generator 7, masking the command's local effects.

Table 2: Examples of the different attack types on `01-Basic`.

3.6 Data Format and Extraction

The Sherlock dataset primarily encompasses packet capture in three scenarios from different vantage points. Additionally, Sherlock provides device logs, context information about events (maintenance and attacks), data point mappings to human-readable identifiers, and ground truth information about the power grid state. To enhance accessibility for researchers, we additionally convert the dataset into the IPAL [30] format, an abstract representation of network data specifically designed for intrusion detection research.

Primarily, we focus on passively extracting the current state of the power grid from network traffic. Therefore, we log the initial system state and parse each intercepted IEC 104 packet to update this system state while mapping abstract Information Object Addresses (IOAs) to human-readable identifiers. This observed state, recorded from a single vantage point, is logged every second. As a result, detection performance on the Sherlock dataset reflects what would be achieved in a passive, real-world deployment.

4 Challenges of IDSs in Power Grids

Providing Sherlock in the IPAL format makes it more accessible to the research community. Additionally, it enables us to benchmark a range of existing IIDSs adapted to IPAL that promise domain-independent industrial intrusion detection. We tested five IIDSs on the Sherlock dataset and present the results in Sec. 4.1:

PASAD [4] interprets a process value as a vector space and assesses the drift compared to the behavior observed during training.

Invariant [12] automatically learns invariants of a process that should hold at all times and alerts if any invariant is violated.

Seq2SeqNN [18] is a neural network trained to predict the next process state that alerts significant deviations from the prediction.

SIMPLE [29] assesses process values' plausibility based on extrema, changes, and distributions observed during training.

GeCo [31] learns a state-space model of the entire process and alerts upon deviations from predicted behavior.

Based on these results, we then identify the specific challenges of intrusion detection in the power grid domain in Sec. 4.2.

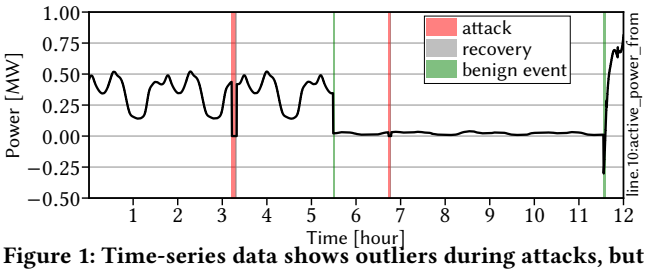


Figure 1: Time-series data shows outliers during attacks, but benign switching event also drastically change the behavior over prolonged times.

4.1 Evaluation Results

In total, we trained five IIDSs for the relatively small `01-Basic` scenario, which features few stations and data points compared to more realistic networks such as the `02-Semiurban` and `03-Rural` scenarios. Table 3 presents our evaluation results. Of these IIDSs, PASAD and Invariant did not produce useful results. Instead, they generated one continuous alert for most of the test data.

The other three IIDSs also only achieve mediocre attack detection performance. A key challenge lies in the dynamic nature of power grids, which exhibit constantly shifting stable configurations driven by factors such as maintenance activities, energy generation and consumption patterns, and the state of connected superordinate grids. Figure 1 clearly illustrates this phenomenon. Two attacks lead to clear outliers in the measured power. However, the two highlighted switching events drastically change the power lines' characteristics, such that trivial intrusion detection mechanisms have a hard time identifying attacks.

To still investigate the potential of process-aware intrusion detection in power grids, we filtered data points that substantially changed between the test and training phases. While this filtering approach facilitates assessing process-aware IDS performance under idealized conditions, it is not feasible in real-world deployments where all measurements may eventually be affected by such changes in the power grid. Nevertheless, this approach provides insight into how process-aware IDS could perform if supplemented with contextual information, such as schedules for planned maintenance activities.

After filtering, Seq2SeqNN, SIMPLE, and GeCo show strong detection performance. An effective IIDS should produce near-zero false alarms, detect most attacks, and do so with minimal delay. While process-aware intrusion detection shows promising potential, the power grid domain still faces several unresolved challenges.

4.2 Challenges in Power Grids

Our evaluation shows the potential of process-aware intrusion detection for power grids. The achieved detection performance will likely improve with further research into detection methodologies facilitated by Sherlock. However, these results were achieved only after filtering sensor values impacted by maintenance activities or switching operations. Our initial benchmarking highlights six challenges in implementing process-aware intrusion detection in power grids, which also exist in other domains to varying degrees.

Some IIDSs faced difficulties during training even on the small-scale `01-Basic` reference grid, despite its relatively limited number

IIDS	Detected Attacks	False Alarms*	Average TTD
PASAD [4]	0/18	1(0)	–
Invariant [12]	0/18	1(0)	–
Seq2SeqNN [18]	7/18	10(0)	149.63s
SIMPLE [29]	16/18	33(33)	81.43s
GeCo [31]	15/18	94(30)	75.89s

* (x) shows how many of these alarms are cause by benign events

Table 3: Detection performance of select IIDS in the *Cigre MV* scenario after heavily filtering measurements that are affected by switching operations and maintenance.

of substations. Furthermore, each substation in Sherlock transmits a minimal amount of data as we employ a single-phase power grid model and, due to the steady-state simulation, abstract from features such as the power grid's utility frequency. IIDSs designed for power grids may reach scalability limitations in practical deployments, a shared problem with other industrial domains [11].

Challenge 1 – Scalability
Intrusion detection must be capable of handling the frequent transmission of hundreds to thousands of data points that are generated by power grid operations.

Beyond the sheer volume of training data, it is impossible to observe all possible states during the training. IIDSs typically aim to learn the cyclic and repetitive behavior of cyber-physical processes in order to detect anomalies. Ideally, the training phase would span multiple cycles to capture this behavior. However, in power grids, this cyclic behavior is only partially present due to factors such as daily weather changes, seasonal variations, maintenance operations, multiple stable configurations, and the ongoing integration of new components (*e.g.*, wind turbines). As a result, some perfectly valid grid configurations may never be observed during training.

Challenge 2 – Training Data Limitation
Even with attack-free training data collected over an extended period, some entirely valid grid configurations would likely remain unobserved.

The operation of a power grid introduces regular benign anomalies, which pose an additional challenge for intrusion detection. Maintenance operations, for example, often require switching off specific power lines and redirecting power flows. Additionally, changes in power demand and generation, including those in subordinate power grids, as well as equipment failures, can necessitate adjustments to the power grid configuration to prevent overloading particular lines. Ideally, these benign changes should not trigger alarms in an IIDS, or at least not result in prolonged false alarms.

Challenge 3 – Benign Anomalous Behavior
Benign anomalous behavior should be anticipated and not trigger (prolonged) false alarms, ensuring dependable surveillance.

An additional challenge for intrusion detection in power grids is the need to consider multiple vantage points. While the control center acts as the primary sink for most measurements and the origin for control commands, a network capture taken just in front of it does not provide a complete picture. Some communications

may be missed, network-based intrusion detection could be compromised, and a strategically positioned attacker could manipulate data at individual vantage points. To address these issues, intrusion detection systems should incorporate multiple vantage points while minimizing communication overhead between them.

Challenge 4 – Vantage Points
To ensure reliable intrusion detection, multiple vantage points should be considered while minimizing the resulting communication overhead.

Multiple vantage points can also help in localizing and understanding the origin of an anomaly. In addition to confirming that these effects are due to a cyberattack, IIDSs should ideally assist in localizing the attack's origin. While actionability is, in general, desirable for IIDSs [11], power grids span vast geographical areas, making it time-consuming to travel between substations and investigate potential signs of an attack (*e.g.*, compromised devices).

Challenge 5 – Actionability
IIDSs should not only detect attacks but also aid in understanding and pinpointing their origin, facilitating a quick resolution.

Finally, different intrusion detection mechanisms are required to detect attacks as reliably as possible [20]. Process-aware intrusion detection can quickly identify anomalous behavior, even in the absence of changes in network traffic, such as when a supply chain attack compromises an RTU that confirms but does not execute commands. In contrast, network-based mechanisms may detect the attachment of new devices to the network (*e.g.*, by observing unexpected ARP requests), thereby identifying attacks before they manipulate the process. Meanwhile, host-based mechanisms may detect manipulated firmware in advance.

Challenge 6 – Multi-layer Intrusion Detection
A holistic surveillance of power grids is only achievable by combining process-aware, network-based, and host-based intrusion detection. These approaches should ideally complement and support each other to enhance detection performance.

5 Conclusion

We present the Sherlock dataset to advance research on process-aware intrusion detection in power grids. The dataset encompasses three scenarios of realistically sized power grids, passively capturing network traffic at multiple vantage points during normal operations and periods influenced by cyberattacks. We extract process state information using human-readable data point identifiers in the IPAL format. This format also facilitates testing and validation, as demonstrated by our evaluation of five general-purpose industrial intrusion detection methods on the dataset. Our initial findings identify six key challenges for intrusion detection research for power grids, such as overlapping cyclic behaviors based on time of day, season, and weather, which complicate the identification of benign characteristics. We envision that the Sherlock dataset will assist the research community in tackling these challenges in the future and thus contribute to the security of critical infrastructures.

Acknowledgments

This paper was supported by the EDA Cyber R&T project "Cyber Electromagnetic Resilience Evaluation on Replicated Environment (CERERE)", funded by Italy and Germany.

References

[1] Chuadhry Mujeeb Ahmed et al. 2017. WADI: A Water Distribution Testbed for Research in the Design of Secure Cyber Physical Systems. In *WS on Cyber-Physical Systems for Smart Water Networks*.
[2] Chuadhry Mujeeb Ahmed and Nandha Kumar Kandasamy. 2021. A Comprehensive Dataset from a Smart Grid Testbed for Machine Learning Based CPS Security Research. In *CPS4CIP*. Cham.
[3] Tejasvi Alladi et al. 2020. Industrial control systems: Cyberattack trends and countermeasures. *Comput. Commun.* 155 (2020).
[4] Wissam Aoudi et al. 2018. Truth Will Out: Departure-Based Process-Level Detection of Stealthy Attacks on Control Systems. In *ACM CCS*.
[5] Lennart Bader. 2024. PowerOwl: A Deterministic Heuristical Approach for Power Grid Modeling. https://powerowl.org, last accessed: December 13, 2024.
[6] Lennart Bader et al. 2023. Comprehensively Analyzing the Impact of Cyberattacks on Power Grids. In *IEEE EuroS&P*.
[7] Amaury Beaudet et al. 2020. Process-Aware Model-based Intrusion Detection System on Filtering Approach: Further Investigations. In *IEEE ICIT*.
[8] Partha P Biswas et al. 2019. A Synthesized Dataset for Cybersecurity Study of IEC 61850 based Substation. In *IEEE SmartGridComm*.
[9] Mauro Conti et al. 2021. A Survey on Industrial Control System Testbeds and Datasets for Security Research. *IEEE Commun. Surv. Tutor.* 23, 4 (2021).
[10] ESET Research. 2022. *Industroyer2: Industroyer reloaded.* https://www.welivesecurity.com/2022/04/12/industroyer2-industroyer-reloaded/
[11] Sandro Etalle. 2017. From Intrusion Detection to Software Design. In *ESORICS*.
[12] Cheng Feng et al. 2019. A Systematic Framework to Generate Invariants for Anomaly Detection in Industrial Control Systems.. In *NDSS*.
[13] Federal Office for Information Security (BSI). 2023. *The State of IT Security in Germany in 2023.* https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/Publications/Securitysituation/IT-Security-Situation-in-Germany-2023.html.
[14] Federal Office for Information Security (BSI). 2024. *The State of IT Security in Germany in 2024.* https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/Publications/Securitysituation/IT-Security-Situation-in-Germany-2024.html.
[15] Brendan Galloway and Gerhard P. Hancke. 2013. Introduction to Industrial Control Networks. *IEEE Commun. Surv. Tutor.* 15, 2 (2013).
[16] Jonathan Goh et al. 2016. A Dataset to Support Research in the Design of Secure Water Treatment Systems. In *CRITIS*.
[17] Won-Seok Hwang et al. 2022. "Do You Know Existing Accuracy Metrics Overrate Time-Series Anomaly Detections?". In *ACM/SIGAPP SAC*.
[18] Jonguk Kim et al. 2020. Anomaly Detection for Industrial Control Systems Using Sequence-to-Sequence Neural Networks. In *ESORICS Workshops*. Springer.
[19] Dominik Kus et al. 2022. A False Sense of Security? Revisiting the State of Machine Learning-Based Industrial Intrusion Detection. In *ACM CPSS*.
[20] Olav Lamberts et al. 2023. SoK: Evaluations in Industrial Intrusion Detection Research. *Journal of Systems Research* (2023).
[21] Steffen Meinecke et al. 2020. SimBench—A Benchmark Dataset of Electric Power Systems to Compare Innovative Solutions Based on Power Flow Analysis. *Energies* 13, 12 (2020).
[22] Vetrivel S. Rajkumar et al. 2023. Cyber Attacks on Power Grids: Causes and Propagation of Cascading Failures. *IEEE Access* 11 (2023).
[23] Martin Serror et al. 2021. Challenges and Opportunities in Securing the Industrial Internet of Things. *IEEE Transactions on Industrial Informatics* 17, 5 (2021).
[24] Hyeok-Ki Shin et al. 2020. HAI 1.0: HIL-based Augmented ICS Security Dataset. In *CSET*.
[25] Kai Strunz et al. 2006. Developing Benchmark Models for Integrating Distributed Energy Resources. In *Power Engineering Society General Meeting*.
[26] Riccardo Taormina et al. 2018. Battle of the Attack Detection Algorithms: Disclosing Cyber Attacks on Water Distribution Networks. *Journal of Water Resources Planning and Management* 144, 8 (2018).
[27] Leon Thurner et al. 2018. Pandapower—An Open-Source Python Tool for Convenient Modeling, Analysis, and Optimization of Electric Power Systems. *IEEE Transactions on Power Systems* 33, 6 (2018).
[28] David E. Whitehead et al. 2017. Ukraine Cyber-induced Power Outage: Analysis and Practical Mitigation Strategies. In *CPRE*.
[29] Konrad Wolsing et al. 2022. Can Industrial Intrusion Detection Be SIMPLE?. In *ESORICS*.
[30] Konrad Wolsing et al. 2022. IPAL: Breaking up Silos of Protocol-dependent and Domain-specific Industrial Intrusion Detection Systems. In *RAID*.
[31] Konrad Wolsing et al. 2025. GeCos Replacing Experts: Generalizable and Comprehensible Industrial Intrusion Detection. In *USENIX Security*.
[32] Sven Zemanek et al. 2022. PowerDuck: A GOOSE Data Set of Cyberattacks in Substations. In *CSET*.