Machine Learning for Information Assurance in Supervisory Control and Data Acquisition Systems

Richard Alcalde, 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: {pb10842p, cc09237p, sg58867n, wo39632n, ctappert, aleider}@pace.edu
richard@alcalde.us

***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. The main goal for these systems 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, they require a system with the ability to constantly monitor all operations and the ability to accordingly respond to attacks. The combination of Machine Learning and Supervisory Control and Data Acquisition Systems can help build a protection/defense system by automating the process of categorizing/classifying malicious events. By using a data-driven approach, the detection of malicious events was explored. The work demonstrated in the text that follows attempts to show how the Industrial Control Systems framework combined with the machine learning algorithm, Reduced Error Pruning Tree, can be used to design a system that monitors and alerts users to cyber-attacks. Following a brief overview of machine learning, Supervisory Control and Data Acquisition Systems, and Adversarial Tactics, Techniques, and Common Knowledge for Industrial Control Systems Framework, this work presents a step-by step analysis of attacks on a gas pipeline. The last part of this work highlights the opportunities found in furthering research for the automation of systems that keep critical infrastructures secure.**

***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 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 advancement of technology, cyber attacks have become so common they supersede natural disasters as the main threat. Industries like chemical plants, assembly lines, and power plants use industrial control systems in their operations in order to detect and respond to threats that potentially risk preventing them from providing quality service 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 [21]. ICSs aim to minimize human intervention; therefore they find a friend in Machine Learning (ML). ML algorithms are able to analyze large amounts of data to identify errors, component deterioration, low quality process optimization, etc. This paper explores 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. [10].

II. Background

A. 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. Another way of defining it is 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 into a multitude of categories but there are three major ones: Supervised Learning, Unsupervised Learning and Reinforcement Learning.

- **Supervised Learning**: Simply put, supervised learning is a process in which we train the machine by using well labeled data—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.

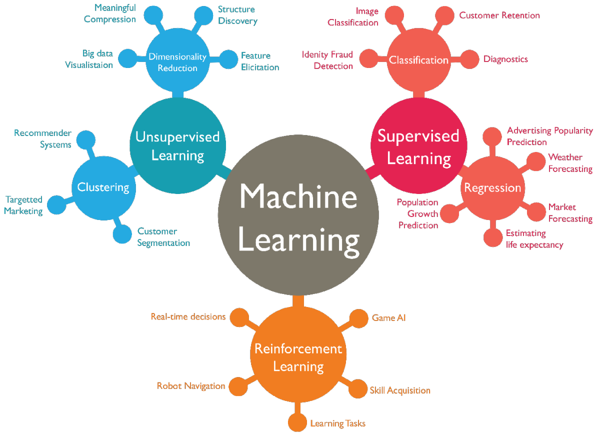


Fig. 1. Machine Learning Overview Diagram. [24]

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. This paper focuses on one of the many opportunities associated with Machine Learning: 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, the term classification is mentioned. 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 role 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 a robust model that can classify the data with high accuracy” [14].

B. 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 generates 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, these algorithms 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).

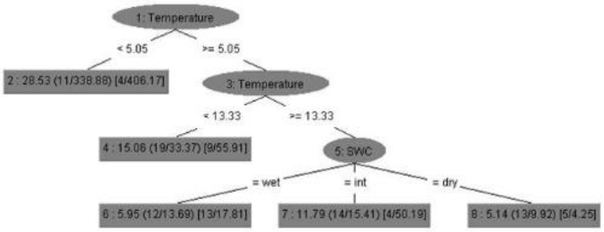


Fig. 2. Reduced Error Pruning Tree Algorithm Diagram. [25]

C. 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, when 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 1990s and early 2000s was the introduction of open system architectures, which let the systems become more easily accessible by vendors. In the same period of time, IT technologies underwent a tremendous transformation which saw Structure query language databases becoming the standard for IT databases. However, SCADA software developers failed to embrace SQL databases which consequently led to SCADA systems being gradually obsolete. Now, while SQL and SCADA paths did not meet at first, along the line they did meet. SCADA software systems have been utilizing the power of SQL databases to improve security, efficiency and reliability. The integration of SQL databases in SCADA software systems largely contributed to its prevalence in so many 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.

III. Literature Review

Information assurance has increasingly become much involved in the SCADA systems. The more this system is used, the more the number of security threats and vulnerabilities grows. "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 denotes 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. Further development of both COTS equipment and of SCADA protocols is 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. Information Assurance has become a center-piece for many network frameworks. It allows for sustainability in security for organizations that incorporate SCADA systems. Given that there is generally a low involvement in personnel in organizations that have SCADA systems, therefore, the rise of intrusion detection systems (IDS) has become more prevalent 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 that 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; finally the third denotes the importance of third party assessment of administration and implementation. SCADA systems influence all tiers of an organization; each level of such 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 more opportunity to ensure the information assets in the organization are secure. SCADA and IDS systems are relying less on personnel and propelling the growth of organizations creating and incorporating a reliable information assurance system 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 controlling components such as electrical, mechanical, hydraulic, and pneumatic. These controllers act together in order to achieve an industrial objective. An ICS framework is critical to the organizations that implement them for two reasons: the first is that they require a system that has a strong emphasis on security; the second is that they compare the actual results to their desired outputs in order to reflect on how efficient the systems are. To ensure that the ICS system is reliable and secure, the design of this system needs to be evaluated. There are a few design considerations for an ICS system, which includes, control timing requirements, geographic distribution, hierarchy, control complexity availability, and impact of failures. With these considerations, the ICS system seems on track towards generating a system prepared for the worst. However, these systems require human involvement. Human error is amongst the top of cybersecurity vulnerabilities for any company. SCADA systems mitigate the errors produced by humans because computers are more statistically ensured to accurately follow instructions appropriate to meet the larger objectives of the system. In addition, the impact of computer related failures can be quantitatively and qualitatively calculated in order to generate path correction for the system. These design considerations provide further support for an automated system. This translates to support for automating intrusion classification in a SCADA system. Human error in misidentification of an intrusion would be mitigated. [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 is an immense variety of attacks that will target an organization. Therefore, ICSs need to implement strong security measures to mitigate cyber-attacks. These security measures have had the pleasant side-effect of providing many datasets concerning the intrusions that an organization has experienced. Weka, an open source machine learning software that allows for datasets to be analyzed and predictions to be drawn from it, allows experimenting with these datasets. Weka can be used to visualize and analyze the various kinds of attacks an organization has withstood. Using Weka, the dataset provided from a gas pipeline will be analyzed using a REPTree algorithm, which will then state which attacks the pipeline has experienced. This literature review supports the classification of cyber intrusions using the REPTree Algorithm, the resource used to achieve our proof of concept for automation of intrusion classification in SCADA systems. This translates into the next section of this paper, as the REPTree proved to be a major part of our study requirements. [18].

IV. 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.

A. Requirements

This study centers around the analysis of SCADA data for post mortem cyber attacks. Through identification of specific parameters in the data and use of a machine learning algorithm, a determination as to what attack occurred will be produced automatically. Using 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. This algorithm will be used to make decisions based on the data points. Using Weka as our data processing tool, we will use REPTree to classify our data into different attack categories. 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.

V. 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 dataset to make a determination on the specific attack, then alert the end-user based on the classifications from the MITRE ATT&CK for ICS Matrix.

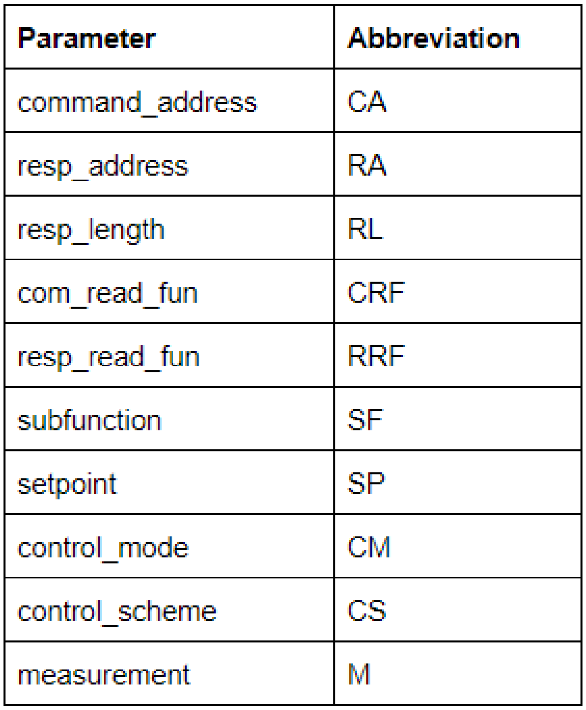
A. ATT&CK ICS Framework

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 infrastructure industries are not uncommon; therefore, the need for a strong cybersecurity system is extremely high. The attacks on the organizations in these industries not only pose a threat to the organizations themselves, but they impact the public and environmental welfare as well [22]. For example, if an unprotected oil refinery were to be a victim of a cyber-attack, the slight possibility of accidents ranging from small leaks to major explosions are highly likely. Not only will this affect the organization itself, it will put lives at risk and create environmental issues. Many modern day industrial control systems do have safety precautions to prevent such disasters, however even the safeties could fail at times. The ATT&CK ICS Framework puts an emphasis on the security measures in these industries by creating protocols, applications, incident responses, and etc. to strengthen an organization's ability to fight cyber intrusions.

This framework also monitors the threat behavior present within the organization. There are several tactics involved in this framework that allow for the SCADA and/or ICS to know how to react when there is a security threat. Some ICSs could also have integrated intrusion detection systems that use this framework as a knowledge base. These tactics include assessing initial access, execution, persistence, evasion, discovery, lateral movement, collection, command and control, inhibit response function, impair process control, and impact. Within these assessment tactics are techniques used in order to act upon a specific issue. Currently, there are a total of 81 accepted techniques. This large number is caused by the variety in solutions to a single security threat. The adoption of this framework is not only secure, it also allows for an organization to have options when it comes to selecting a response to an adversary. [20]

B. Gas Pipeline Dataset

The dataset used for our research was gathered by a laboratory scale industrial control system (ICS) for a gas pipeline hosted by Oak Ridge National Laboratories and Mississippi State University. The laboratory and university partnered to conduct 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]. The parameters are shown below (Fig. 3).



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

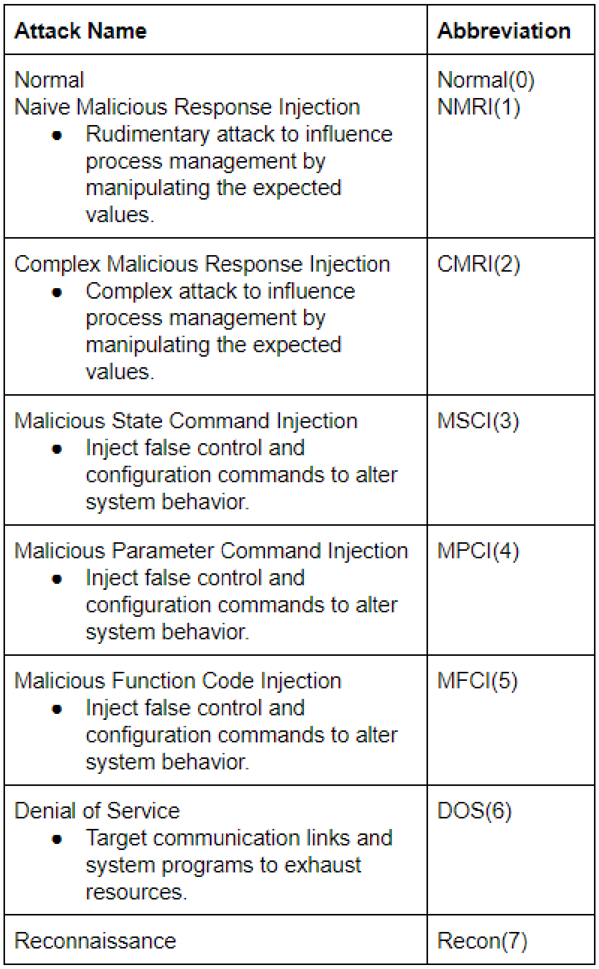
Fig. 3. Data Parameter Table

We found it best to start by using only 10% of our research data. This made the algorithms' ingestion and evaluation of the data more efficient. The REPTree ingested these initial data points and classified them based on the parameters we had edited to our specified preferences.

C. Attacks on SCADA/ICS Systems

There were seven (7) attacks carried out on the lab scale gas pipeline. Note that this number is uncharacteristically low for a gas pipeline. This is only because the attacks were human-generated in a controlled environment. The attacks are listed in a table below with their abbreviations (Fig. 4).

These are some of the most common attacks carried out on SCADA and Industrial Control Systems. These attacks are also renowned as some of the most popular tactics and techniques according to the ATT&CK for ICS framework.



Attack Name	Abbreviation

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

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

Fig. 4. Attack Vector Table

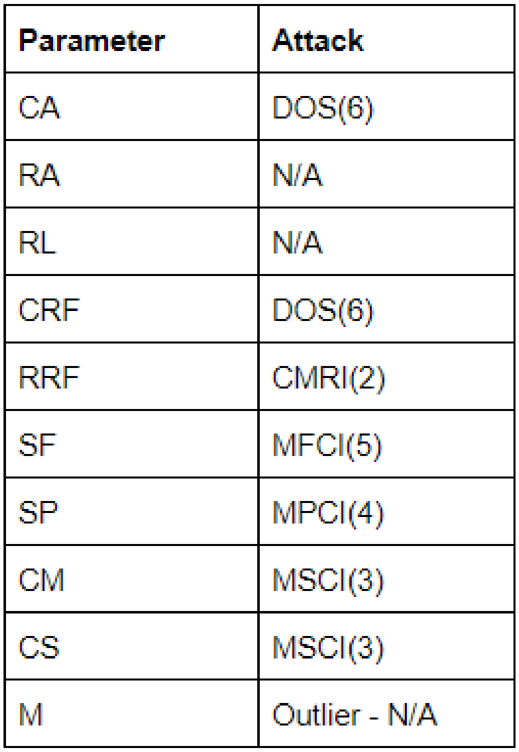
VI. Classification/Analysis

Given the fairly large nature of our dataset, (even the 10% portion was relatively large), and our testing for multiple attack vectors, we decided to use Weka to classify the data. Weka is a data processing tool that implements machine learning algorithms to visualize data. This tool allowed 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.

A. Classification Results and Correlations

As seen in figure 5, we found correlations between certain parameters alerting to certain attack types. When put through the REPTree algorithm, each parameter would return a baseline number. The baseline number acted as a system indicator. If that baseline changed, it tripped an alert for an attack. The correlations we found are as follows:

- Command_address - Baseline: 4, Changes Alert to DOS attack
- Response_address - Baseline: 0, Changes Alert to Recon attack



Parameter	Attack
CA	DOS(6)
RA	N/A
RL	N/A
CRF	DOS(6)
RRF	CMRI(2)
SF	MFCI(5)
SP	MPCI(4)
CM	MSCI(3)
CS	MSCI(3)
M	Outlier - N/A

Fig. 5. Results Table

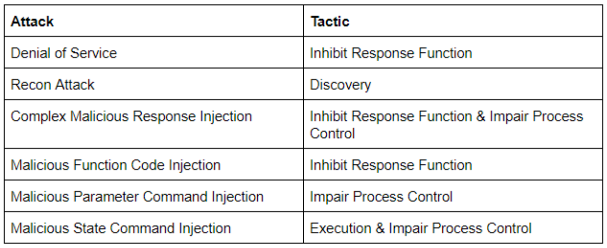
- Response_length - Baseline: 19, Changes Alert to Recon attack
- Comm_read_function - Baseline: 3, Changes Alert to DOS attack
- Resp_read_fun - Baseline: 1, Changes Alert to CMRI attack
- Subfunction - Baseline: 0, Changes Alert to MFCI attack
- Setpoint - Baseline: 20, Changes Alert to MPCI attack
- Control_mode - Baseline: 1, Changes Alert to MSCI attack
- Control Scheme - Baseline: 0, Changes Alert to MSCI attack
- Measurement - Outlier in dataset

The baselines are predefined in the system architecture and vary by dataset. The alerts generated by each parameter were then compared to the ATT&CK for ICS Framework Matrix for further attack classification. This comparison led to a more comprehensive incident response.

Certain common parameters such as Comm_read_function can bring forward a sense of a false positive. For example, an individual can commonly read into a DIO or AIO and not necessarily be a DOS attack. Regular computing systems have this trouble of identifying legitimate traffic and non legitimate traffic. As our research progresses and we continue to analyze new datasets, we are trying to identify correlations with legitimate and illegitimate read functions. Once we identify those types of read function calls more specifically, we can better apply the framework matrix to the intrusion attempt.

B. ATT&CK for ICS Framework Matrix Correlation

We were able to compare the resulting alerts generated by our detection system against the framework. The framework classified attacks and assigned them a tactic and an identification (ID). These tactics were then used to better understand the attacks and assist in assigning an alert level. When proceeding through the matrix with an attack in hand, the attack is correlated to the tactic. If the exact attack is not in the matrix, a generalization concept was used to achieve the same or similar technique. Some attacks correlated with multiple tactics, as they were used for several purposes. These attacks and tactics have been gathered from the ATT&CK for ICS knowledge base [20]. Refer to figure 6 to see the correlation found between the attacks in our data set and the framework matrix.



Attack	Tactic
Denial of Service	Inhibit Response Function
Recon Attack	Discovery
Complex Malicious Response Injection	Inhibit Response Function & Impair Process Control
Malicious Function Code Injection	Inhibit Response Function
Malicious Parameter Command Injection	Impair Process Control
Malicious State Command Injection	Execution & Impair Process Control

Fig. 6. ICS Framework Tactics

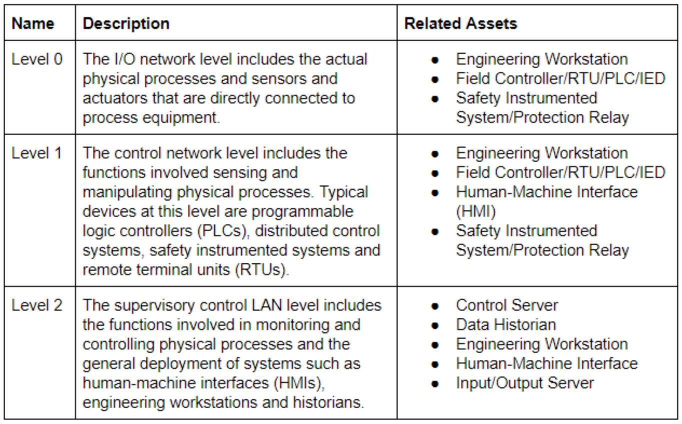
There are a total of 11 tactics in the ATT&CK for ICS framework. However, the dataset we implemented only alluded to 5 of the tactics. The description of our 5 tactics per the ATT&CK for ICS is below [20].

- Inhibit Response Function - The adversary is trying to prevent your safety, protection, quality assurance, and operator intervention functions from responding to a failure, hazard or unsafe state.
- Discovery - The adversary is trying to figure out your ICS environment.
- Impair Process Control - The adversary is trying to manipulate, disable, or damage physical control processes.
- Execution - The adversary is trying to run malicious code.

C. Levels in ATT&CK for ICS Framework

ATT&CK for ICS techniques implement the Purdue Model in order to classify which techniques are applicable to their environment. Not all of the levels in the Purdue Model are incorporated in the ATT&CK for ICS framework. This is due to the many software/hardware platforms, applications, and protocols in the ATT&CK for ICS environments. The currently present levels act as an aid for ATT&CK for ICS users to grasp the techniques that are applicable to their environment [20].

Those enterprise networks associated with levels 3 and 4 of the Purdue Model, could be used as a starting point for adversaries targeting ICS networks ATT&CK for Enterprise describes the tactics, techniques, and procedures (TTP) adversaries use to operate within these networks. Analyzing the TTPs of adversaries can better prepare a network for the types of attacks that they are susceptible to. In Level 2 of ICS networks, where specialized applications are run on Windows and Linux platforms, ATT&CK for Enterprise can describe TTPs of adversaries. There is great use in having the behavior of an adversary under surveillance. This allows for a platform to observe the behavior of it and act upon it adequately. It also serves as a learning technique to prevent similar attacks in the future. This point is considered the interface between the ATT&CK for Enterprise model and the ATT&CK for ICS model [20]. The three levels that are the focus of ATT&CK for ICS are shown in figure 7.



Name	Description	Related Assets
Level 0	The I/O network level includes the actual physical processes and sensors and actuators that are directly connected to process equipment.	• Engineering Workstation • Field Controller/RTU/PLC/IED • Safety Instrumented System/Protection Relay
Level 1	The control network level includes the functions involved sensing and manipulating physical processes. Typical devices at this level are programmable logic controllers (PLCs), distributed control systems, safety instrumented systems and remote terminal units (RTUs).	• Engineering Workstation • Field Controller/RTU/PLC/IED • Human-Machine Interface (HMI) • Safety Instrumented System/Protection Relay
Level 2	The supervisory control LAN level includes the functions involved in monitoring and controlling physical processes and the general deployment of systems such as human-machine interfaces (HMIs), engineering workstations and historians.	• Control Server • Data Historian • Engineering Workstation • Human-Machine Interface • Input/Output Server

Fig. 7. ICS Framework Levels

VII. Conclusion

The result of this research is the ability to successfully identify certain cyberattacks using a machine algorithm while applying the MITRE ATT&CK for ICS framework. Our dataset consisted of a gas pipeline industrial control system with many different parameters that had already suffered a cyberattack. The ICS framework offered a comprehensive matrix to compare our results to and offer insight on how to respond. The different levels in the framework also gave a good indication on how the severity of the incidents should be handled. Overall, with the combination of a ML algorithm and framework we were able to provide a responsive intrusion detection system for SCADA and ICS systems. We plan on continuing our research with expanding our datasets and development of a front facing software to detect these intrusions.

A. Implications

The most important takeaway from this research is that practicing automation for classifications processes would be an invaluable step forward within SCADA dependent industries. Our research showed us the current processes used for identifying and classifying intrusions within SCADA systems. The results of current systems are unfavorable when compared to our solution. This is because currently SCADA systems do not have any kind of automation of classification of intrusions, as per our client and mentor, Richard Alcalde. The lack of classification of intrusions delays the process of reacting to them adequately. It is comparable to going to war with an unknown intruder. If there is no information about the intruders tactics for attack, then the victim is at a large disadvantage. Automating these classifications would speed up the process tremendously, putting the user at an advantage at reacting to the attack.

These aforementioned implications must be mentioned alongside the study's parameters and limitations. As much as this research and the product of this research created a fantastic result, it is necessary to point out that there were certain aspects of our study that limit the effectiveness of the created product. The first of these is resources. In our team, the main resource we had was ourselves, four cybersecurity analysts. We ended up being an excellent random agglomeration of students, but it is important to note our nature. We are online students, with other classes and work to attend to; our time is limited to that which we have for the capstone course. This goes hand in hand with the limited number of data sets used. Our initial scope included testing three different machine learning algorithms, with three different data sets each. However, given our time frame and ease of access to appropriate data sets, we refined our scope to a single algorithm and a partial data set. Our study is efficient currently only for a gas pipeline industrial setting. It would have to be modified to work with other industries, according to their specific regulations.

B. Recommendations and Potential Next Steps

There are a few items to address for further development. These are as follows:

1) Continued development of the Java program to enhance end user experience and support multiple different dataset types and industrial control systems. There are several options for this.
   a) The program can be given a graphic user interface (GUI) system. The majority of successfully deployed applications have an sleek, interactive GUI that makes user experience (UX) favorable. A prototype for this classification program would likely include a simple input-output GUI, where the user inputs their information in one section, and receives the results on the side.
   b) To ease access for the user, the program can be made available as a web-based application. Many applications are made available through the internet. Given the relatively simple input-output nature of the classification program, it might be easy to create a web-based application that would collect user data and provide results based on the classification of this data. One possible effect of this is information can be collected remotely and continuously; the user would only have to reach their specific user login to access their classifications.
   c) The program can either have a cloud-based database or a host-based database. This depends

on the scalability of the program. If the number of users for the java program are low, it may be easier to simply keep it host-based, and periodically update the program on their devices. However, if the number of users for the program are higher and the database will be used on a more world-wide basis, it may be preferable to keep the program cloud based given today's IoT cloud based society.

2) Further research into training the algorithm to more accurately detect false-positives in DOS and Recon attacks. For example, a legitimate user repeatedly inputting a read function could be read as a false-positive, even though the user may just be running reporting or looking at multiple different DIO/AIO. Training the program through larger datasets may decrease the possibility of false-positives.
3) Since the focus of this research was on gas pipeline control systems, another step would be to expand to water, electric, and sewage control systems. These are all critical structures that should have an automated intrusion detection system that incorporates the ATT&CK for ICS framework. The program would have to be either copied and individually trained to fit each industry, or it would have to be broken down and given branches to apply for each industry.
4) The last step required for furthering the study would be to either increase resources available to the team or outsource some of the research. Each team member works around specific work schedules, family schedules, and now COVID-19 [23] schedules. It might be necessary to outsource some of the work in order to get this program developed quickly.

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/SustainableSec 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/

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

[21] E. H. GICSP, M. Assante, and T. Conway, "An abbreviated history of automation & industrial controls systems and cybersecurity," 2014.

[22] "Mitre releases framework for cyber attacks on industrial control systems," Jan 2020. [Online]. Available: https://www.mitre.org/news/press-releases/mitre-releases-framework-for-cyber-attacks-on-industrial-control-systems

[23] C. for Disease and C. Protection, "Coronavirus (covid-19)," 2020.

[24] D. Shewan, "10 companies using machine learning in cool ways," Aug 2019. [Online]. Available: https://www.wordstream.com/blog/ws/2017/07/28/machine-learning-applications

[25] "Description of reptree result," Jan 2015. [Online]. Available: https://weka.8497.n7.nabble.com/Description-of-REPTree-result-td33451.html