Received 1-05-2024; Accepted 22-06-2024; Published 29-06-2024

Innovative Machine Learning Model Design for Predictive IoT Security Attacks

Atdhe Buja atdhe.buja@academyict.net
ICT Academy, Prishtina, Kosovo.

Melinda Pacolli pacollimelinda@gmail.com
ECPD, Prishtina Kosovo.

Donika Bajrami donika.bajrami@academyict.net
ICT Academy, Prishtina, Kosovo.

Philip Polstra ppolstra@commonwealthu.edu
Bloomsburg University of Pennsylvania, PA, USA.

Akihiko Mutoh mutoh@tsukijihongwanji.jp
Tsukijihongwanji, Tokyo, Japan.

Corresponding Author: Atdhe Buja

Abstract

The research focuses on designing a predictive model for Internet of Things (IoT) attack identification using historical IoT data from the Global Cyber Alliance's (GCA) Automated IoT Defense Ecosystem (AIDE). This research goes into the design of an enhanced machine-learning model, to predict potential security breaches. The process involved a thorough data science lifecycle, overall data preprocessing, feature selection, and engineering. The study's main objective is to design a model to classify IoT activities and events, distinctive among normal operations and indicators of potential cyber-attacks. The model design incorporates distinct features like command frequency, login success, geo-distance calculations, credentials tried, and protocol encodings to enhance predictive accuracy. The model uses algorithms like logistic regression and random forests to explore their efficacy in binary and multiclass classifications. The research emphasizes the critical role of the model's capability to proactively address IoT security challenges. Offering early alerts is crucial in allowing timely countermeasures, herein strengthening IoT ecosystems against cyber threats. The model's accuracy in predicting IoT attacks, reduces the likelihood of successful breaches, thus safeguarding sensitive data and infrastructure. Furthermore, it assists as a key tool for policymakers and security professionals, providing insight into occurring threat patterns and enabling the development of robust defense strategies.

Keywords: IoT Security, Predictive model, Machine learning, Data science, Cyber-attacks.

Citation: Atdhe Buja, et al. Innovative Machine Learning Model Design for Predictive IoT Security Attacks Advances in Artificial Intelligence and Machine Learning. 2024;4(2):139.

1. INTRODUCTION

In the latest times, the increase of Internet of Things (IoT) devices has transformed modern life, from smart home automation to industrial process optimization [1, 2]. Still, along with the enhancements, the IoT view has become very vulnerable to cyber threats, presenting meaningful challenges to cybersecurity practitioners and researchers [3, 4]. The integration nature of IoT devices, followed by their limit to computational resources and distinct communication protocols, introduces rare vulnerabilities that can be exploited by malicious actors. Consequently, the need for robust cybersecurity countermeasures customized to the IoT ecosystem is demanding.

This study addresses the issue of IoT security utilizing a thorough analysis of attack tactics identification using data obtained from the GCA Automated IoT Defense Ecosystem (AIDE) [5]. By benefiting enhanced data science techniques and machine learning (ML) algorithms, our study focuses on disclosing patterns and recurring strategies used by attackers to obtain unauthorized access and compromise IoT systems. Across a systematic investigation of main features such as command frequency, login success, geo-distance calculations, and protocol encodings, we look to provide important insights into the evolving threat landscape linked to IoT devices.

The intention of this research study consists of its potential to notify the design of proactive defense mechanisms and countermeasures customized to mitigate IoT security threats. By comprehending the tactics and techniques used by malicious actors, cybersecurity practitioners can better expect and respond to cyber-attacks, in that way protecting critical IoT infrastructure and maintaining the integrity of connected systems. Also, our findings have an impact beyond the field of cybersecurity and broaden to areas such as policy-making, regulatory compliance, and the future design of IoT devices.

In the following sections, we introduce a thorough methodology defining our approach to data collection, preprocessing, feature selection, and model design [6, 7]. Afterward, we deep into the analysis of attack tactics identification and [8], explain the insights obtained from our research. At last, we discuss the impact of our findings and offer recommendations for advancing IoT security posture. Regardless of the increasing detection of IoT security risks, there is a gap in the capacity to proactively identify and mitigate potential attacks before they happen. In response to this gap, this paper focuses on designing a predictive model capable of identifying patterns, trends, and anomalies in historical IoT attack data.

2. METHODOLOGY

The methodology employed in this study adhered to solid scientific principles to ensure the validity and reliability of the research findings [9, 10]. The methodology embraces a thorough data science lifecycle and cybersecurity, including data preprocessing, feature selection, and engineering. Using the GCA Automated IoT Defense Ecosystem (AIDE) [5], dataset, the research identifies the main features applicable to IoT security, such as command frequency, login success, geo-distance calculations, credentials tried, and protocol encodings. These features are carefully selected and engineered to enhance the model's predictive accuracy. The model design concerns the investigation of several Machine Learning (ML) algorithms, as well as logistic regression and random forests. The referred algorithms are assessed for their efficacy in binary and multiclass classifications, taking into account the complexity of IoT security threats. Below is a high-level overview of the steps carried off.

2.1 Data Acquisition

The first step of the methodology involved the collection of data from an operational honey farm by found collaboration with GCA [5], serves as a centralized repository of IoT attack data. Python was utilized to efficiently obtain data in chunks from the AIDE, which contained a significant dataset (54,835,849 records) for a certain period (1st May 2023 – 31st July 2023). Every record in the dataset holds information on specific events related to IoT attacks, as well as timestamps, unknown commands, urls, hashes, version, commands, credentials, and loggedin. This way includes retrieving datasets in chunks sequentially and processing them iteratively. Python's versatility and scalability allowed the development of custom scripts to automate the retrieval process, ensuring optimal performance and resource utilization.

2.2 Data Preprocessing Techniques

Previously the data is acquired, it goes through data preprocessing to ensure the quality and suitability of the dataset and prepare it for analysis, and further model design. This involved several steps, including data cleaning, feature scaling, feature encoding, and data balancing. By applying these data preprocessing techniques, we were able to ease potential biases and ensure the robustness of the following analysis, and model design.

2.3 Feature Selection and Engineering

A vital step of the methodology involved the selection and engineering of relevant features for analysis and further model design. Based on the complexity of IoT attack data, it was necessary to identify relevant features that could provide important insights into attack tactics identification, and distribution. Utilization of statistical analysis, and domain expertise, we discovered a subset of features that were considered most instructional for obtaining the research objectives. Custom and many Python libraries were utilized to simplify the analysis process. Custom Python scripts were written to automate repetitive tasks and refine the whole analysis. Also, visualization libraries such as Pandas, Matplotlib, and Seaborn were engaged to generate perceptive visualizations, confirming the interpretation and communication of analysis results. In general, the methodology utilized in this study fences data acquisition, comprehensive data preprocessing, thoughtful feature selection and engineering, and productive data analysis using Python scripts and libraries. By overseeing feature selection and engineering, the dataset was improved to include informative features that are related to predicting IoT security attacks.

3. ATTACK TACTICS IDENTIFICATION ANALYSIS OF AIDE IOT ATTACK DATA

In this section, we involve attack tactics identification analysis methods to inspect the rich dataset obtained from the GCA AIDE. The analysis of attack tactics identification was a vital aspect of this research, directed at unraveling patterns and recurring strategies used by attackers within the GCA AIDE dataset. Our objective is to uncover attack tactics identification or recurring patterns in IoT attacks, using key features such as timestamps (@timestamp), unknown commands, URLs, hashes, version, commands, and login credentials (loggedin, and credentials).

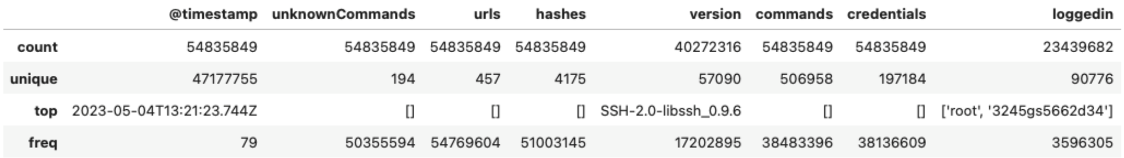
The analysis is led by the research question:

- What are the common attack tactics used by attackers?

In our study, we engaged several attack tactics identification analysis techniques (pattern recognition, descriptive analysis, and combination analysis) to gain insights into the IoT attack tactics or recurring patterns. We use a methodology based on attack tactics identification analysis techniques, including pattern recognition, descriptive analysis, and combination analysis.

3.1 Pattern Recognition

Pattern recognition techniques were engaged to find characteristic patterns suggestive of distinct attack tactics within the GCA AIDE dataset. By examining a series of events and their temporal relationships, common attack behavior patterns were discovered. This intricates discovering a series of commands, URLs, or login attempts that often lead to or companion particular types of attacks. The analysis engaged pattern recognition techniques to reveal characteristic patterns within the GCA AIDE IoT attack tracking dataset. Particularly, by the inspection of commands executed, URLs accessed, unknown commands, and related hashes, diverse patterns reflective of many attack tactics were discovered. We examine the first few rows of the dataset which provides insights into several fields vital for understanding the nature of the attack data to pinpoint patterns and attack tactics. Next, we use the method (data.describe) to result in summary statistics for the numerical columns, offering insights into the dataset distribution shown in Figure 1. The dataset has 54,835,849 records, each obtaining specific events related to IoT attacks. Main fields (‘@timestamps’, ‘unknown commands’, ‘URLs’, ‘hashes’, ‘version’, ‘commands’, ‘credentials’, and ‘loggedin’) provide vital insights into attack timestamps, commands executed, url and hashes of malware files, library version, and login attempts. Notice, that the timestamps (54,835,849) field shows the temporal context of a total count with the most frequent timestamp (’2023-05-04T13:21:23.744Z’ freq 79), indicating attention of activity at that time. Then, the unknown commands (freq 50,355,594) field exhibits empty brackets ’[]’, meaning a lack of fully emulated commands, probably indicating reconnaissance or failed attempts. Also, the urls (freq 54,769,604) show sources from where malware files were likely downloaded. Unusually, the most frequent record is an empty bracket ’[]’, indicating that most of the records lack associated URLs. Similarly, the hashes (freq 51,003,145) field means a lack of associated hashes in most cases. So, the version (40,272,316) field appears, with the very frequent version recorded (’SSH-2.0-libssh_0.9.6’ by freq 17,202,895), indicating widespread exploit of this specific version by attackers. The commands (freq 38,483,396) field like the unknown commands field shows a lack of fully emulated commands and empty brackets ’[]’. Following, the login attempts (‘credentials’ freq 38,136,609) indicating a high frequency of failed login attempts, and (‘loggedin’, ’root’, ’3245gs5662d34’ freq 3,596,305) show insights into compromised credentials used by attackers, and report successful login attempts (23,439,682). The analysis uncovers a meaningful part of the dataset described by empty fields, indicating a widespread presence of reconnaissance activities or failed attempts. The frequent exploit of specific SSH library versions and common username/password combinations emphasizes possible targets and vulnerabilities exploited by attackers. The lack of URLs and hashes in the most of records indicates a need for enhanced detection and logging mechanisms to obtain thorough attack data. Insights from this dataset can advise the design of proactive defense plans, including advanced monitoring of SSH activity and reinforcing authentication mechanisms to mitigate unauthorized access attempts.



	@timestamp	unknownCommands	urls	hashes	version	commands	credentials	loggedin
count	54835849	54835849	54835849	54835849	40272316	54835849	54835849	23439682
unique	47177755	194	457	4175	57090	506958	197184	90776
top	2023-05-04T13:21:23.744Z	[]	[]	[]	SSH-2.0-libssh_0.9.6	[]	[]	['root', '3245gs5662d34']
freq	79	50355594	54769604	51003145	17202895	38483396	38136609	3596305

Figure 1: Summary statistics of the dataset distribution.

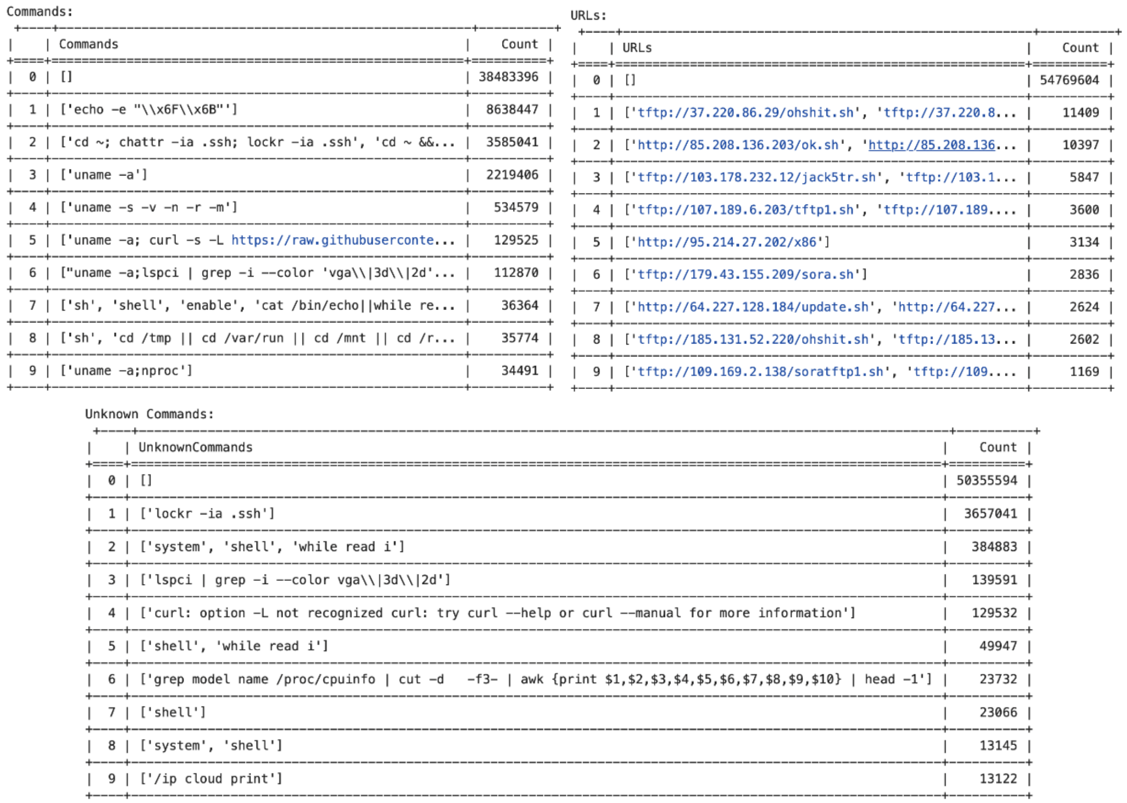
3.2 Descriptive Analysis

Descriptive analysis supplied insights into the distribution and characteristics of main features within the dataset. Visualization techniques were applied to inspect the distribution of features over time and detect any anomalies or flaws. In this descriptive analysis, the frequency of every attack tactic is measured to realize the recurring patterns or tactics utilized by attackers. Figure 2, presents a comprehensive distribution and frequency of each attack tactic or attack vector (commands, urls, unknown commands, and hashes). Further, provide the findings for every field:

Commands: The most of records (38,483,396) do not define any commands executed. The use of the command echo (-e ”\\x6F\\x6B” executed 8,638,447 times) means a high frequency of this command, outputs the string ”ok” in a Unix-like environment, certainly means of automated processes or bot activity. Other outstanding command sequences (cd ~; chattr -ia .ssh; lockr -ia .ssh executed 3,585,041 times) indicate tries to manipulate SSH configurations or keys. Additional commands (uname -a and uname -s -v -n -r -m) were applied frequently, clearly meaning reconnaissance or exploitation activities.

URLs: Most records (54,769,604) do not have related URLs. Though, some URLs were accessed frequently, (tftp://37.220.86.29/ohshit.sh 11,409 times and http://85.208.136.203/ok.sh 10,397 times). These URLs may serve as command-and-control (C2) servers [11, 12], or sources of malware. As we know from practical experience C2 servers usually send commands to jeopardized devices or systems to execute malicious actions from the urls provided in the first place.

Unknown Commands: Almost all records (50,355,594) do not have any unknown commands. Noteworthy unknown commands (lockr -ia .ssh and command sequences ‘system’, ‘shell’, and



```
Commands:
+----+-------------------------------------------------------+----------+
|    | Commands                                              |    Count |
+====+=======================================================+==========+
|  0 | []                                                    | 38483396 |
+----+-------------------------------------------------------+----------+
|  1 | ['echo -e "\\x6F\\x6B"']                              |  8638447 |
+----+-------------------------------------------------------+----------+
|  2 | ['cd ~; chattr -ia .ssh; lockr -ia .ssh', 'cd ~ &&... |  3585041 |
+----+-------------------------------------------------------+----------+
|  3 | ['uname -a']                                          |  2219406 |
+----+-------------------------------------------------------+----------+
|  4 | ['uname -s -v -n -r -m']                              |   534579 |
+----+-------------------------------------------------------+----------+
|  5 | ['uname -a; curl -s -L https://raw.githubuserconte... |   129525 |
+----+-------------------------------------------------------+----------+
|  6 | ["uname -a;lspci | grep -i --color 'vga\\|3d\\|2d'... |   112870 |
+----+-------------------------------------------------------+----------+
|  7 | ['sh', 'shell', 'enable', 'cat /bin/echo||while re... |    36364 |
+----+-------------------------------------------------------+----------+
|  8 | ['sh', 'cd /tmp || cd /var/run || cd /mnt || cd /r... |    35774 |
+----+-------------------------------------------------------+----------+
|  9 | ['uname -a;nproc']                                    |    34491 |
+----+-------------------------------------------------------+----------+
```

```
URLs:
+----+-------------------------------------------------------+----------+
|    | URLs                                                  |    Count |
+====+=======================================================+==========+
|  0 | []                                                    | 54769604 |
+----+-------------------------------------------------------+----------+
|  1 | ['tftp://37.220.86.29/ohshit.sh', 'tftp://37.220.8... |    11409 |
+----+-------------------------------------------------------+----------+
|  2 | ['http://85.208.136.203/ok.sh', 'http://85.208.136... |    10397 |
+----+-------------------------------------------------------+----------+
|  3 | ['tftp://103.178.232.12/jack5tr.sh', 'tftp://103.1... |     5847 |
+----+-------------------------------------------------------+----------+
|  4 | ['tftp://107.189.6.203/tftp1.sh', 'tftp://107.189.... |     3600 |
+----+-------------------------------------------------------+----------+
|  5 | ['http://95.214.27.202/x86']                          |     3134 |
+----+-------------------------------------------------------+----------+
|  6 | ['tftp://179.43.155.209/sora.sh']                     |     2836 |
+----+-------------------------------------------------------+----------+
|  7 | ['http://64.227.128.184/update.sh', 'http://64.227... |     2624 |
+----+-------------------------------------------------------+----------+
|  8 | ['tftp://185.131.52.220/ohshit.sh', 'tftp://185.13... |     2602 |
+----+-------------------------------------------------------+----------+
|  9 | ['tftp://109.169.2.138/soratftp1.sh', 'tftp://109.... |     1169 |
+----+-------------------------------------------------------+----------+
```

```
Unknown Commands:
+----+----------------------------------------------------------------------------------------------------------+----------+
|    | UnknownCommands                                                                                          |    Count |
+====+==========================================================================================================+==========+
|  0 | []                                                                                                       | 50355594 |
+----+----------------------------------------------------------------------------------------------------------+----------+
|  1 | ['lockr -ia .ssh']                                                                                       |  3657041 |
+----+----------------------------------------------------------------------------------------------------------+----------+
|  2 | ['system', 'shell', 'while read i']                                                                      |   384883 |
+----+----------------------------------------------------------------------------------------------------------+----------+
|  3 | ['lspci | grep -i --color vga\\|3d\\|2d']                                                                |   139591 |
+----+----------------------------------------------------------------------------------------------------------+----------+
|  4 | ['curl: option -L not recognized curl: try curl --help or curl --manual for more information']           |   129532 |
+----+----------------------------------------------------------------------------------------------------------+----------+
|  5 | ['shell', 'while read i']                                                                                |    49947 |
+----+----------------------------------------------------------------------------------------------------------+----------+
|  6 | ['grep model name /proc/cpuinfo | cut -d   -f3- | awk {print $1,$2,$3,$4,$5,$6,$7,$8,$9,$10} | head -1'] |    23732 |
+----+----------------------------------------------------------------------------------------------------------+----------+
|  7 | ['shell']                                                                                                |    23066 |
+----+----------------------------------------------------------------------------------------------------------+----------+
|  8 | ['system', 'shell']                                                                                      |    13145 |
+----+----------------------------------------------------------------------------------------------------------+----------+
|  9 | ['/ip cloud print']                                                                                      |    13122 |
+----+----------------------------------------------------------------------------------------------------------+----------+
```

Figure 2: Distribution and frequency of each attack tactic and attack vector.

'while read i') were executed repeatedly, meaning tried to exploit system vulnerabilities or execute unemulated commands.

Hashes: Most of the records (51,003,145) do not have related hashes. One hash, such (a8460f446be540410004b1a8db4083773fa46f7fe76fa84219c93daa1669f8f2), was found (3,649,179), indicating the frequent use of specific malware files or payloads.

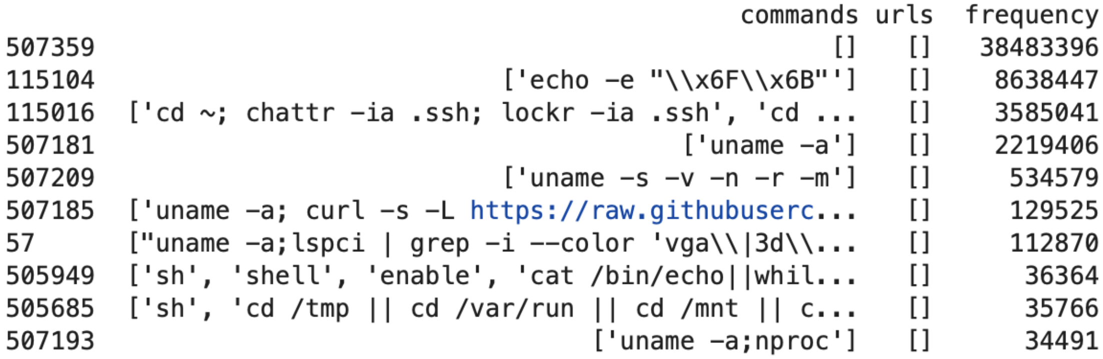
The descriptive analysis uncovers common attack tactics, as well as specific commands executed, accessed URLs, unknown command tries, and common malware hashes used. Some commands and URLs show with high frequency, signifying their meaning in the attacker's tactics. Unknown commands and specific hashes simply try to exploit system vulnerabilities or deploy a certain malware. The findings give insights into attackers' manner and can guide defense strategies and beyond examinations. Knowing these patterns is vital for developing effective defense mechanisms and proactive countermeasures. The results of our descriptive analysis illuminate the common patterns that the attackers have used, and this answers our research question. It emphasizes the importance of understanding the attacker's playbook, allowing us to expect and counter threats proactively.

3.3 Combination Analysis

Combination analysis involves the investigation of relations among various features to uncover hidden relationships and attack patterns. This analysis aimed to identify intricate attack tactics that may involve collective stages or components. The thorough analysis of attack patterns uncovered critical insights into common IoT attack tactics engaged by attackers or recurring patterns. In this analysis patterns, and combinations of fields (commands, URLs, unknown commands, and hashes) are inspected to pinpoint correlations and recurring behaviors through attackers.

Commands corresponding with URLs:

The analysis uncovers that some commands are executed without any related URLs, meaning self-standing activities. Figure 3 provides insights, commands (echo -e ”\\x6F\\x6B” and order starting with cd ~; chattr -ia .ssh; lockr -ia .ssh...) are often executed without accessing URLs, meaning possible reconnaissance or system manipulation tactics.



```
                                                    commands urls  frequency
 507359                                                   []   []   38483396
 115104                             ['echo -e "\\x6F\\x6B"']   []    8638447
 115016  ['cd ~; chattr -ia .ssh; lockr -ia .ssh', 'cd ...   []    3585041
 507181                                        ['uname -a']   []    2219406
 507209                             ['uname -s -v -n -r -m']   []     534579
 507185  ['uname -a; curl -s -L https://raw.githubuserc...   []     129525
 57      ["uname -a;lspci | grep -i --color 'vga\\|3d\\...   []     112870
 505949  ['sh', 'shell', 'enable', 'cat /bin/echo||whil...   []      36364
 505685  ['sh', 'cd /tmp || cd /var/run || cd /mnt || c...   []      35766
 507193                                  ['uname -a;nproc']   []      34491
```

Figure 3: Group by ’commands’ and ’urls’ and count the frequency

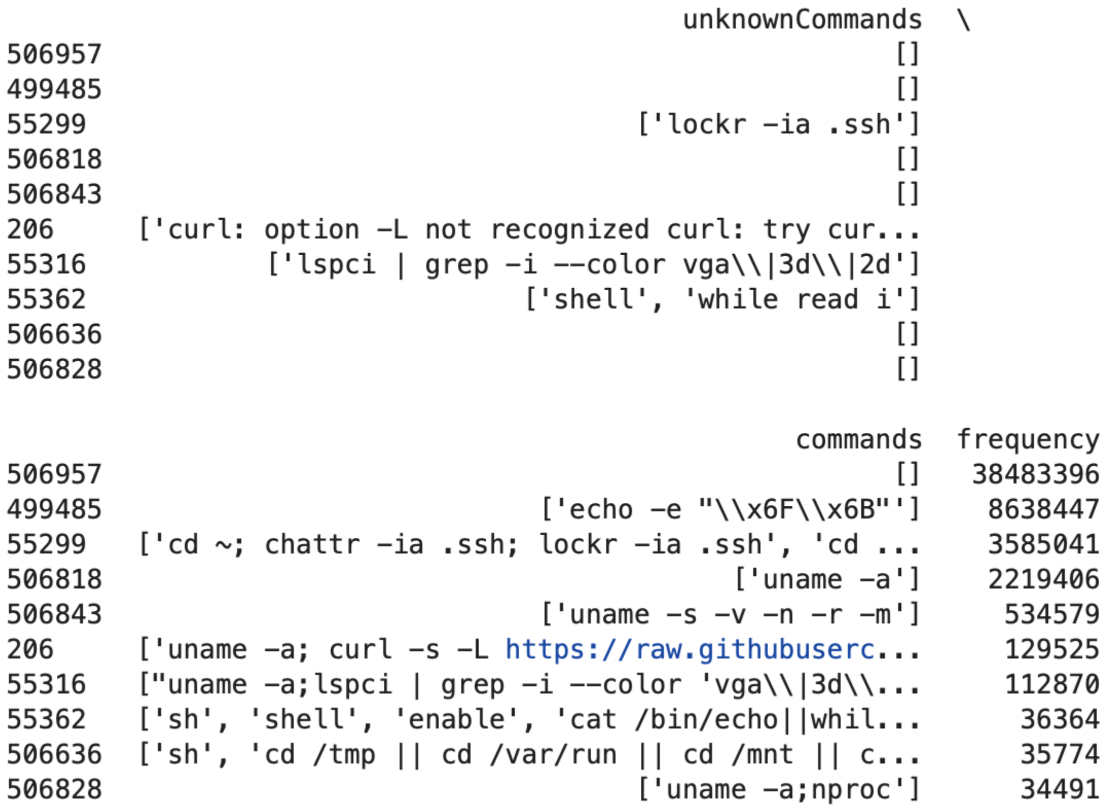
Unknown Commands succeeded by Known Commands:

Certain unknown commands, (lockr -ia .ssh) shown in Figure 4, are often succeeded by known command sequences, meaning a pattern where attackers try a set of actions before executing extensive commands. This pattern hints at a structured approach to exploit vulnerabilities or gain potential unauthorized access.

Hashes corresponding with Commands:

Especial hashes are often related to specific commands, meaning potential malware files or payloads used in combination with certain actions. Figure 5 provides detailed insights into the hash (a8460f446be540410004b1a8db4083773fa46f7fe76fa84219c93daa1669f8f2) is often related to the order (cd ~; chattr -ia .ssh; lockr -ia .ssh...) indicating a coherent distribution of malware after modifying SSH configurations.

The combination analysis results provide detailed insight into the attacker’s tactics and potential relations. Additionally, we understand that attackers repeatedly execute particular commands solely, indicating reconnaissance steps or system manipulation tactics. Certain unknown commands lead to known commands, meaning a systematic testing of tactics or exploiting vulnerabilities. Some files



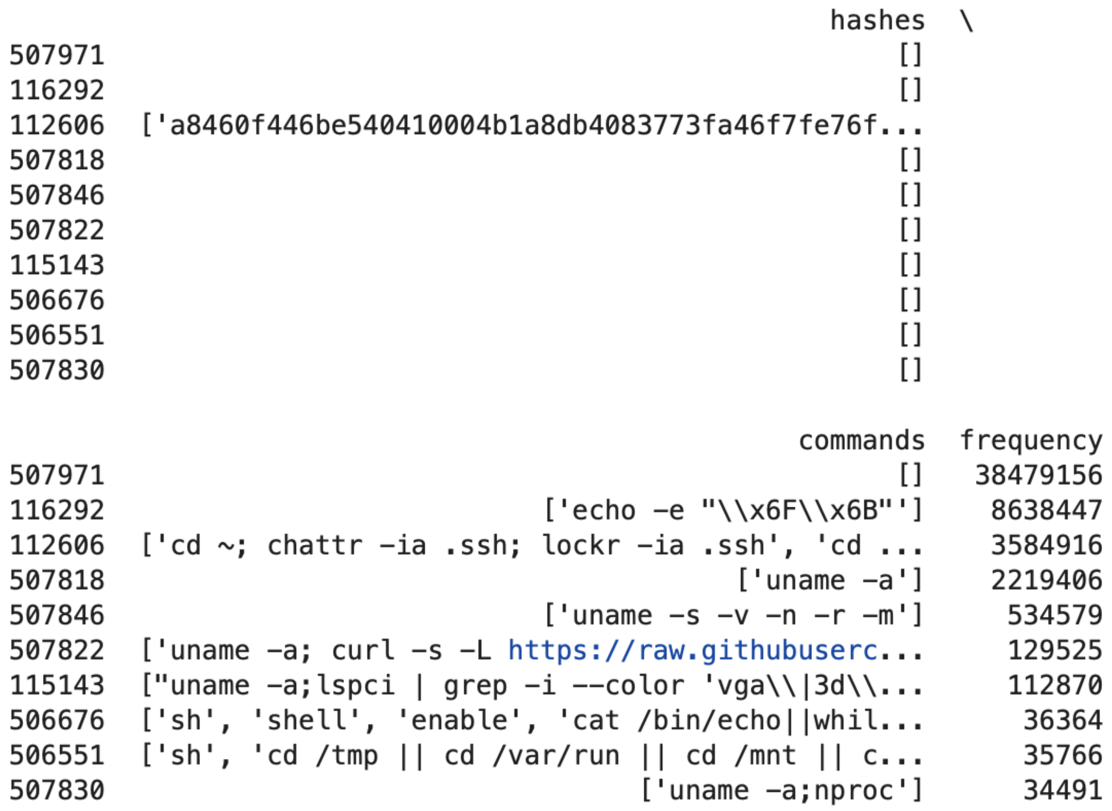
```
                                          unknownCommands  \
506957                                                 []   
499485                                                 []   
55299                                ['lockr -ia .ssh']   
506818                                                 []   
506843                                                 []   
206     ['curl: option -L not recognized curl: try cur...   
55316           ['lspci | grep -i --color vga\\|3d\\|2d']   
55362                           ['shell', 'while read i']   
506636                                                 []   
506828                                                 []   

                                                 commands  frequency
506957                                                 []   38483396
499485                           ['echo -e "\\x6F\\x6B"']    8638447
55299   ['cd ~; chattr -ia .ssh; lockr -ia .ssh', 'cd ...    3585041
506818                                       ['uname -a']    2219406
506843                           ['uname -s -v -n -r -m']     534579
206     ['uname -a; curl -s -L https://raw.githubuserc...     129525
55316   ["uname -a;lspci | grep -i --color 'vga\\|3d\\...     112870
55362   ['sh', 'shell', 'enable', 'cat /bin/echo||whil...      36364
506636  ['sh', 'cd /tmp || cd /var/run || cd /mnt || c...      35774
506828                                 ['uname -a;nproc']      34491
```

Figure 4: Group by ’unknownCommands’ and ’commands’ and count the frequency

(appeared by hashes) are often related to specific commands, indicating that these files may be key components of specific attack patterns. Cross-referencing the hashes with any threat intelligence databases might give more insights into the nature of the files. Realizing the limits of this research to have access to any threat intelligence solution, is a lack of provided detailed data on hashes.

Regarding the recommendations that emerge from these analysis results, they will serve security teams, engineers, and data analysts to enhance the IoT security posture. Security teams must monitor the most often commands, URLs, and hashes as they are characteristics of common attack tactics. Real-time monitoring for these often patterns can guide early threat detection approaches. Moreover, cross-referencing hashes with threat intelligence supply additional insights into the nature of related files and their maliciousness which were captured. The combination analysis stresses key insights into attacker manners, identifying the value of understanding patterns (command-URL relations) and the use of unknown commands. These findings reinforce the need for proactive defense countermeasures, continuous monitoring, and resilient threat detection to successfully mitigate advanced cyber threats.



```
                                                     hashes  \
507971                                                   []
116292                                                   []
112606  ['a8460f446be540410004b1a8db4083773fa46f7fe76f...
507818                                                   []
507846                                                   []
507822                                                   []
115143                                                   []
506676                                                   []
506551                                                   []
507830                                                   []

                                                   commands  frequency
507971                                                   []   38479156
116292                             ['echo -e "\\x6F\\x6B"']    8638447
112606  ['cd ~; chattr -ia .ssh; lockr -ia .ssh', 'cd ...    3584916
507818                                        ['uname -a']    2219406
507846                             ['uname -s -v -n -r -m']     534579
507822  ['uname -a; curl -s -L https://raw.githubuserc...     129525
115143  ["uname -a;lspci | grep -i --color 'vga\\|3d\\...     112870
506676  ['sh', 'shell', 'enable', 'cat /bin/echo||whil...      36364
506551  ['sh', 'cd /tmp || cd /var/run || cd /mnt || c...      35766
507830                                   ['uname -a;nproc']      34491
```

Figure 5: Group by ‘hashes’ and ‘commands’ and count the frequency

4. PRELIMINARY MODEL DESIGN

In this section, we introduce the preliminary model design, based on the logical steps undertaken to conclude the model from the results of the analysis conducted. Performing analysis (Exploratory Data Analysis EDA) [13, 14], which includes the attack tactics identification and time-series analysis of AIDE IoT attack data [8], helps to understand the characteristics of the dataset. The design includes activities of data preprocessing, feature selection, and suitable algorithms for classification.

Data preprocessing - the foremost step is intricate meticulous data preprocessing to assure the quality of the dataset for model design by eliminating features that could cause overfitting. This was carried out by addressing missing values, outliers, and errors to advance data integrity. Imputation techniques addressed those, then, fields (’endTime’ and ’startTime’) were converted to an acceptable format (datetime) to support model design. Also, checks were performed for any missing values to identify and rectify potential data discrepancies. Finally, the processed dataset was saved to a CSV file.

Feature selection - is headed to identify the most appropriate features for the model [15], based on the analysis results and the structure of the dataset we have. Features were selected to obtain the foremost aspects of IoT attack behavior. These features consist of:

1. **Duration of attack**: Calculate the duration of each attack and employ fields ('startTime' and 'endTime') to integrate temporal information into the model.
2. **Successful login indicator**: Generate a binary feature-based ('loggedin') field to discern between successful and unsuccessful login attempts, having insights into authentication-based attack patterns.
3. **Geo-distance**: Calculate the geographical distance between the attacker and the target employing the latitude and longitude ('geoip' and 'hostGeoip'), allowing the taking in of spatial context.
4. **Command frequency**: Count the number of unique commands utilized in each session to obtain the frequency of command usage as a possible indicator of attack behavior.
5. **Credentials attempt**: Counting the number of unique ('credentials') attempts in each session to assess the wideness of credential-based attacks.
6. **Protocol encoding**: Converting ('protocol') ssh or telnet into a binary encoded format to describe protocol usage.

Algorithm selection - further, the choice of algorithms for the model was grounded on their usefulness for analyzing IoT attack data and their potential to supply insights into attack tactics and patterns. Suggested algorithms logistic regression and random forests were considered for their ability to handle binary and multiclass classifications appropriate for distinguishing between attack and non-attack events. In general, the preliminary model design focuses on establishing a basis for future research work on model development, using insights from the analysis to produce a robust framework for comprehending and mitigating IoT security threats.

5. DISCUSSION

The study employed a thorough methodology to assure the credibility and solidity of the research findings, using a robust data science lifecycle and cybersecurity principles. Anyway, there is a gap in the capacity to proactively identify and mitigate potential attacks before they happen. In response to this gap, this paper aims to design a predictive model capable of identifying patterns, trends, and anomalies in historical IoT attack data. The analysis includes pattern recognition, descriptive analysis, and combination analysis techniques to gather insights into the tactics used by attackers.

Pattern recognition techniques [16, 17], uncovered characteristic patterns indicative of diverse attack tactics within the dataset. The analysis discovered common commands executed, URLs accessed, unknown commands attempted, and hashes related to malware files. These findings illuminate the tactics used by attackers, such as reconnaissance steps, system manipulation tries, and exploitation of vulnerabilities. Then, the descriptive analysis [18, 19], offers insights into the distribution and frequency of attack vectors, stressing the meaning of certain commands, URLs, unknown commands, and hashes in attackers' strategies. The combination analysis exposed complex relationships through various features, identifying correlations and recurring behaviors shown by attackers [20, 21]. Refer to, commands that were executed without accessing URLs, indicating self-standing activities or introductory reconnaissance. Furthermore, certain unknown commands were often succeeded by known command sets, meaning a structured approach to exploiting vulnerabilities or gaining unauthorized access. files (appeared by hashes) were often related to specific commands, indicating vital components of attack patterns.

The insights harvested from these analyses have meaningful inferences for advancing IoT security posture and designing proactive defense strategies. Security teams can benefit from these findings by monitoring constantly commands, URLs, and hashes, facilitating early detection of potential threats. Also, cross-referencing hashes with any threat intelligence databases can supply further insights into the nature of files and their maliciousness.

Lastly, this study contributes precious insights into attacker behaviors and tactics targeting IoT devices, highlighting the significance of understanding attack patterns and engaging proactive defense countermeasures. By benefiting from enhanced analytics techniques and thorough datasets, organizations can advance their cybersecurity posture and mitigate evolving cyber threats. Progressing forward, continuous monitoring, threat intelligence consolidation, and resilient defense mechanisms will be vital in the protection of IoT ecosystems against advanced cyber-attacks.

6. CONCLUSION

This research seeks to bond the gap in IoT security by designing a predictive model able to proactively identify potential attacks using historical IoT data from the Global Cyber Alliance's Automated IoT Defense Ecosystem (AIDE). By using enhanced data science techniques and machine learning (ML) algorithms, we have proposed a design of an advanced model that displays likely capabilities in enlightened patterns, trends, and anomalies characteristic of cyber-attacks on IoT infrastructure. Our model design integrates critical features such as command frequency, login success, geo-distance calculations, credentials attempted, and protocol encodings, all focused on enhancing predictive accuracy. By posing algorithms like logistic regression and random forests, we have shown the efficacy of binary and multiclass classifications in identifying and grouping IoT activities and events.

The intent of our research is to its potential to feed with early alerts and proactive defense mechanisms, important for protecting IoT data and infrastructure against advanced cyber threats. By equipping stakeholders with the tools to predict and counteract emerging threats, our model works as a key asset for cybersecurity practitioners, IoT device manufacturers, policymakers, and regulatory bodies. Additionally, our findings impact the informing policy-making, regulatory compliance, and the future design of IoT devices. Noted, the limitation of the research is the lack of access to threat intelligence solutions, which could have supplied deeper insights into the nature of the discovered malware files and their related risks. Future research work will focus on refining the model design and advancing Machine Learning (ML) and Artificial Intelligence (AI) model development, training, testing, and evaluation processes. Further exploration in this line will cover the way for the enhancements of a more robust machine learning (ML) model for predicting or identifying IoT attacks. Additionally, it could be possible to create the model into a user-friendly interface, such as a finite dashboard, to provide stakeholders with easy access to key metrics. Potentially, this model can be implemented within the industrial infrastructure in various directions from information technology technical aspect (integrated within an intrusion detection as a rule, agentless model, etc.).

In summary, our research highlights the demanding role of predictive modeling in sustaining the resilience and cybersecurity of IoT ecosystems. By benefiting from data-driven insights and robust cyber defense strategies, we can efficiently mitigate the risks put forth by cyber threats and ensure the maintenance of integrity and operation of IoT infrastructure.

7. ACKNOWLEDGMENT

We would like to thank Global Cyber Alliance (GCA) https://globalcyberalliance.org/ for sharing the data with us. The ICT Academy https://www.academyict.net supported the work under the Research & Innovation Department in partnership with GCA.

References

[1] Buja A, Apostolova M, Luma A, Januzaj Y. Cyber Security Standards for the Industrial Internet of Things (Iiot)– A Systematic Review. International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA). 2022:1-16.

[2] Rajmohan T, Nguyen PH, Ferry N. A Decade of Research on Patterns and Architectures for Iot Security. Cybersecurity. 2022;5:2.

[3] Buja A, Apostolova M, Luma A. Enhancing Cyber Security in Industrial Internet of Things Systems: An Experimental Assessment. 12th Mediterranean Conference on Embedded Computing (MECO). 2023.

[4] Badr Y, Zhu X, Alraja MN. Security and Privacy in the Internet of Things: Threats and Challenges. Serv Oriented Comput Appl. 2021;15:257-271.

[5] https://www.globalcyberalliance.org/

[6] Li J, Othman MS, Chen H, Yusuf LM. Optimizing Iot Intrusion Detection System: Feature Selection Versus Feature Extraction in Machine Learning. J Big Data. 2024;11:36.

[7] Singh R, Ujjwal RL. Feature Selection Methods for IoT Intrusion Detection System: Comparative Study. Comp Intell. 2023:227-236.

[8] Buja A, Pacolli M, Bajrami D, Polstra P, Mutoh A. Time-Series Analysis on Aide Iot Attack Data Unraveling Trends and Patterns for Enhanced Security. Adv Artif Intell Mach Learn;2024;4:2233-2243.

[9] https://research-methodology.net/research-methodology/research-types/applied-research/

[10] Nielsen C, Lund M, Montemari M, Paolone F, Massaro M, et al . Business Models. Routledge. 2018.

[11] Haider RZ, Aslam B, Abbas H, Iqbal Z. C2-Eye: Framework for Detecting Command and Control (C2) Connection of Supply Chain Attacks. Int J Inf Secur. 2024.

[12] https://martina.lindorfer.in/files/papers/c2miner_asiaccs24.pdf

[13] Khan A, Cotton C. Efficient Attack Detection in Iot Devices Using Feature Engineering-Less Machine Learning. Int J Comput Sci Inf Technol. 2022;14:47-64.

[14] Jullian O, Otero B, Rodriguez E, Gutierrez N, Antona H, Canal R. Deep-Learning Based Detection for Cyber-Attacks in Iot Networks: A Distributed Attack Detection Framework. J Netw Syst Manage. 2023;31:33.

[15] Sarhan M, Layeghy S, Portmann M. Feature Analysis for Machine Learning-Based Iot Intrusion Detection. Cryptogr Sec. 2022. https://arxiv.org/abs/2108.12732

[16] Haque S, El-Moussa F, Komninos N, Muttukrishnan R. A Systematic Review of Data-Driven Attack Detection Trends in Iot. Sensors (Basel). Aug. 2023;23:7191.

[17] Rajmohan T, Nguyen PH, Ferry N. A Decade of Research on Patterns and Architectures for Iot Security. Cybersecurity. 2022;5:2

[18] Gueye T, Wang Y, Rehman M, Mushtaq RT, Zahoor S. A Novel Method to Detect Cyber-Attacks in Iot/Iiot Devices on the Modbus Protocol Using Deep Learning. Clust Comput. 2023;26:2947-2473

[19] Dalal S, Lilhore UK, Faujdar N, Simaiya S, Ayadi M, et al. Next-Generation Cyber Attack Prediction for Iot Systems: Leveraging Multi-Class Svm and Optimized Chaid Decision Tree. J Cloud Comput. 2023;12:137.

[20] Seong TB, Ponnusamy V, Zaman Jhanjhi N, Annur R, Talib MN. A Comparative Analysis on Traditional Wired Datasets and the Need for Wireless Datasets for Iot Wireless Intrusion Detection. IJEECS.2021;22:1165-1176.

[21] Saied M, Guirguis S, Madbouly M. A Comparative Analysis of Using Ensemble Trees for Botnet Detection and Classification in IOT. Sci Rep. 2023;13:21632.