Machine Learning for Information Assurance in SCADA Systems

Steve H. Kim, Peter Bayiokos, Constanza Cabrera-Mendoza, Sabrin Kaur Guron,
Wildenslo Osias, Charles C. Tappert and Avery Leider
Seidenberg School of Computer Science and Information Systems, Pace University
Pleasantville, NY 10570, USA
Email: {sk48149w, pb10842p, cc09237p, sg58867n, wo39632n, ctappert, aleider}@pace.edu

***Abstract*—Among the threats facing today's critical infrastructures such as power, water, telecommunication, and gas systems industries are cyber attacks. To protect and defend themselves against cyber threats these aforementioned infrastructures use Industrial Control Systems (ICS). An Industrial Control Systems' main goal is to improve efficiency and controllability while minimizing human input. Malicious events caused by cyber attacks are better managed when they are detected early on. Hence, the need to have a system that is able to constantly monitor all operations and able to accordingly respond to attacks is extremely vital. Machine Learning along with Supervisory Control and Data Acquisition Systems (SCADA), can help build a protection/defense system by automating the process of categorizing/classifying malicious events. By using a data-driven approach, we explore the detection of malicious events.**

***Index Terms*—SCADA systems, infrastructures, inherent risk, ICS framework, machine learning techniques, pattern identification, classification, reinforcement learning, data set, algorithms.**

I. Introduction

To a lot of people, turning the light on and off at home is a very trivial act. However, behind this seemingly trivial act of turning the light on and off is a whole infrastructure that is actively being managed and monitored. A few decades ago, the main threat from delivering power to people's homes would be an unfortunate natural event. Nowadays, with the advance of technology, cyber attacks might supersede natural disasters as the main threat. Hence, industries like chemical plants, assembly lines, or power plants use industrial control systems in their operations in order to detect and respond to threats that might prevent them from providing quality services to their customers. Industrial control systems (ICS) ensure the smooth operations of industrial environments with minimal human interventions. There are two major types of ICS: the Distributed Control System (DCS), where the system is divided into distributed and decentralized subsystems each responsible for its own local process, and the Supervisory Control and Data Acquisition (SCADA), where the control of the entire system is centralized and the system typically spans over a large geographical area [22]. Because ICS aims to minimize human intervention, it finds a friend in Machine Learning (ML). Through MLs use of algorithms, it is able to analyze large amounts of data to identify errors, component deterioration, low quality process optimization, etc. This paper will explore the use of Machine Learning and its combination with the ATT&CK for ICS framework in automating the process of categorizing/classifying malicious events in a SCADA system. This will be done through the following methodology. [10].

A. Methodology

The objective of this research is to highlight the importance of automating the process of categorizing and classifying malicious intrusion events in SCADA systems. This will be achieved through the following consecutive processes.

1) Classify data sets taken from SCADA systems using a data set classification tool to give each data point a specific classification.
2) Thorough analysis of different machine learning methods to identify the machine learning method that best suits the research requirements.
3) Develop the machines' relative 'intelligence' by introducing previously classified data sets. (This is comparable to giving the machine an understanding of the concept of "anomaly" vs. "normal".)
4) Apply the ATT&CK and ICS Framework to the machine learning algorithm in order to automate the process of specifically categorizing/classifying malicious intrusion events in a SCADA system. This will specify what the intrusion is based on the various regulations held under the ATT&CK and ICS Frameworks.
5) Create a final prototype of the automated machine learning intrusion classifier that is permissibly sellable to potential clients.

B. Overview of Machine Learning

Coined by Arthur Samuel, a pioneer in the field of artificial intelligence, Machine Learning can be defined as giving computers the ability to act, through the use of statistical techniques and data, without them being explicitly programmed. Otherwise stated, Machine Learning "is concerned with the question of how to construct computer programs that automatically improve with experience... A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience " [11] as stated by Mitchell in his book Machine Learning. Machine Learning is subdivided in many categories but the three major ones are: Supervised Learning, Unsupervised Learning and Reinforcement Learning. Please see Figure 1 for a diagram describing Machine Learning.

- **Supervised Learning**: Simply put, supervised learning is a process in which we train the machine by using data that is well labeled—data that is tagged with the correct answer. For example, a group of pictures of apples with the tag apple on them would help the computer come up with the rules to classify pictures of apples. Supervised Learning can be, in turn, sub-categorized into classification and regression.
- **Unsupervised Learning**: Unsupervised learning refers to training the machine using non-tagged, non-labeled or non-classified data, thereby allowing the algorithm to act on the data without directions. For example, asking the computer to identify and group items that are frequently bought together on an e-commerce website involves unsupervised machine learning. Unsupervised Learning can be grouped into clustering and association.
- **Reinforcement Learning**: This type of learning focuses on taking appropriate action for the purpose of maximizing reward in a specific circumstance. Reinforcement Learning is mostly used in the search for best possible behavior or path in a certain situation. For example, a robot for accomplishing a certain task as requested.

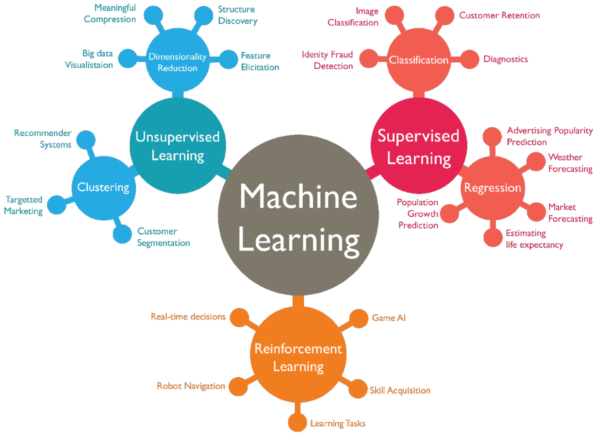


Fig. 1. Machine Learning Overview Diagram.

From web search engines to photo tagging and spam detectors, Machine Learning is being used in many products and services and has become increasingly impactful in the technological realm. As Machine Learning gets closer to our everyday lives, researchers and tech enthusiasts search to better understand the opportunities and the challenges that are associated with it. In this paper, we focus on one of the many opportunities associated with Machine Learning, which is the use of "machine learning algorithms to detect malicious network traffic in SCADA systems and for other intrusion and anomaly detection purposes" [12]. As a subdivision of Supervised Machine Learning, we mention the term classification. Machine Learning Classification is defined as "a process of categorizing a given set of data into classes, It can be performed on both structured or unstructured data" [13]. Classification plays a crucial in the effort to protect data from unauthorized access in that it helps classify data. Given the large amount of redundant data included in network traffic it is pivotal "to develop robust model that can classify the data with high accuracy" [14].

C. Overview of the Reduced Error Pruning Tree Algorithm (REPTree)

The REPTree algorithm is a decision tree learner based on the C4.5 algorithm. C4.5 is used primarily in data mining for decision tree classifiers. The algorithm will generate a decision based on a sample of data.The REPTree produces both a classification or a continuous outcome, which makes it unique over C4.5. Additionally the REPTree offers reduced error pruning, which replaces each node with it's most popular class. This effort of pruning is simple and efficient [15]. Essentially what these types of algorithms do is analyze a piece of data, and then make a decision based on two different factors. Each decision factor produces its own result, which makes it a tree with branches. See an example of the reduced error pruning tree algorithm below (Fig. 2).

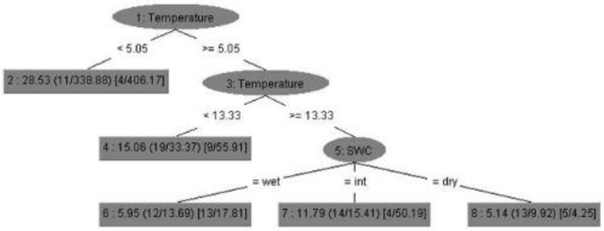


Fig. 2. Reduced Error Pruning Tree Algorithm Diagram.

D. Overview of SCADA Systems

Supervisory Control and Data Acquisition Systems, (SCADA Systems), are collections of hardware and software components used to supervise and control plants, either from a local or a remote location, through examination, collection, and process of data in real time. These SCADA Systems work hand in hand with Human Machine Interface software (HMI), to facilitate indirect user control of hardware in the plants. Remote Terminal Units (RTUs) and Programmable Logic Controllers (PLCs) further enhance the ability of the user to indirectly analyze and respond to plant events. SCADA was introduced to lower the need for onsite personnel at plants. Before, personnel were required on a 24 hr schedule in order to monitor and maintain plant machinery. SCADA eliminated this need by giving personnel the ability to remotely control and supervise plant processes. As industrial plants grew, a larger importance was placed on automation and reliable SCADA systems; focus and money was thrown into this field and it boomed into the subtle yet extremely important industry it is today.

The initial breaths of SCADA systems began back in the 1950s, where processors were first being used to automate control of machinery at plants. The 1960s provided telemetry, (the remote record and transmission of instrument readings), and further facilitated the rise of SCADA systems. SCADA was officially recognized in the 1970s, but was highly inefficient. The SCADA systems were comparable to standalone mainframes, large and cumbersome to manage and maintain. The '80s and '90s saw SCADA system development through the inclusion of Local Area Networking, but still didn't connect efficiently over longer distances because of needed wiring. These wire relying systems were named distributed SCADA systems.

The most impactful change these systems saw in the late '90s and early 2000s was the introduction of open system architectures, which let the systems become more easily accessible by vendors. These Networked SCADA systems almost became obsolete once SQL Databases were introduced. SQL Databases were not anticipated to be combined with SCADA systems, pushing them into a "past technology" category rather quickly. Once these SQL Databases were adapted and incorporated, along with web-based applications, SCADA Systems became what they are today: a staple of various, if not all, plant industries. Present day operator interactions now include real time facilitated responses to SCADA system queues based on field collected data and system analysis from almost anywhere in the world. This system-wide improvement also incorporated trend analysis, company mandated record keeping, plant process automation, and an immense decrease in required personnel. [3] Today, the most common applications of SCADA Systems include the following industries and plants: telecommunications, water and waste control, energy, oil and gas refining and transportation.

II. Literature Review

Information assurance has increasingly become much involved in the supervisory control and data acquisition (SCADA) systems. With the further use of this system, there brings a growing number of security threats and vulnerabilities. "Security Issues in SCADA Networks" highlights the alarming issues present in SCADA networks and the challenges that need to be tackled in order to improve these systems. This paper states that the reason behind the lack of security in organization's that implement SCADA networks is due to the use of commercial off-the-shelf (COTS) hardware and software to develop devices for operating in SCADA networks. With the use of this equipment, the further development of SCADA protocols are needed to create a more secure environment. Even though this paper highlights the security issues SCADA systems hold, there is an increasing effort placed on the information assurance of this system. Ensuring information assurance has become a center-piece to many network frameworks and it allows for sustainability in security for those organizations that incorporate SCADA systems. In addition, there is generally a low involvement in personnel in organizations that have SCADA systems, therefore, the rise in intrusion detection systems (IDS) are becoming more present in these organizations as a result [1].

The paper "Sustainable Security for Infrastructure SCADA" emphasizes the various security structures set in place in SCADA systems and how these systems are built in order to keep the organization's information secure. The paper states there are three elements in sustainable security in SCADA systems: the first is to secure implementations of technology and procedures managed by effective security administration including enforcement and audit; the second is better security technology, including SCADA-specific capabilities; and finally, third party assessment of administration and implementation. SCADA systems influence all tiers of an organization and each level of an organization has some sort of effect on each other. In a SCADA system, the IT control framework has an influence on the security policy, the security policy affects the security plan, and finally the security plan affects the implementation guidance of the security plan. An emphasized security policy, creates for more opportunity to ensure the information assets in the organization are secure. SCADA and IDS systems are relying less on personnel and as a result, the growth in creating a reliable information assurance system is being incorporated into the organizations' security policies. The SCADA security policy needs to have a reliable control objective for the organization to ensure a secure SCADA design, implementation, and operation [2].

According to the "Guide to Industrial Control Systems (ICS) Security," industrial control systems can be defined as the several types of control systems, including SCADA systems, distributed control systems (DCS), and other control system configurations, found in the industrial sectors and critical infrastructures. The ICS framework can be found in many industries such as electric, water and wastewater, oil and natural gas, chemical, pharmaceutical, and food and beverage. This control system consists of a combination of many different control components such as: electrical, mechanical, hydraulic, and pneumatic, that act together in order to achieve an industrial objective such as manufacturing for example. An ICS framework is critical to the organizations that implement them because they require a system that has a strong emphasis on security and they rely on their desired outputs to reflect how efficient the systems are. In an ICS framework the desired output is called the process. In addition, the control part of the system is the specifications required from the desired output and performance. This paper states "control can be fully automated or may include a human in the loop. Systems can be configured to operate open-loop, closed-loop, and manual mode." Open loop control systems are where the output is controlled by the established settings. Closed-loop control settings is where the output has an effect on the input in order to maintain the desired objective. Finally, manual mode is where the system is completely controlled by humans [19].

The paper "A Survey of Approaches Combining Safety and Security for Industrial Control Systems," discusses the growing use of ICS in many organizations and the many security issues that pose a threat to this framework. This paper emphasizes the increasing number of information technologies and communication devices that are being integrated into modern control systems, which increases the degree of complexity and interconnection among systems. There can be various sorts of attacks that will target an organization, therefore, ICSs need to utilize security measures to mitigate cyber-attacks. Having this in mind, there are many data sets that can be analyzed in order to detect the intrusions an organization experienced. The tool Weka is an open source machine learning software that allows for data sets to be analyzed and predictions to be drawn from it. In this case, it allows for the data sets being experimented on, to be visualized and analyzed in order to see the various kinds of attacks an organization has withstood [18].

III. Project Requirements

For this project we will be analyzing SCADA data for post mortem cyber attacks. Using the different parameters in the data and a machine learning algorithm, a determination as to what attack occurred will happen automatically. From this determination, we will compare it against the ATT&CK for ICS framework to provide the type of attack technique and tactic.

The specific machine learning algorithm being used is the Reduced Error Pruning Tree to make decisions based on the data points. Using Weka as our data processing tool we will use REPTree to classify our data to determine the attack that occurred. Once we have a determination of the attacks occurring we can design a system that will create functional alerts based off of the ATT&CK for ICS framework. The framework will also assist in assigning a severity level to the attack ranging from zero to two.

IV. Preliminary Findings

Through our research we hope to create an automated system that applies the MITRE ATT&CK framework for Industrial Control Systems to better alert on cyber attacks carried out on SCADA and ICS systems. The machine learning algorithm will scan through our data set to make a determination on the specific attack, then alert based on the classifications on the MITRE ATT&CK for ICS Matrix.

A. ATT&CK ICS Framework

The ATT&CK ICS Framework is a knowledge base that describes the actions an adversary may take while operating within an ICS network. Attacks on the various infrastructure industries are not uncommon, therefore, the need for a strong cyber security system is necessary. The attacks that are performed on the organizations in these industries not only pose a threat to the organization itself, but it impacts the public and environmental welfare as well [**?**]. For example, if an oil refinery were to be a victim of a cyber-attack, the possibility of an explosion or other sorts of accidents are highly likely, since there was not any control over the system that was maintaining the oil systems. This will not only affect the organization itself, but, it will put lives at risk and create environmental issues. The ATT&CK ICS Framework puts an extra emphasis on the security measures in these industries by creating protocols, applications, incident responses, and etc. to strengthen an organization's ability to fight intrusions. This framework also monitors the threat behavior present in the organization. There are several tactics involved in this framework that allow for the system to know how to react when there is a security threat. These tactics include initial access, execution, persistence, evasion, discovery, lateral movement, collection, command and control, inhibit response function, impair process control, and impact. Within these tactics are techniques used in order to act upon a specific issue. In total there are a total of 81 techniques because there can be more than one solution to a security threat or a technique may provide more security to several areas of focus. The adoption of this framework is not only secure, but it allows for an organization to have options when it comes to selecting a response to an adversary [20].

B. Gas Pipeline Dataset

To begin our research we found it best to use 10% of our data to make it more efficient for the algorithm to ingest and read the data. The REPTree ingests all of our data points and makes a decision based on the parameters we gave it. Our data set was gathered by a laboratory scale industrial control system (ICS) for a gas pipeline hosted by Oak Ridge National Laboratories and Mississippi State University. Together the laboratory and university conducted a series of cyber attacks to the lab gas pipeline ICS. There are a total of 27 parameters used in this ICS, however for our research we only needed 10. We made this determination by first removing all non-changing parameters. Additionally we tested each remaining parameter and evaluated its performance against the REPTree. The parameters that had negligible performance against the algorithm were later removed, to give us our 10 parameters for testing [16]. Please find the parameters in the table below (Table 1).

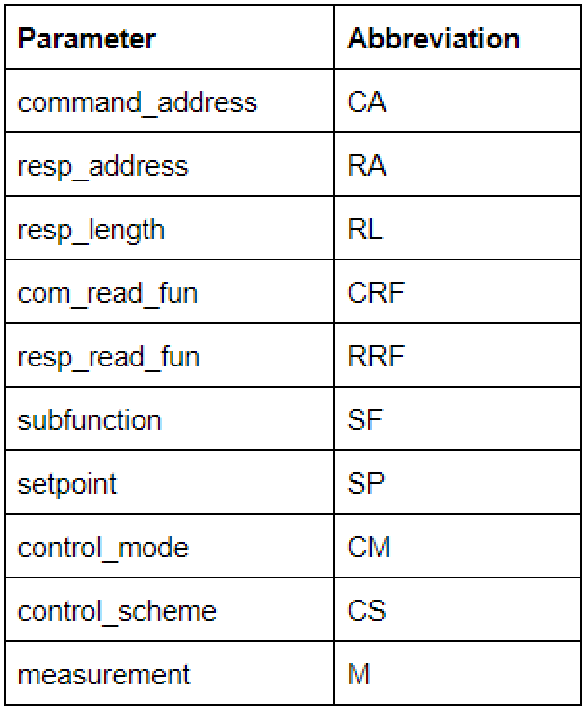
C. Attacks on SCADA/ICS Systems

Since our data was created in a lab controlled environment there were seven attacks carried out on the lab scale gas pipeline. The attacks are listed in a table below with their abbreviations (Table 2).

These are some of the most common attacks carried out on SCADA and Industrial control systems. Also, these attacks are seen as some of the most popular tactics and techniques according to the ATT&CK for ICS framework.

V. Classification/Analysis

Since our data set is fairly large and we are testing for multiple attack vectors we've used Weka to classify the data. Weka is a data processing tool that implements machine learning algorithms to visualize data. This tool allows us to classify our data using machine learning algorithms to then make a determination on the type of attack based on the ATT&CK ICS framework.



Parameter	Abbreviation
command_address	CA
resp_address	RA
resp_length	RL
com_read_fun	CRF
resp_read_fun	RRF
subfunction	SF
setpoint	SP
control_mode	CM
control_scheme	CS
measurement	M

Fig. 3. Data Parameter Table

References

[1] V. M. Igure, S. A. Laughter, and R. D. Williams, "Security issues in scada networks," *computers & security*, vol. 25, no. 7, pp. 498–506, 2006.

[2] J. Stamp, P. Campbell, J. DePoy, J. Dillinger, and W. Young, "Sustainable security for infrastructure scada," *Sandia National Laboratories, Albuquerque, New Mexico (www. sandia. gov/scada/documents/Sustainable Sec urity. pdf)*, 2003.

[3] *What is SCADA?* YouTube, Jun 2019. [Online]. Available: https://www.youtube.com/watch?v=nlFM1q9QPJw

[4] P. Gupta, "Naive bayes in machine learning," Nov 2017. [Online]. Available: https://towardsdatascience.com/naive-bayes-in-machine-learning-f49cc8f831b4

[5] B. Miller and D. Rowe, "A survey scada of and critical infrastructure incidents," in *Proceedings of the 1st Annual conference on Research in information technology*, 2012, pp. 51–56.

[6] M. Hentea, "Improving security for scada control systems," *Interdisciplinary Journal of Information, Knowledge, and Management*, vol. 3, no. 1, pp. 73–86, 2008.

[7] E. J. Byres, M. Franz, and D. Miller, "The use of attack trees in assessing vulnerabilities in scada systems," in *Proceedings of the international infrastructure survivability workshop*. Citeseer, 2004, pp. 3–10.

[8] C. Davis, J. Tate, H. Okhravi, C. Grier, T. Overbye, and D. Nicol, "Scada cyber security testbed development," in *2006 38th North American Power Symposium*. IEEE, 2006, pp. 483–488.

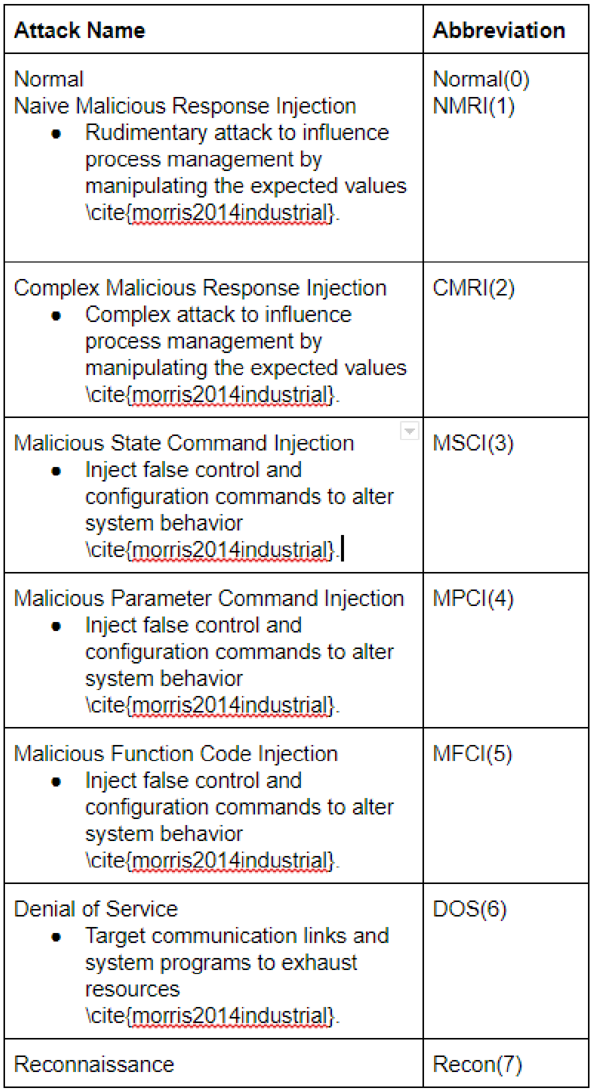
[9] B. G. Becker, "Visualizing decision table classifiers," in *Proceedings IEEE Symposium on Information Visualization (Cat. No. 98TB100258)*. IEEE, 1998, pp. 102–105.

[10] "Guide to idps." [Online]. Available: https://www.esecurityplanet.com/products/top-intrusion-detection-prevention-systems.html

[11] T. M. Mitchell, *Machine Learning*. McGraw-Hill, 1997.

[12] A. AYODEJI, "Machine learning approach to industrial control system health monitoring and cyber security: Similarities, conflicts and limitations."

[13] "Classification in machine learning: Classification algorithms," Dec 2019. [Online]. Available: https://www.edureka.co/blog/classification-in-machine-learning/



Attack Name	Abbreviation

Normal
Naive Malicious Response Injection
• Rudimentary attack to influence process management by manipulating the expected values \cite{morris2014industrial}.
	Normal(0)
	NMRI(1)

Complex Malicious Response Injection • Complex attack to influence process management by manipulating the expected values \cite{morris2014industrial}.	CMRI(2)
Malicious State Command Injection • Inject false control and configuration commands to alter system behavior \cite{morris2014industrial}.	MSCI(3)
Malicious Parameter Command Injection • Inject false control and configuration commands to alter system behavior \cite{morris2014industrial}.	MPCI(4)
Malicious Function Code Injection • Inject false control and configuration commands to alter system behavior \cite{morris2014industrial}.	MFCI(5)
Denial of Service • Target communication links and system programs to exhaust resources \cite{morris2014industrial}.	DOS(6)
Reconnaissance	Recon(7)

Fig. 4. Attack Vector Table

[14] D. P. Mohapatra and S. Patnaik, *Intelligent Computing, Networking, and Informatics Proceedings of the International Conference on Advanced Computing, Networking, and Informatics, India, June 2013*. Springer India, 2014.

[15] M. B. Al Snousy, H. M. El-Deeb, K. Badran, and I. A. Al Khlil, "Suite of decision tree-based classification algorithms on cancer gene expression data," *Egyptian Informatics Journal*, vol. 12, no. 2, pp. 73–82, 2011.

[16] J. Hsu, D. Mudd, and Z. Thornton, "Mississippi state university project report-scada anomaly detection," 2014.

[17] T. Morris and W. Gao, "Industrial control system traffic data sets for intrusion detection research," in *International Conference on Critical Infrastructure Protection*. Springer, 2014, pp. 65–78.

[18] S. Kriaa, L. Pietre-Cambacedes, M. Bouissou, and Y. Halgand, "A survey of approaches combining safety and security for industrial control systems," *Reliability engineering & system safety*, vol. 139, pp. 156–178, 2015.

[19] K. Stouffer, J. Falco, and K. Scarfone, "Guide to industrial control systems (ics) security," *NIST special publication*, vol. 800, no. 82, pp. 16–16, 2011.

[20] "Attck® for industrial control systems." [Online]. Available: https://collaborate.mitre.org/attackics/index.php/Main_Page

[21]

[22] E. H. GICSP, M. Assante, and T. Conway, "An abbreviated history of

automation & industrial controls systems and cybersecurity," 2014.

[23] "An abbreviated history of automation industrial controls systems and cybersecurity."