Workshop Proceedings of the 19th International Conference on Intelligent Environments (IE2023)
G. Bekaroo et al. (Eds.)

doi:10.3233/AISE230004

A Survey of Smart Grid Intrusion Detection Datasets

Mohammed M. ALANI [a]
[a] *Cybersecurity Research Lab, Toronto Metropolitan University, Toronto, Canada*

Thar BAKER [b]
[b] *School of Architecture, Technology and Engineering, The University of Brighton, Brighton, UK*
[1] ORCiD ID: Mohammed M. Alani https://orcid.org/0000-0002-4324-1774, Thar Baker https://orcid.org/0000-0002-5166-4873

Abstract. Due to the rapid growth of smart grid applications all over the globe, it has become a more attractive target to malicious actors. Countries and stakeholders (e.g., governments) spend billions of dollars on ensuring the continuity and security of their smart grids for strategic and operational reasons. In fact, the risk associated with compromising a smart grid is considered among the highest in the cybersecurity world. This paper surveys a group of well-known smart grid intrusion detection datasets that are used in the development of machine learning-based intrusion detection systems. The study presents an analysis of these datasets and provides recommendations for researchers utilizing them.

Keywords. smart grid; IIoT; security; intrusion detection; datasets

1. Introduction

The rapid developments in the field of Internet-of-Things (IoT) made it an integral part of our daily lives. With a wide varieties of applications ranging from wearable health devices to self-driving cars and Industrial IoT (IIoT), these devices are automating many daily tasks for the end users. The current smart grid applications, in term of electricity and power consumption management, fall in utilizing the Advanced Meter Infrastructure (AMI) (*a.k.a.* smart meter) to mainly gain knowledge about the actual need of power and availability (demand-response). Other important use of smart grid/meter is to verify outage and restoration events. Table 1 summarizes four of the AMI applications.

One particular area witnessed noticeable adoption and investments in recent years is the smart grid. The smart grid is defined in the European Unions Smart Grids Strategic Research Agenda as:

> *A Smart Grid is an electricity network that can intelligently integrate the actions of all users connected to it generators, consumers, and those that do*

Table 1. Smart grid characteristics and purposes.

Applications	Characteristics	Data sending interval	Data rate per node	Data size (bytes)	Delay	Reliability
Outage management	Event based, Delay Tolerant	1 per meter per power lost	56 kbps	25	2 s	>98%
Demand response	Semi periodic (Delay tolerant)/Event based (Mission Critical)	1 per device per Broadcast request	14-100 kbps	100	500 ms-1 min	>99.5%
Distribution automation	Semi-Periodic, Delay Sensitive	1 per device per 12 h	9.6-56 kbps	150-200	25-100 ms	>99.5%
Meter reads	Periodic, Delay Tolerant	5 min, 10 min, 15 min, 30 min, 1 h	10 kbps to 128 kbps	200	2-5 s	>98%

both in order to efficiently deliver sustainable, economic and secure electricity supplies.[1]

Other researchers define the smart grid as a power supply network that is based on digital communication technologies that control the operation of the grid [2]. There exist organizations, i.e., Electrical Power Research Institute (EPRI), the National Institute of Standards and Technology (NIST), and European Commission Research (ECR), which are working towards developing comprehensive frameworks, communication specifications and standards based on the existing ICT standards to realize the vision of smart grid. Just like any other power generation and supply network, the smart grid has transmission lines, substations, and transformers. The transmission of signals in smart grid are mainly distinguished based on the transmission medium, thus being divided into wired and wireless. Power Line Communication (PLC) is a popular wired technology that utilizes the existing power lines for signal transmission. Wireless communication technologies such as ZigBee, WLAN, cellular communication, WiMAX, can be used for connecting distant and unreachable areas. Therefore, smart grid However, it has the capacity to control all of these components, and more, digitally. This enables the smart grid the capacity to perform intelligent monitoring, control, communication, and self-healing when an issue occurs. These advancements made the smart grid a very appealing development that is being increasingly adopted in many countries around the world.

Figure 1 shows the expected growth in the smart grid technology market size in the coming three years. The market is speculated to grow by over 15% in the coming three years to exceed 55 billion USD. Despite the recent advances towards making smart grid a reality, there exist several open issues such as interoperability, cyber and physical security, lack of communication and architectural standards, etc., that require further research and development efforts. More importantly to mention, this growth makes the smart grid an attractive target for malicious actors. While compromising a home IoT device, such as a smart light, or a smart camera, might be considered a reasonable risk, the risk is significantly higher when the target is a smart grid or a nuclear power plant.

Many of the protocols used in managing the smart grid are either dedicated smart grid protocols, or Supervisory Control and Data Acquisition (SCADA) or

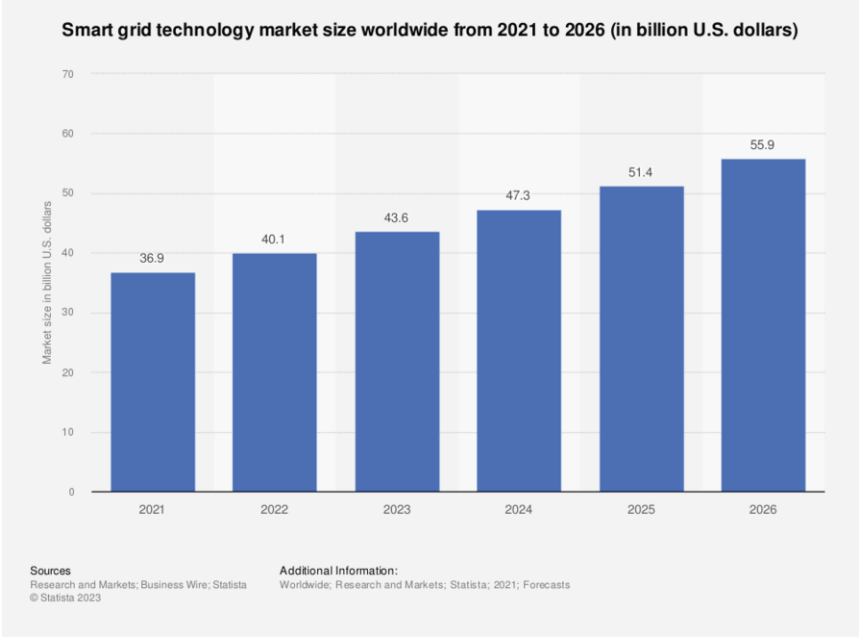

Figure 1. Smart grid technology market size worldwide from 2021 to 2026 [3].

Industrial IoT (IIoT) protocols. In addition, the amount, speed, and sensitivity of the data exchanged in the smart grid is noticeably different from other computer environments, and other IoT contexts. This makes generalized network or IoT intrusion detection datasets, such as CIC-IDS-2017 [4], UNSW-NB15 [5], and TON_IoT [6], less likely to capture the unique nature of network traffic, and attacks on the smart grid. Due to the above mentioned factors, it became increasingly necessary to develop dedicated datasets that would capture the essence of the smart grid and its unique traffic characteristics and features. Thus, many datasets were made publicly available for researchers to study the threat landscape of the smart grid.

This paper presents a summarized review of the intrusion detection datasets available publicly for the smart grid. The study presents comparative analysis of these datasets and makes recommendations to researchers utilizing these datasets in building machine learning-based intrusion detection solutions.

The following section covers the basics of the communication protocols used on the smart grid. Section 3 discusses the different available datasets, while Section 4 presents a comparative analysis of the reviewed datasets. The last section provides conclusions derived from the comparative analysis along with a set of recommendations for researchers.

2. Smart Grid Network Protocols

While the smart grid remains connected to other parts of the Internet using the plain old Internet Protocol (IP) [7], most of the functional operations within the

smart grid use other specialized protocols that are better fit for the purpose. The following subsections discuss four protocols that are considered the most widely used in the smart grid context [8].

2.1. IEC 60870-5-104

IEC 60870-5 is a set of industrial communications protocols, with -104 being its most popular transport protocol in smart grids. IEC 60870-5-104, referred to as IEC104 for short, is an extension of IEC 60870-5-101 protocol with developments in its transport, network, datalink, and physical layer services.

While this protocol has gained popularity, there have been many security concerns associated with it. This pushed IEC to issue an updated security standard name IEC 62351 that implements encryption and network monitoring to address man-in-the-middle (MITM), and replay attacks [9].

2.2. IEEE 1815 (DNP3)

Distributed Network Protocol 3 (DNP3) is a set of protocols designed to facilitate the communication between process automation system's components. The protocol was derived from IEC 60870-5 before it was finalized as a standard. In this protocol, a SCADA master station, or stations, are connected to Remote Terminal Units (RTUs), and Intelligent Electronic Devices (IEDs) via any type of communication networks. RTUs would be intelligent enough to be able to send sensing data and receive commands from the SCADA master using any type of network (mostly computer networks). IEDs would be the smart devices connected to the RTU such as sensors, actuators, or Programmable Logic Controllers (PLCs) [10].

While DNP3 is generally a popular SCADA networking protocol, it has gained popularity in smart grids in Northern America.

2.3. IEEE 2030.5 (SEP2)

IEEE 2030.5 Smart Energy Profile 2.0 (SEP2) is a communication protocol designed to facilitate communication between the smart grid and consumers. It was first coined in 2016 [11] as an IoT-based protocol to facilitate the exchange of essential data such as energy usage, pricing, and demand response. The protocol gained popularity in enabling secure communication between varying ecosystems of smart grid consumer devices.

2.4. OpenADR

Open Automated Demand Response (OpenADR) is an open standard for managing grid-to-consumer communications. The main concept behind this protocol is to turn off high power-consumption devices during the peak time. In its earliest form, it was released in 2009 [12]. It enables communicating information such as pricing, demand response, and energy use, as the case in IEEE 2030.5 (SEP2).

3. Security Datasets

Many previous smart grid security-related publications rely on the power and transmission information to detect attacks after they happen. In this study, we focus on the network security aspects of these systems to detect the attack originating in the data network, rather than detecting the outcome of the attack on the power network.

One of the oldest intrusion detection datasets that is still being currently used in research studies is NSL-KDD dataset that was introduced in 2009 [13]. Despite its popularity in smart grid intrusion detection research, as well as general intrusion detection applications, it is worth noting that the dataset is obsolete, and it does not include traffic captured in a smart grid, or IIoT environments. It's main advantage is that it captures a wide range of attacks, and a high number of samples (1,074,992 samples, including about 80% attack and 20% benign). The dataset was an improvement of the data that was originally presented in KDD dataset back in 1999.

In 2012, the University of New Brunswick presented ISCX-2012, another dataset used in several other smart grid intrusion detection publications [14]. During a seven-day period, 19 network flow features were captured for the purpose of creating this dataset. The attacks were performed on eight different application layer protocols such as File Transfer Protocol (FTP), Netbios, and Domain Name System (DNS). The total number of network flow samples captured and extracted was 2,450,324. While at its time, this dataset was considered comprehensive, currently, however, it is considered outdated.

In 2015, Moustafa presented UNSW-NB15 intrusion detection dataset [5]. This extensive dataset captured 2,540,004 network flow samples. The dataset was synthetically created using IXIA PerfectStorm tool in the Cyber Range Lab of UNSW Canberra. The dataset included nine types of attacks:

- Fuzzing.
- Analysis.
- Backdoors.
- DoS.
- Exploitation.
- Generic Attacks.
- Reconnaissance.
- Shell codes.
- Worms.

The dataset was created by extracting 49 network flow features from the captured packets. Argus and Zeek were used to extract those features. While the dataset seems interesting for intrusion detection solutions, it does not capture smart grid, or IIoT traffic.

In 2022, Radoglou-Grammatikis *et al.* presented DNP3 intrusion detection dataset [15]. The dataset was collected from network captures from a testbed of eight industrial entities, with one human machine interface. These eight entities represented different types of DNP3 slave units such as RTUs, and IEDs. Three additional machines were used as attacking machines. These machines performed the following attacks over the period of several days:

- DNP3 Disable Unsolicited Message Attack.
- DNP3 Cold Restart Attack.
- DNP3 Warm Restart Attack.
- DNP3 Enumerate Attack.
- DNP3 Info Attack.
- Data Initialization Attack.
- MITM-DoS Attack.
- DNP3 Replay Attack.
- DNP3 Step Application Attack.

The dataset features were extracted using CICFlowMeter [16] and a custom DNP3 Python Parser. The network flow timeout was considered 120 seconds.

The resulting dataset included 40,420 network flows with 99 features for each network flow. The labels used were nine labels; eight attack labels, and one normal flow label.

The same authors presented, in 2022 as well, an intrusion detection dataset for IEC 60870-5-104 in [17]. In a similar methodology to [15], this dataset was built up using seven industrial entities, one human machine interface, and three attacking machines. The industrial entities used IEC TestServer [18], while the human interface machine utilized QTester104 [19]. The attacks performed were:

- MITM Drop.
- C_CI_NA_1 (Counter Interrogation command in the control direction).
- C_SC_NA_1 (Sending unauthorised C_SC_NA_1 60870-5-104 packets to the target system).
- C_SE_NA_1 (Sending unauthorised IEC 60870-5-104 C_SE_NA_1 packets to the target system).
- C_RD_NA_1 (Sending unauthorised IEC 60870-5-104 C_RD_NA_1 packets to the target system).
- C_RP_NA_1 (Sending unauthorised IEC 60870-5-104 C_RP_NA_1 packets to the target system).
- M_SP_NA_1_DoS (Flooding the target system with IEC 60870-5-104 M_SP_NA_1 packets).
- C_CI_NA_1_DoS (Flooding the target system with IEC 60870-5-104 C_CI_NA_1 packets).
- C_SE_NA_1_DoS (Floods the target system with IEC 60870-5-104 C_SE_NA_1 packets).
- C_SC_NA_1_DoS (Flooding the target system with IEC 60870-5-104 C_SC_NA_1 packets).
- C_RD_NA_1_DoS (Flooding the target system with IEC 60870-5-104 C_RD_NA_1 packets).
- C RP NA 1 DoS (Flooding the target system with IEC 60870-5-104 C RP NA 1 packets).

The features were extracted using CICFlowMeter and a custom IEC 60870-5-104 Python Parser. The number of features extracted was 83 features from CICFlowmeter, and 111 features from the custom parser, while the total number of network flows was 6,828. The data was labelled as normal traffic, in addition to 12 attack labels.

Table 2. Comparing smart grid intrusion detection datasets.

Dataset	Protocol(s)	Features	Samples	Attack(s)	Multiclass	Balanced	SG
NSL-KDD [13]	IP	41	1,074,992	4	✓	✗	✗
ISCX-2012 [14]	IP, HTTP, FTP, Netbios, DNS, etc.	19	2,450,324	5	✓	✗	✗
UNSW-NB15 [5]	IP, SSH, HTTP, etc.	49	2,540,004	9	✓	✓	✗
DNP3 [15]	DNP3 and IP	99	40,420	9	✓	✗	✓
EIC 60870-5-104 [17]	EIC 60870-5-104 and IP	83 + 111	6,828	12	✓	✗	✓

4. Comparative Analysis

As mentioned in Section 3, many of the datasets used in smart grid intrusion detection research were not actually captured in a smart grid setting. In fact, a large portion of the research on smart grid security is conducted using datasets that are used for network intrusion detection [20].

Table 2 shows a comparative summary of the reviewed datasets. The summary lists the types of protocols captured in the dataset, the number of features, the number of samples, the number of attack types, whether the dataset is labelled as a binary (attack vs. benign) or multiclass (each attack type is defined in the label), whether the classes are balanced in the dataset, and whether the dataset was captured in a smart grid environment.

As discussed earlier, NSL-KDD samples were collected in the late 1990s. This means that the dataset is outdated, and does not cover the most recent types of attacks. The same note can be said about ISCX-2012, as it was collected in 2012. In addition, both datasets are general intrusion detection datasets that do not capture IoT, IIoT, or smart grid traffic. While there are several attacks that are common between a classical computer network, and the smart grid, the smart grid utilizes many protocols that are considered the backbone of the grid. Those protocols are not conventionally used in none industrial environments, such as Internet-of-Medical-Things (IoMT), and smart home applications.

While UNSW-NB15 is considered a slightly newer and more comprehensive dataset, it also misses the types of protocols and traffic used in the smart grid environment. While many researchers ignore this fact and utilize such generalized datasets, it is noteworthy to mention that the nature of the smart grid network traffic is somewhat unique. The smart grid traffic can be generally characterized as small bursts of control information, continuous slow flows of sensing data, and protocol-specific commands and packet headers. Such unique traits makes the attacks on the smart grid significantly different from the ones performed on a corporate or personal computer network.

The remaining two datasets, DNP3 and EIC 60870-5-104, capture the essence of smart grid traffic by being focused on SCADA protocols that are popularly used in smart grids. While DNP3 presents a larger number of samples, it presents 9 types of attacks. In comparison, EIC 60870-5-104 presents a smaller number of samples, with a higher number of attacks captured. However, the number of samples in DNP3 is significantly larger, and hence, it is more suitable for machine learning-based applications. Both of these datasets suffer from class imbalance.

However, this can be addressed using techniques such as random over-sampling of minority classes [21].

5. Conclusions

In this paper, we presented a summarized survey of datasets used for smart grid intrusion detection solutions. In this survey, we briefly reviewed smart grid protocols, and presented a comparative analysis of five popularly used datasets.

While the analysis did not present one particular dataset that can be utilized to build reliable smart grid intrusion detection systems, DNP3 presents itself as the most suitable candidate with a reasonable number of samples, and a focus on SCADA protocol that is commonly used in smart grids.

References

[1] Smart grid regional group; 2023. [Online; accessed 8. Apr. 2023]. Available from: https://energy.ec.europa.eu/topics/infrastructure/projects-common-interest/selection-process-old/smart-grid-regional-group_en#:~:text=A%20smart%20grid%20is%20defined,economically%20efficient%20and%20sustainable%20power.

[2] Fang X, Misra S, Xue G, Yang D. Smart grid—The new and improved power grid: A survey. IEEE communications surveys & tutorials. 2011;14(4):944-80.

[3] Smart grids worldwide - Statista. Statista. 2023 apr. [Online; accessed 6. Apr. 2023], https://www.statista.com/study/111848/smart-grids-worldwide.

[4] Sharafaldin I, Lashkari AH, Ghorbani AA. Toward generating a new intrusion detection dataset and intrusion traffic characterization. ICISSp. 2018;1:108-16.

[5] Moustafa N, Slay J. UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In: 2015 military communications and information systems conference (MilCIS). IEEE; 2015. p. 1-6.

[6] Moustafa N. A new distributed architecture for evaluating AI-based security systems at the edge: Network TON_IoT datasets. Sustainable Cities and Society. 2021;72:102994.

[7] Alani MM, Alani MM. Tcp/ip model. Guide to OSI and TCP/IP models. 2014:19-50.

[8] Tightiz L, Yang H. A comprehensive review on IoT protocols' features in smart grid communication. Energies. 2020;13(11):2762.

[9] Maynard P, McLaughlin K, Haberler B. Towards understanding man-in-the-middle attacks on iec 60870-5-104 scada networks. In: 2nd International Symposium for ICS & SCADA Cyber Security Research 2014 (ICS-CSR 2014) 2; 2014. p. 30-42.

[10] Clarke G, Reynders D, Wright E. Practical modern SCADA protocols: DNP3, 60870.5 and related systems. Newnes; 2004.

[11] IEEE 2030.5 (Smart Energy Profile 2.0): An Overview and Applicability to Distributed Energy Resources (DER) - IEEE Smart Grid, available from: https://smartgrid.ieee.org/resources/webinars/non-bulk-generation/ieee-2030-5-smart-energy-profile-2-0-an-overview-and-applicability-to-distributed-energy-resources-der; 2023. [Online; accessed 7. Apr. 2023].

[12] Watson DS, Piette MA, Sezgen O, Motegi N. Machine to machine (M2M) technology in demand responsive commercial buildings. In: Handbook of Web Based Energy Information and Control Systems. River Publishers; 2020. p. 429-46.

[13] Tavallaee M, Bagheri E, Lu W, Ghorbani AA. A detailed analysis of the KDD CUP 99 data set. In: 2009 IEEE symposium on computational intelligence for security and defense applications. Ieee; 2009. p. 1-6.

[14] Shiravi A, Shiravi H, Tavallaee M, Ghorbani AA. Toward developing a systematic approach to generate benchmark datasets for intrusion detection. computers & security. 2012;31(3):357-74.

[15] Radoglou-Grammatikis P, Kelli V, Lagkas T, Argyriou V, Sarigiannidis P. DNP3 Intrusion Detection Dataset. IEEE Dataport; 2022. Available from: https://dx.doi.org/10.21227/s7h0-b081.

[16] ahlashkari. CICFlowMeter; 2023. [Online; accessed 7. Apr. 2023]. Available from: https://github.com/ahlashkari/CICFlowMeter.

[17] Radoglou-Grammatikis P, Rompolos K, Lagkas T, Argyriou V, Sarigiannidis P. IEC 60870-5-104 Intrusion Detection Dataset. IEEE Dataport; 2022. Available from: https://dx.doi.org/10.21227/fj7s-f281.

[18] SourceForge; 2023. [Online; accessed 7. Apr. 2023]. Available from: https://sourceforge.net/projects/iecserver.

[19] SourceForge; 2023. [Online; accessed 7. Apr. 2023]. Available from: https://sourceforge.net/projects/qtester104.

[20] Radoglou-Grammatikis PI, Sarigiannidis PG. Securing the smart grid: A comprehensive compilation of intrusion detection and prevention systems. IEEE Access. 2019;7:46595-620.

[21] Raschka S, Liu YH, Mirjalili V, Dzhulgakov D. Machine Learning with PyTorch and Scikit-Learn: Develop machine learning and deep learning models with Python. Packt Publishing Ltd; 2022.