On IT and OT Cybersecurity Datasets for Machine Learning-Based Intrusion Detection in Industrial Control Systems

Mohammad Pasha Shabanfar, Yiheng Zhao, Jun Yan(✉), and Mohsen Ghafouri

Concordia University, Montreal, Canada
{mohammadpasha.shabanfar,yiheng.zhao}@mail.concordia.ca,
{jun.yan,mohsen.ghafouri}@concordia.ca

Abstract. Intrusion detection plays a pivotal role in the cybersecurity of industrial control systems (ICS) to safeguard the safety of individuals, communities, and nations. Lately, intrusion detection models based on machine learning have been adopted to improve the detection of cyberattacks. However, there is a lack of a systematic approach to selecting the appropriate dataset for training these models. An appropriately selected dataset should be based on the needed collection environment, i.e., Information Technology (IT) and Operational Technology (OT), and include required specifications of the under-study ICS, e.g., deployed protocols. On this basis, this paper classifies the existing intrusion detection datasets into IT and OT datasets. The IT datasets are investigated from the perspectives of attack/normal traffic inclusion and their anonymity, number of packets, duration, and kind of traffic. On the other hand, the OT datasets are studied based on features such as data protocols, distribution, and data domain. Then, we have discussed the gap between the method of detection and the selection of the appropriate dataset in terms of (i) performance indicators, i.e., detection time and imbalanced distribution of data, and (ii) use case, i.e., summarizing communication layers, protocols, and attack types contained in datasets. Finally, the essential features for constructing an effective cybersecurity dataset are discussed to illustrate how to establish an ideal dataset accordingly.

Keywords: Information Technology · Operational Technology · Datasets · Cybersecurity · Intrusion Detection System

1 Introduction

According to the statistics reported for cybersecurity, the damages caused by cyberattacks are expected to reach up to three trillion by 2021, with the probability of executing zero-day exploits one per day. Moreover, the amount of information stored in private and public clouds operated by data-driven companies, such as Amazon Web Services, Facebook, and Twitter has been increased

X. Li et al. (Eds.): SmartGift 2024, LNICST 600, pp. 36–53, 2025.
https://doi.org/10.1007/978-3-031-78806-2_3

a hundred times by 2022 [1]. As a result, there would be a need for appropriate detection systems.

Machine learning (ML) methods are one of the commonly used solutions that are increasingly popular and effective in detecting malware and cyber attacks; however, selecting an efficient dataset for training them is essential. Knowing the dataset collection environment and their associated use cases would help researchers to choose the most appropriate dataset for training their methods. According to the kind of dataset testbed, we can classify them into two subsets of IT and OT.

Due to the importance of Information Technology (IT) security, much effort has been spent researching intrusion and insider threat detection [2]. Many papers have been published for security-related data, detecting attacks, etc. All of them need a network-based testbed. During these years, some good IT datasets have been published to evaluate the detection methods' power. Given a labeled dataset in which each data point is assigned to the class normal or attack, the number of detected attacks or false alarms may be used as evaluation criteria [2].

On the other hand, Operational Technology (OT), which includes Industrial Control Systems (ICSs), plays an influential role in managing and supervising processes in the industry, such as water, energy, gas, chemical, etc. Although improving technology affected deterring attacks, the risk of cyberattacks is still increasing. To respond to these security threats targeting ICSs, a security technology that reflects the ICS operating environment is needed [3]. Industrial Detection System (IDS) is in charge of detecting suspicious activities and cyberattacks. Generally, IDS monitors the environment and triggers alerts following any suspicious activity. Moreover, IDS adoption in ICS is being influenced by the increasing number of ICS attacks and their consequence. As a result, several ICS datasets have been published in different domains (such as gas pipelines, power systems, etc.) that give us useful information.

The ICS experimental environment generally consists of three levels, which have been shown in Fig. 1 [3]. Devices should be located and set up when building the environment. In addition, a system for collecting various data is arranged during the ICS operation.

OT consists of compassionate information regarding industrial process operations. Unlike the IT domain, industries are reluctant to share their confidential and sensitive operational data for analysis. Therefore, The researchers are forced to utilize the publicly available ICS datasets, which are outdated and lack the right classification of their use cases [4].

In recent years, cyber security solutions have started to deploy big data analytics to correlate security events across multiple data sources, providing, amongst others, early detection of suspicious activities. Methods employed in cyber data analytics are predominantly based on ML, which needs appropriate data used for specific use cases [5].

However, there is a gap between choosing a dataset and using an appropriate method. This gap can be divided in terms of performance indicators and use cases. Detection time and imbalanced datasets are the most important perfor-

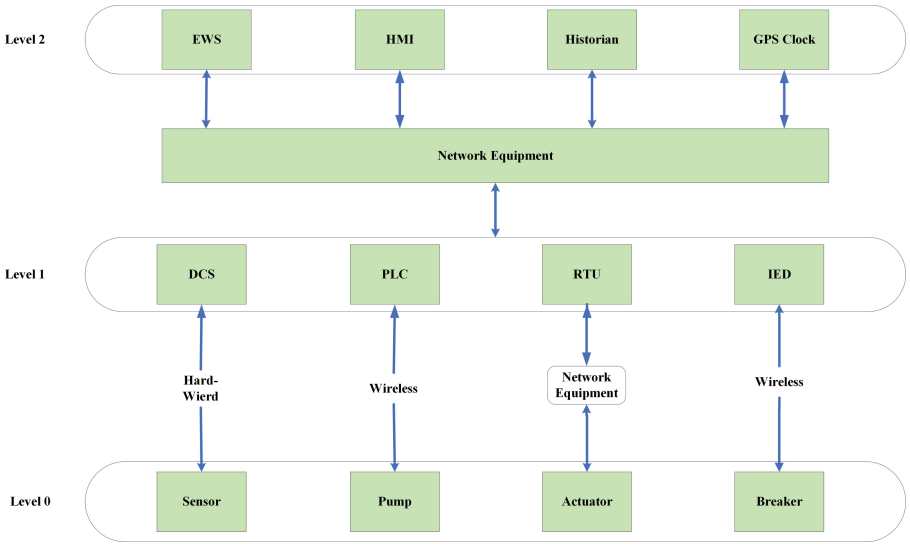

Fig. 1. ICS Environment Rating

mance indicators. Imbalanced data could lead to inefficient results for testing an ML method. On the other hand, a prolonged detection time for a cyber attack might result in overwhelming damage or crashing a big part of a system. In terms of use cases, IT and OT datasets are two different types of datasets. As a result, communication layers, protocols, and attack types contained in different datasets are summarized, which can help researchers more conveniently choose the dataset appropriate for their objectives. Furthermore, performing a detection algorithm for datasets with different mapping OSI layers will yield inaccurate results. Thus, knowing this mapping helps researchers design more effective algorithms for testing their on-target dataset.

In this paper, we introduced the following:

- A new mapping of IT and OT datasets has been introduced that gives better details of each dataset to researchers for designing more effective detection methods for cybersecurity attacks.
- Two performance indicators, i.e., detection time and imbalanced distribution of data, have been studied to assist in selecting suitable ML algorithms for various systems and explore how imbalanced data affects these algorithms' performance.
- Critical features for generating an ideal dataset have been investigated. These features can be useful to generate a real-world dataset that researchers can use to train and evaluate their detection methods in their application systems.

The rest of the paper is organized as follows. In Sect. 2, we discuss the concept of IT and OT first and then compare the IT cybersecurity datasets with OT cybersecurity datasets. After that, we discuss IT security and OT security to classify datasets into two subsets of IT and OT datasets, which will be investigated in Sect. 3. In addition, we analyze some popular ML algorithms to evaluate their performances on the detection time of attacks and evaluate the impact of imbalanced data distribution on the performance of these algorithms in Sect. 4. Moreover, we also delve into additional crucial characteristics of datasets, with the aim of facilitating the selection of the most appropriate dataset tailored to various specific targets in Sect. 4. Then, we introduced how to construct an ideal cybersecurity dataset based on these essential features in Sect. 5. Finally, We have summarized our work in Sect. 6.

2 IT and OT Security in ICS

In this section, first, we discuss the concept of IT and OT and then, compare the IT cybersecurity datasets with OT cybersecurity datasets.

2.1 IT Security in ICS

Security in IT systems is understood conceptually in the academic literature and to a degree in practice in the enterprise environment. IT security measures have evolved over the past two decades from a binary 'secure or not secure' measurement to one based on risk. Since risk management is already a functioning business requirement, the risk management concept has made it easier to integrate security into business decisions. For instance, the CIA triad serves as the foundation for nearly all IT security solutions, which is a description of IT security that goes beyond the scope and focus of this essay [6].

IT security is widely used on the public Internet as well as ICS for many applications, e.g., email, voice-over-IP, in energy, transportation, healthcare, and many other sectors. These networks facilitate internal and external communication for employees, suppliers, and customers. Backend offices of energy companies are another example of maintaining corporate IT networks that handle administrative tasks, finance, human resources, and other business operations. These networks often include servers hosting enterprise applications and databases. For instance, IT networks in the oil and gas industry support exploration, extraction, and production activities. This includes communication between remote drilling sites, data centers, and corporate offices for managing exploration data and production processes.

2.2 OT Security in ICS

Systems employed in manufacturing, transportation, critical infrastructure, cyber-physical systems, and other areas are often referred to as OT systems. These systems are where computers manage operational procedures and make

data accessible to the business [6]. OT communications are widely used in ICS right now. Supervisory control and data acquisition (SCADA) systems are used to monitor and control industrial processes and infrastructure. They use various OT communication protocols to gather data from sensors and control equipment. Distributed Control Systems (DCSs) are used in manufacturing and process industries to control and automate production processes. They rely on dedicated communication networks for real-time control [7].

OT systems have two sources of security specification, one for general-purpose deployments and a second set of requirements driven separately by infrastructure segmentation [6]. Regarding specific sector recommendations on security in the context of control systems, many distinct standards may or may not be applicable depending on the business and other variables [6]. Furthermore, Programmable Logic Control (PLC) regulates machinery and equipment in industrial settings. They use OT communication protocols to receive input signals and send control commands. Also, Generic Object-Oriented Substation Events (GOOSE) [8] communication is primarily associated with OT. GOOSE is a part of the International Electrotechnical Commission (IEC) 61850 suite of standards and specifies the communication of electrical substation events. It is a messaging protocol used in electrical power systems, particularly substation automation and protection systems [9].

2.3 Comparing IT Security with OT Security in ICS

The importance of OT security is as well as IT security. While the systems may not be completely developed from technological security capabilities, they are from a regulatory standpoint. The focus of IT security is on protecting information, networks, and computer systems, while OT systems are related to the control and automation of physical processes, such as manufacturing, industrial machinery, and critical infrastructure. Therefore, even though OT systems (i) may not have many technical level controls, such as access control systems and cryptography, and (ii) though they may not have much to no forensic and logging capability, these features are all governed by regulatory decree. OT personnel are put in a cognitive cage where regulatory compliance trumps security considerations when regulatory edict is used instead of serious security functionality. Cognitively, OT employees may confuse security with regulatory compliance, which is difficult to spot before a significant preventable incident. Figure 2 shows some differences between IT and OT regarding priority, risks, networks, and protocols [10].

3 Datasets for IT and OT Security in ICS

The selection of an appropriate dataset for training ML algorithms for intrusion detection is of paramount importance. To assist researchers in better-selecting datasets to train their ML algorithms based on their application environment, i.e., IT and OT, we have categorized the existing significant intrusion detection

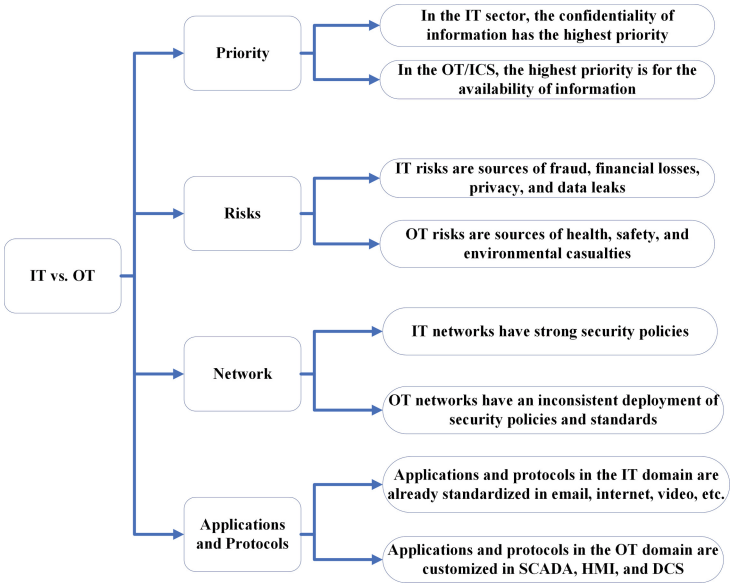

Fig. 2. Differences Between IT and OT

datasets into IT and OT datasets. In this section, we first mention some differences between the IT dataset and the OT dataset. Then, we discuss the existing important IT and OT datasets.

3.1 Comparison of IT Dataset and OT Dataset

The differences between IT and OT security datasets lie in the data collected, the systems and environments they pertain to, and the specific cybersecurity challenges they address. The most significant difference between IT and OT cybersecurity datasets is the environment in which they operate to generate data. OT cybersecurity safeguards industrial environments, typically involving machinery, PLCs, and communication across industrial protocols. OT systems do not run on regular operating systems, often lack traditional security tools, and are usually programmed differently from conventional computers. Conversely, IT cybersecurity protects common devices, such as networks, computers, keyboards, printers, and smartphones. It secures everyday environments like servers using standard solutions, such as antivirus and firewalls, as well as popular communication protocols like Hypertext Transfer Protocol (HTTP).

There are different purposes for IT and OT security based on what they aim to achieve for organizations. The primary objective of OT cybersecurity is to ensure the availability and safety of critical equipment and processes. It maintains physical systems that require meticulous, ongoing control to prevent significant financial damage caused by ceased production. IT cybersecurity focuses more on confidentiality by helping organizations store and transmit data securely.

Another noteworthy difference between OT and IT datasets is the type of security events they defend against. OT cybersecurity datasets are typically generated to put in place to prevent highly-destructive events. OT systems generally have fewer entry points, yet the magnitude of a compromise is comparatively greater-even a minor incident can result in vast financial losses and can affect an entire nation through a power outage or water contamination. IT systems tend to have more gateways and entry points due to the Internet, which a cybercriminal can exploit, which means more security risks and vulnerabilities.

Although IT datasets and OT datasets have some differences, the cyber-attack types of these two types of datasets are similar. There are four general classes of attacks against the integrity, availability, confidentiality, access control, authentication, and non-repudiation security aspects [11]. These attacks include interruption attacks, interception attacks, modification attacks, and fabrication attacks.

An interruption attack includes both hardware-based DoS attacks and software-based DoS attacks [12]. DoS attacks and distributed DoS (DDoS) attacks occur when an attacker hacks several machines (or zombies) and uses up network resources. This overloads the target's bandwidth and causes genuine traffic to be slowed down or dropped. DoS attacks, for example, can result in missed or delayed measurements from EPES devices that rely on real-time measurement data. This leads to incomplete failure of network measurement devices, erroneous forecasts of the transmission system status, and delayed response to power system issues.

Information traveling over the network between devices is accessible to an interception attack. These attacks may take two forms: passive and active. Packet sniffing attacks are a type of interception attack. In which attackers can gain access to the contents of the Phasor measurement unit (PMU) or smart meter Transmission Control Protocol (TCP)/Internet Protocol (IP) packets that are sent across the EPES network using software programs such as Wireshark [11].

Modification attacks use network security flaws to hijack, change, or contaminate a genuine process. Man-in-the-middle (MITM) attacks are one type of modification attack. In MITM attacks, the attacker poses as the legitimate target to both the legitimate client and server during the protocol session [11].

In fabrication attacks, the attacker forges an identity on the IT network and uses it to send fake data that, if improperly verified, could be accepted by other network devices. System spoofing is a type of fabrication attack. System Spoofing: Data accuracy in the IT network is critical for efficient and reliable operation. System spoofing injects fabricated (inaccurate) data into the control centers.

3.2 IT Datasets

An IT dataset can be developed by collecting information from varied sources, such as network traffic flows that contain information about the host, user behavior, and system configurations [13]. This information is required to study various network attack patterns and abnormal activity. The network activity is collected

through a router or network switch. After collecting the incoming and outgoing network traffic, network flow analysis is performed to study the traffic. Flow analysis can be described as analyzing the network packet information such as source IP address, destination IP address, source port number, destination port number, and type of network services, to name a few. The network host delivers the system configurations and user information that cannot be extracted from the network flow analysis [14].

According to the categorization, some public IT datasets frequently used for intrusion detection in ICS are:

CICIDS 2017. CICIDS 2017 [15], generated over a span of five days in an emulated environment, encompasses network traffic presented in both packet-based and bidirectional flow-based formats. The dataset comprises extensive attributes, exceeding 80 for each flow, accompanied by supplementary metadata concerning IP addresses and attack details. It encompasses a wide range of attack types, including but not limited to SSH brute force, Heartbleed, botnet, DoS, DDoS, web, and infiltration attacks.

CIC DOS. The CIC DoS datasetcite [16], sourced from the Canadian Institute for Cybersecurity, was developed with the aim of constructing an intrusion detection dataset featuring application layer DoS attacks. To achieve this, the researchers conducted eight distinct application layer DoS attacks. To generate normal user behavior data, they merged the obtained traces with attack-free traffic extracted from the ISCX 2012 dataset.

DARPA, KDD CUP, NSL-KDD. The DARPA 1998/99 datasets [17], widely recognized as the primary datasets for intrusion detection, were crafted at the MIT Lincoln Lab in an emulated network environment. Comprising packet-based network traffic data, the DARPA 1998 dataset spans seven weeks, while the DARPA 1999 dataset covers five weeks.

KDD CUP 99 [18], derived from the DARPA 98 dataset, ranks among the most extensively employed datasets for intrusion detection purposes. This dataset includes fundamental attributes concerning TCP connections and higher-level features, such as the count of unsuccessful login attempts, though it omits IP addresses.

NSL-KDD [2], an evolved dataset, was created as a response to duplicate data concerns within KDD CUP. This dataset, stemming from the original KDD cup99 dataset, was born after Tavallaee et al.'s analysis of the KDD training and test sets, which unveiled duplicate network packets accounting for around 78

DDOS 2016. The DDOS 2016 dataset [19], constructed in 2016 through the utilization of the NS2 network simulator, adopts a packet-based format. Unfortunately, specific details regarding the simulated network environment remain undisclosed. Within the DDoS 2016 dataset, attention is primarily directed toward various categories of DDoS attacks. In addition to normal network traffic, this dataset encompasses four distinct DDoS attack types: UDP flood, smurf, HTTP flood, and SIDDOS.

UNSW-NB 15. The UNSW-NB15 dataset, as outlined in [19], comprises both regular and malicious network traffic, presented in a packet-based format. This dataset was generated within a confined emulated environment over a period of 31 h, utilizing the IXIA Perfect Storm tool. It encompasses a diverse array of attack categories, including but not limited to backdoors, DoS (Denial of Service), exploits, fuzzers, and worms, forming nine distinct attack families. UNSW-NB15 comes equipped with predefined partitions for training and testing purposes, with a total of 45 unique IP addresses included in the dataset.

Table 1. General Information of IT Datasets

Dataset	Normal Traffic	Attack Traffic	Anonymity	Count	Duration	Kind of Traffic
CIC DoS	yes	yes	none	4.6 GB packets	1 day	emulated
CICIDS 2017	yes	yes	none	3.1 M flows	5 days	emulated
DARPA	yes	yes	none	n.s.	7.5 weeks	emulated
DDoS 2016	yes	yes	yes	2.1 M packets	n.s.	synthetic
KDD Cup 99	yes	yes	none	5 M points	none	emulated
NSL-KDD	yes	yes	none	150 K points	n.s.	emulated
UNSW-NB15	yes	yes	none	2 M points	31 h	emulated

A detailed overview of IT data sets is shown in Table 1. According to Table 1, general information on IT datasets, such as normal and attack data, the amount of the datasets, and their kind of traffic, has been shown. Moreover, most datasets have been generated in an emulated environment, and there are fewer datasets with real network environments. The presence of specific attack scenarios is an important aspect when searching for a network-based data set. According to [2], which describes the specific attacks within IT datasets, DoS, DDoS, port scans, and botnets are the most popular attacks used in the datasets to simulate an abnormal situation in the network.

3.3 OT Datasets

This section discusses public OT datasets used in several surveys to detect attacks by implementing different algorithms and methods. ICSs are one of the most effective tools to prevent cyberattacks. The key components of the ICS include SCADA, Human Machine Interface (HMI), PLC, Remote Terminal Unit, and DCS. A SCADA system helps collect data from field sensors, enabling us to control the system through HMI software [10]. OT monitors all industrial systems, and ICS relates to the security of industrial systems. Thus, ICS datasets are a type of subset of OT datasets.

Here are some public OT datasets used frequently for detecting algorithms to compare them based on four categories that we will discuss in the next subsection.

Morris et al. Datasets [3]. For their research on intrusion detection, Morris et al. have made five separate datasets about the production of electricity, gas, and water available. The Morris datasets can be used for ML in creating intrusion detection systems because they all provide labels in common. The Morris-1 dataset includes 37 scenarios for power system events that consider the number of intelligent electronic device (IED) operations and typical and unusual occurrences in the testbed for power systems comprising generators, IEDs, breakers, switches, and routers. The RS-232, or Ethernet interface in the gas pipeline testbed, is connected in the Morris-2, Morris-3, and Morris-4 datasets to enable Modbus protocol connection between the control device and the HMI. Every dataset has network data information that has some header information removed.

SWaT [3]. The SWaT dataset encompasses sensor data, actuators, PLC input/output (I/O) signals, and network traffic, which were recorded over a duration of four days during an assault scenario and seven days under regular operational conditions. It is worth noting that the SWaT datasets represent one of the most extensive data collections within a substantial testbed. SWaT has meticulously crafted a total of 36 attack scenarios, encompassing both field signals and network traffic. Each attack scenario was meticulously designed by specifying the targeted devices and physical points, with each attack being individually structured. These attack scenarios are meticulously aligned with the operational principles of the physical system. Furthermore, the datasets are well-suited for monitoring research, as they are categorized into distinct segments based on the physical layer and the network layer.

Lemay [3]. Lemay et al. have contributed a network traffic dataset focused on covert channel command and control within the SCADA domain. For the creation of the testing environment, a SCADA network was established utilizing the publicly available SCADA Sandbox tool. Additionally, two master terminal units were implemented through SCADA BR. The dataset encompasses Modbus/TCP communication, involving the connection of three controllers and four field devices per controller.

Rodofile et a. Dataset [3]. It comprises two elevated reservoir tanks, six consumer tanks, two raw water tanks, and a return tank. It contains chemical dosing systems, booster pumps, valves, instrumentation, and analyzers. WADI is controlled by 3 PLCs that operate over 100 network sensors. Moreover, the testbed is equipped with a SCADA system. WADI consists of three main processes: (i) P1 (Primary supply and analysis), (ii) P2 (Elevated reservoir with Domestic grid and leak detection), and (iii) P3 (Return process). Its use cases are to show that the detection mechanism applies to real-world ICS data and to see whether any attack methodology is transferable from a scenario in which simulated data are used to another scenario in which real data are used.

WADI [20]. The WADI testbed comprises a comprehensive facility encompassing two elevated reservoir tanks, six consumer tanks, two raw water tanks, and a return tank. Within this setup, you'll find an array of essential components, including chemical dosing systems, booster pumps, valves, instrumentation, and

analyzers. The control of WADI is managed by three PLCs, each communicating with over 100 network sensors. Additionally, the testbed is equipped with a SCADA system to facilitate monitoring and control. WADI's operations revolve around three primary processes: P1 (Primary supply and analysis), P2 (Elevated reservoir with Domestic grid and leak detection), and P3 (Return process). The primary objectives of this testbed are twofold: firstly, to demonstrate the applicability of the detection mechanism in real-world ICS data, and secondly, to explore the transferability of attack methodologies from scenarios involving simulated data to scenarios employing real data.

EPIC [21]. Data from the EPIC testbed encompasses eight distinct scenarios during normal operation, each scenario spanning approximately 30 min. The data collected includes sensor and actuator information, meticulously recorded in an Excel spreadsheet, and network traffic data, which has been archived in *.pcap* files. EPIC represents a power testbed that faithfully replicates a compact real-world smart grid system, encompassing four essential stages: generation, transmission, microgrid, and smart home. Each stage is under the control of its dedicated PLC/controller. Additionally, communication channels are established between the SCADA system, the DCS, the energy management system (EMS), and each PLC/controller.

WUSTL [22]. This dataset encompasses network data derived from the Industrial Internet of Things (IIoT) for the purpose of cybersecurity research. The primary objective of this testbed is to replicate real-world industrial systems with maximum fidelity, enabling the execution of genuine cyber-attacks for research purposes. A substantial volume of data, totaling 2.7 GB, was accumulated over a period of approximately 53 h. Prior to its release, the dataset underwent thorough preprocessing and cleaning procedures.

Here, we briefly compare the OT datasets. We compare datasets based on their public information, data domain, and size of normal and attack data. Each dataset is collected from its own experimental environment in a specific or complex domain. To specify our analysis target, we limited our study to the ICS-related datasets that can be accessed publicly. Table 2 describes the data domain and size of the normal and attack of some of the datasets as an example.

Table 2. Data Domain and Dimensions of Normal and Attack of OT Datasets

Dataset	Data Domain	Num. of Normal Traffic (%)	Num. of Attack Traffic (%)
Morris5	EMS	16,362(92.09)	1,405(7.91)
Lemay	SCADA	395,298(87.86)	54,321(12.14)
Rodofile	Mining Refinery	1,137,294(63.09)	665,463(36.91)

4 Gaps in Developing ML-Based Algorithms with Existing IT and OT Security Datasets for Diverse Use Cases

Nowadays, the use of ML methods for intrusion detection in cybersecurity has become increasingly important and effective. However, there is a gap between choosing a dataset and using an appropriate method in terms of performance indicators and use cases. This is the point that we will discuss in this section.

4.1 Performance Indicators

In this section, we first evaluated the detection time of various ML algorithms on NSL-KDD and UNSBW-NB15 datasets to choose a suitable algorithm for use cases from different collection environments. This is due to the varied detection time requirements imposed by diverse use cases originating from different collection environments. Then, to illustrate the adverse impact of imbalanced datasets on the intrusion detection performance of the algorithm, we evaluated the intrusion detection performance of the CNN-LSTM algorithm on the original imbalanced Morris Power and CICIDS 2017 datasets, as well as on the Morris Power and CICIDS 2017 datasets that had been preprocessed to achieve balance.

Detection Time. The detection time of different ML algorithms on NSL-KDD and UNSBW-NB15 datasets is shown in Table 3 [23]. According to Table 3, the detection time of different algorithms in descending order is Support Vector Machine (SVM), K-nearest Neighbors (KNN), Gradient Boosting Tree (GBT), Logistic Regression (LG), and Gaussian Naive Bayes (GNB) on these two datasets and the same algorithm has a longer detection time on UNSBW-NB15 dataset since UNSBW-NB15 dataset has more features than NSL-KDD dataset. Therefore, these algorithms, except SVM, can be used for use cases in information system environments that do not require short detection time. However, GNB is an appropriate choice for use cases in operational system environments with high real-time algorithm requirements [23].

Table 3. Detection Time Models Comparison on NSL-KDD and UNSBW-NB15 Datasets

Models	Detection Time(s)	
	NSL-KDD	UNSBW-NB15
Gradient Boosting Tree (GBT)	0.41	0.96
K-nearest Neighbors (KNN)	1.79	5.55
Logistic Regression (LG)	0.24	0.81
Gaussian Naive Bayes (GNB)	0.06	0.21
Support Vector Machine (SVM)	67.26	634.11

Imbalanced Distribution of Data. To compare the intrusion detection results of ML algorithms on imbalanced and balanced datasets, we first need to process the imbalanced dataset into the balanced dataset. Therefore, we used undersampling and oversampling techniques to process two common imbalanced datasets, Morris Power and CICIDS 2017, into balanced datasets. Undersampling is a technique for lowering the proportion of the majority class [24]. This method is adopted when the number of elements belonging to the majority class is rather high. Oversampling, on the other hand, increases the minority class's percentage by randomly reproducing it [24]. Table 4 shows the number of normal and attack data before and after implementing these two technologies on Morris power and CICIDS 2017 datasets. It can be observed that after using undersampling and oversampling techniques, the imbalance of the dataset has been greatly alleviated.

Secondly, to evaluate the impact of imbalanced datasets on algorithm intrusion detection performance, We trained the CNN-LSTM algorithm on imbalanced Morris power, balanced Morris power obtained through undersampling, and balanced Morris power obtained through oversampling datasets, and then evaluated their F1-SCORE performance on the test set of the original Morris Power dataset. We also used the same experimental method on the CICIDS 2017 dataset.

The experimental results are shown in Table 5. The authors of [24] used the CNN-LSTM model to experiment with these balanced datasets. According to that, the CNN-LSTM achieves higher F1-Score results on both undersampled and oversampled balanced datasets compared to the original imbalanced dataset. Specifically, when using the undersampling technique to train the CNN-LSTM on the balanced Morris power dataset, compared to training on the original imbalanced Morris power dataset, this F1-Score was improved by 8.03 on the test set of the original Morris power dataset. Therefore, the dataset should be well-balanced regarding the number of malicious data samples vs. benign traffic samples to achieve adequate results when we use an ML method. However, [25] suggests that resampling to full balance is generally not the optimal resampling rate, at least when the test set is balanced. Furthermore, the optimal resampling rate varies from domain to domain and resampling strategy to resampling strategy.

Table 4. Data Distribution of Morris Power and CICIDS 2017 Datasets

Technique	Morris power		CICIDS 2017	
	Normal	Attack	Normal	Attack
Unbalanced	15,471	38,583	625,757	218,251
Undersampling	15,471	19,338	291,001	218,251
Oversampling	32,425	38,583	625,757	457,547

Table 5. Intrusion Detection Results on Imbalanced and Balanced Morris Power and CICIDS 2017 Datasets

Datasets	Technique	F1-Score
Morris Power	Unbalanced	58.06
	Undersampling	66.09
	Oversampling	64.18
CICIDS2017	Unbalanced	98.44
	Undersampling	99.34
	Oversampling	99.46

4.2 Use Cases

Using the right dataset that results in better accuracy of results comes from being aware of the complete use cases of each cybersecurity dataset. This is part of the gap that we discussed before. Apart from discerning the categorization of this dataset as either IT or OT, this section explains other essential features of the datasets that can be considered when choosing the closest dataset to different targets.

In order to select an appropriate dataset for training an ML algorithm to achieve the desired intrusion detection performance, it is essential not only to determine the system in which the algorithm will be employed, i.e., IT or OT, but also to identify the specific communication layer that the algorithm is intended to monitor for intrusion detection. This is crucial because cyber intrusions are typically executed by exploiting vulnerabilities within a specific layer of communication. We also need to determine which specific protocols and attack types of intrusions the algorithm needs to detect, as multiple protocols may also be included in the same communication layer, and various attacks may also be implemented based on a single protocol. Therefore, communication layers, protocols, and types of attacks on the datasets need to be considered to choose the closest dataset to the researcher's target. To help researchers select datasets based on their goals, we have summarized communication layers, protocols, and types of attack contained in the existing important datasets, and the results are shown in Table 6.

5 How to Build and Use Datasets for Combined IT-OT Security in ICS

In this section, we first discuss the limitations of existing important datasets and then discuss how to establish an ideal dataset based on these essential features.

As shown in Table 6, we have shown the essential features that can determine the applicable scenarios of existing important datasets. First, it can be found that the collection environment of these datasets comes from a single system,

Table 6. Essential Features Determining the Applicability Scenarios of Existing Datasets

Datasets	Systems	OSI Layers	Protocols	Attack Types
CIC DoS	IT	Application	HTTP	DoS
CICIDS 2017	IT	Transport Network	TCP IP	Brute Force DoS DDoS Heartbleed Web Infiltration Botnet
DARPA	IT	Transport Network	TCP IP	DoS Privilege escalation Probing
DDoS 2016	IT	Application Network	TCP UDP ICMP HTTP	DDoS
KDD Cup 99	IT	Transport Network	TCP IP	DoS Privilege Escalation Probing
NSL-KDD	IT	Transpor Network	TCP IP	DoS Privilege Escalation Probing
UNSW-NB15	IT	Transport Network	TCP IP	Backdoors DoS Exploits Fuzzers Worms
Morris	OT	Application	MODBUS	Malicious Response Injection DoS
Lemay	OT	Application	MODBUS	Exploits Fingerprinting Unauthorized Command
SWAT	OT	Application Session Network	CIP Ethernet IP	False Data Injection
Rodofile	OT	Application Presentation Session	S7Comm	Reconnaissance
WuSTL	OT	Transport Application	TCP IP	Port scanner Address scanner Device identification Aggressive model device Exploit
EPIC	OT	Data Link	GOOSE MMS	False Data Injection
WADI	OT	Application Session Network	CIP Ethernet IP HSPA	False Data Injection

i.e., IT or OT. Then, these datasets contain limited communication layers, protocols, and types of attacks. For example, algorithms trained on the CIS DOS dataset can only be used to detect DoS attacks against the HTTP protocol at the application layer on IT systems. Therefore, it is necessary to establish an ideal dataset on which algorithms trained can detect as many different cyber intrusions as possible on both IT and OT.

According to the essential characteristics in Table 6, the ML algorithm developed can be applied in the IT system or the OT system based on the goal of the method. We need enough data in both the normal and attack categories. To do this, the researchers should capture necessary network packets from the host and destination for flow analysis and dataset generation. Then, To record realistic attack scenarios, the data collector should have a thorough understanding of the network topology and how networking devices are configured in the testing environment. Collecting all network packets is sometimes not essential. In some cases, we only need to get particular traffic between hosts such as GOOSE or packets that are transmitted to honeynets. Then, the dataset samples should be mapped to various layers of the OSI layers to ensure that algorithms trained on the ideal dataset can detect attacks against different OSI layers. In ML, labels and annotations are essential for supervised learning. Each data point in the dataset should have well-documented labels that specify the type of attack, whether it is benign or malicious, and other pertinent details such as metadata. Thus, labelling the dataset can be an efficient way to train the ML methods well. Also, the data samples should include as many different protocols as possible, as network intrusions may be based on various protocols. Finally, data samples based on different protocols should also strive to include a wide variety of attack types, ensuring that algorithms trained on the ideal dataset can detect different types of attacks.

6 Conclusion

In this work, we classified cybersecurity datasets into two subsets, IT and OT datasets, and investigated them. Then, we discussed the most applicable existing datasets in both subsets of IT and OT and explained their related features and methods for their generation. In addition, we analyzed some popular ML algorithms to assess their performance in detecting anomalies in two different datasets and explore the impact of imbalanced data distribution on the performance of these algorithms. In addition, we have summarized the essential features of existing datasets that can assist researchers in choosing the closest dataset to their target. Furthermore, we also introduce how to build an ideal dataset based on these essential features.

References

1. Gómez, Á.L.P., et al.: On the generation of anomaly detection datasets in industrial control systems. IEEE Access **7**, 177460–177473 (2019)
2. Ring, M., Wunderlich, S., Scheuring, D., Landes, D., Hotho, A.: A survey of network-based intrusion detection data sets. Comput. Secur. **86**, 147–167 (2019)
3. 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: 13th International Conference, CRITIS: Kaunas, 24–26 September 2018, Revised Selected Papers 13, vol. 2019, pp. 154–166. Springer (2018)

4. Mubarak, S., Habaebi, M.H., Islam, M.R., Khan, S.: ICS cyber attack detection with ensemble machine learning and dpi using cyber-kit datasets. In: 8th International Conference on Computer and Communication Engineering (ICCCE), vol. 2021, pp. 349–354. IEEE (2021)
5. Lin, Q., Verwer, S., Kooij, R., Mathur, A.: Using datasets from industrial control systems for cyber security research and education. In: Critical Information Infrastructures Security: 14th International Conference, CRITIS: Linköping, 23–25 September 2019, Revised Selected Papers 14, vol. 2020, pp. 122–133. Springer (2019)
6. Conklin, W.A.: It vs. OT security: a time to consider a change in CIA to include resilience. In: 2016 49th Hawaii International Conference on System Sciences (HICSS), pp. 2642–2647. IEEE (2016)
7. Murray, G., Johnstone, M.N., Valli, C.: The convergence of IT and OT in critical infrastructure (2017)
8. Kush, N.S., Ahmed, E., Branagan, M., Foo, E.: Poisoned goose: exploiting the goose protocol. In: Proceedings of the Twelfth Australasian Information Security Conference (AISC 2014) [Conferences in Research and Practice in Information Technology, vol. 149, pp. 17–22]. Australian Computer Society (2014)
9. Hoyos, J., Dehus, M., Brown, T.X.: Exploiting the goose protocol: a practical attack on cyber-infrastructure. In: IEEE Globecom Workshops, vol. 2012, pp. 1508–1513. IEEE (2012)
10. 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. **37**(1) (2021)
11. Bedi, G., Venayagamoorthy, G.K., Singh, R., Brooks, R.R., Wang, K.-C.: Review of internet of things (IoT) in electric power and energy systems. IEEE Internet Things J. **5**(2), 847–870 (2018)
12. Beasley, C., Zhong, X., Deng, J., Brooks, R., Venayagamoorthy, G.K.: A survey of electric power synchrophasor network cyber security. In: IEEE PES Innovative Smart Grid Technologies, Europe, pp. 1–5. IEEE (2014)
13. Koch, R.: Towards next-generation intrusion detection. In: 2011 3rd International Conference on Cyber Conflict, pp. 1–18. IEEE (2011)
14. Thakkar, A., Lohiya, R.: A review of the advancement in intrusion detection datasets. Procedia Comput. Sci. **167**, 636–645 (2020)
15. Sharafaldin, I., Lashkari, A.H., Ghorbani, A.A.: Toward generating a new intrusion detection dataset and intrusion traffic characterization. ICISSp **1**, 108–116 (2018)
16. Jazi, H.H., Gonzalez, H., Stakhanova, N., Ghorbani, A.A.: Detecting http-based application layer dos attacks on web servers in the presence of sampling. Comput. Netw. **121**, 25–36 (2017)
17. Lippmann, R.P., et al.: Evaluating intrusion detection systems: the 1998 Darpa off-line intrusion detection evaluation. In: Proceedings DARPA Information Survivability Conference and Exposition (DISCEX 2000), vol. 2, pp. 12–26. IEEE (2000)
18. Lippmann, R., Haines, J.W., Fried, D.J., Korba, J., Das, K.: The 1999 darpa off-line intrusion detection evaluation. Comput. Netw. **34**(4), 579–595 (2000)
19. Moustafa, N., Slay, J.: Unsw-nb15: a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set). In: Military Communications and Information Systems Conference (MilCIS), vol. 2015, pp. 1–6. IEEE (2015)
20. Erba, A., et al.: Constrained concealment attacks against reconstruction-based anomaly detectors in industrial control systems. In: Annual Computer Security Applications Conference, pp. 480–495 (2020)

21. Shen, G., Wang, W., Mu, Q., Pu, Y., Qin, Y., Yu, M.: Data-driven cybersecurity knowledge graph construction for industrial control system security. Wirel. Commun. Mob. Comput. **2020**, 1–13 (2020)
22. Diaba, S.Y., et al.: Scada securing system using deep learning to prevent cyber infiltration. Neural Networks (2023)
23. Zhou, Y., Han, M., Liu, L., He, J.S., Wang, Y.: Deep learning approach for cyberattack detection. In: IEEE INFOCOM 2018-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), pp. 262–267. IEEE (2018)
24. Balla, A., Habaebi, M.H., Elsheikh, E.A., Islam, M.R., Suliman, F.: The effect of dataset imbalance on the performance of Scada intrusion detection systems. Sensors **23**(2), 758 (2023)
25. Estabrooks, A., Jo, T., Japkowicz, N.: A multiple resampling method for learning from imbalanced data sets. Comput. Intell. **20**(1), 18–36 (2004)