REGISTRA Cyber Threat Analysis based on ANN using Event Profiles

[1]Bondada Naveen Kumar, [2]D.S.Ramkiran

[1,2]Dept. of Computer Science & Engineering, KIET, Kakinada, AP, India

Abstract

Cyber Supply Chain (CSC) system is complex which involves different sub-systems performing various tasks. Security in supply chain is challenging due to the inherent vulnerabilities and threats from any part of the system which can be exploited at any point within the supply chain. This can cause a severe disruption on the overall business continuity. Therefore, it is paramount important to understand and predicate the threats so that organization can undertake necessary control measures for the supply chain security. Cyber Threat Intelligence (CTI) provides an intelligence analysis to discover unknown to known threats using various properties including threat actor skill and motivation, Tactics, Techniques, and Procedure (TT and P), and Indicator of Compromise (IoC). This paper aims to analyse and predicate threats to improve cyber supply chain security. We have applied Cyber Threat Intelligence (CTI) with Machine Learning (ML) techniques to analyse and predict the threats based on the CTI properties. That allows to identify the inherent CSC vulnerabilities so that appropriate control actions can be undertaken for the overall cybersecurity improvement. To demonstrate the applicability of our approach, CTI data is gathered and a number of ML algorithms, i.e., Logistic Regression (LG), Support Vector Machine (SVM), Random Forest (RF), and Decision Tree (DT), are used to develop predictive analytics using the Microsoft Malware Prediction dataset.

Parameters and vulnerabilities and Indicators of compromise (IoC) as output parameters. The results relating to the prediction reveal that Spyware/Ransomware and spear phishing are the most predictable threats in CSC. We have also recommended relevant controls to tackle these threats. We advocate using CTI data for the ML predicate model for the overall CSC cyber security improvement.

Keywords

TF-IDF--Term Frequency–Inverse Document Frequency, SVM-Support Vector Machine, CSC-Cyber Supply Chain, CTI-Cyber Threat Intelligence.

I. Introduction

Cyber Supply Chain (CSC) security is critical for reliable service delivery and ensure overall business continuity of Smart CPS. CSC systems by its inherently is complex and vulnerabilities within CSC system environment can cascade from a source node to a number of target nodes of the overall cyber physical system (CPS). A recent NCSC report highlights a list of CSC attacks by exploiting vulnerabilities that exist within the systems. Several organizations outsource part of their business and data to the thirdparty service providers that could lead any potential threat. There are several examples for successful CSC attacks. For instance, Dragonfly, a Cyber Espionage group, is well known for targeting CSC organization . The Saudi Aramco power station attack halted its operation due to a massive cyberattack . There are existing works that consider CSC threats and risks but a lack of focus on threat intelligence properties for the overall cyber security improvement. Further, it is also essential to predict the cyberattack trends so that the organization can take the timely decision for its countermeasure. Predictive analytics not only provide an understanding of the TTPs, motives and intents of the threat actors but also assist situational awareness of current supply system vulnerabilities. This paper aims to improve the cybersecurity of CSC by specifically focusing on integrating Cyber Threat Intelligence (CTI) and Machine Learning (ML) techniques to predicate cyberattack patterns on CSC systems and recommend suitable controls to tackle the attacks. The novelty of our work is threefold:

1. Firstly, we consider Cyber Threat Intelligence(CTI) for systematic gathering and analysis of information about the threat actor and cyber-attack by using various concepts such as threat actor skill, motivation, IoC, TTP and incidents. The reason for considering CTI is that it provides evidence-based knowledge relating to the known attacks. This information is further used to discover unknown attacks so that threats can be well understood and mitigated. CTI provides intelligence information with the aim of preventing attacks as well as shorten time to discover new attacks.
2. Secondly, we applied ML techniques and classification algorithms and mapped with the CTI propreteis to predict the attacks. We use several classification algorithms such as Logistic Regression (LG), Support Vector Machine (SVM), Random Forest (RF) and Decision Tree (DT) for this purpose. We follow CTI properties such as Indicator of Compromise (IoC) and Tactics, Techniques and Procedure (TTP) for the attack predication.
3. Finally, we consider widely used cyberattack dataset to predict the potential attacks . The predication focuses on determining threats relating to Advance Persistent Threat (APT), command and control and industrial espionage which are relevant for CSC. The result shows the integration of CTI and ML techniques can effectively be used to predict cyberattacks and identification of CSC systems vulnerabilities. Furthermore, our prediction reveals a total accuracy of 85% for the TPR and FPR. The results also indicate that LG and SVM produced the highest accuracy in terms of threat predication.

II. Literature Survey

The production of renewable energy is increasing worldwide. To integrate renewable sources in electrical smart grids able to adapt to changes in power usage in heterogeneous local zones, it is necessary to accurately predict the power production that can be achieved from renewable energy sources. By using such predictions, it is possible to plan the power production from non-renewable energy plants to properly allocate the produced power and compensate possible unbalances. In particular, it is important to predict the unbalance between the power produced and the actual power intake at a local level (zones). In this paper, we propose a novel method for predicting the sign of the unbalance between the power produced by renewable sources and the power intake at the local level, considering zones composed of multiple power plants and with heterogeneous characteristics. The method uses a set of historical features and is based on Computational Intelligence techniques able to learn the relationship between historical data and the power unbalance in heterogeneous geographical regions. As a case study, we evaluated the proposed method using data collected by a player in the energy market over a period of seven months. In this preliminary study, we evaluated different configurations of the proposed method, achieving results considered as satisfactory by a player in the energy market. predict future cyber attack trends.

A. Scope- based

A threat could be anything that leads to interruption, disruption or destruction of any valuable service or asset within an organization's technology ecosystem. Whether of "human" or "nonhuman" origin, a cyberthreat analysis must scrutinize each potential vector that may bring about conceivable security risk to a system or asset. To support an organization's efforts to identify, remediate and prepare for potential threat provides a structured, repeatable process. The outputs of the process are combined with the knowledge of internal data and external guidance and recommendations concerning the vulnerabilities pertinent to a particular organization. Finally, the vulnerabilities identified are evaluated to define their probability of occurring and their potential impact. In other words, this threat-oriented approach to defending against cyberattacks represents a transition from a state of reactive security to a proactive one. Ultimately, an organization identifies how it can better protect the availability, confidentiality and integrity of its technology assets without affecting their usability and functionality.

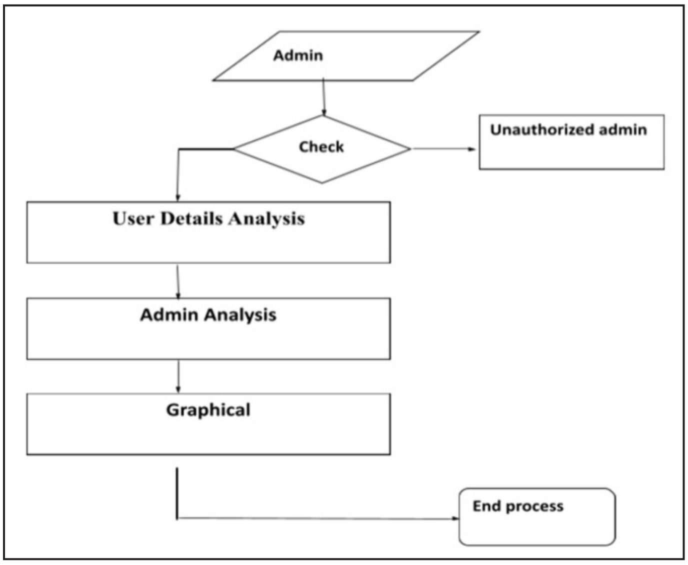


Fig. 1: Flow Diagram of admin

B. Motivation

It's easy to assume that hackers are trying to get into your network the "old-fashioned" way. You might picture them hacking your network trying to get your passwords and usernames or breaking through your firewall protection. While some hackers will do this (it's easy for them if you use simple passwords), many of today's cybercriminals rely on social engineering.

The most common form of social engineering is the phishing scam. The criminal sends you or your employees an e-mail, hoping someone will click a link or open an attached file. Cybercriminals have gotten VERY sophisticated. These e-mails can mimic the look of a legitimate e-mail from a legitimate business, such as the local bank you work with or another company you buy from (or that buys from you). Social engineering is all about tricking people.

III. Existing System

Cyber security professionals continually defend computer systems against different types of cyber threats. Cyber attacks hit businesses and private systems every day, and the variety of attacks has increased quickly. According to former Cisco CEO John Chambers, "There are two types of companies: those that have been hacked, and those who don't yet know they have been hacked." The motives for cyber attacks are many. One is money. Cyber attackers may take a system offline and demand payment to restore its functionality. Ransomware, an attack that requires payment to restore services, is now more sophisticated than ever.

Corporations are vulnerable to cyber attacks, but individuals are targets too, often because they store personal information on their mobile phones and use insecure public networks.

Tracking evolving and increasing cyber attacks is key to better cyber security. As cyber security professionals work to increase their knowledge of threats and cyber security information, earning an online cyber security master's degree can be invaluable.

IV. Proposed System

This paper aims to improve the cybersecurity of CSC by specifically focusing on integrating Cyber Threat Intelligence (CTI) and Machine Learning (ML) techniques to predicate cyberattack patterns on CSC systems and recommend suitable controls to tackle the attacks

V. Architecture

The architecture provides the entire process flow of the system.

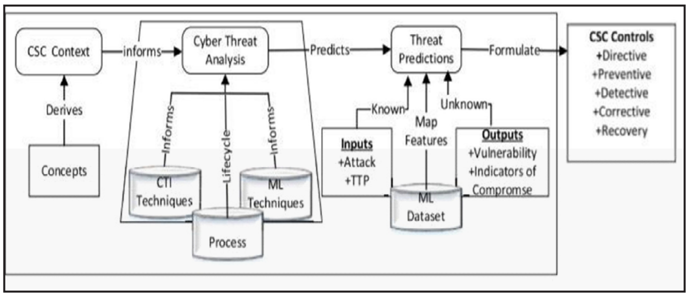


Fig. 2: Architecture diagram

VI. Implementation

A. Algorithms Used

1. Modules Used in Project

a) Tensorflow: TensorFlow is a free and open-sourcesoftware library for dataflow and differentiable programming across a range of tasks. It is a symbolic math library, and is also used for machine learning applications such as neural networks. It is used for both research and production at Google. TensorFlow was developed by the Google Brain team for internal Google use. It was released under the Apache 2.0open-source license on November 9, 2015.

b) Numpy: Numpy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays.

It is the fundamental package for scientific computing with Python. It contains various features including these important ones:

- A powerful N-dimensional array object
- Sophisticated (broadcasting) functions

- Tools for integrating C/C++ and Fortran code
- Useful linear algebra, Fourier transform, and random number capabilities
- Besides its obvious scientific uses, Numpy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined using Numpy which allows Numpy to seamlessly and speedily integrate with a wide variety of databases.

c) Pandas: Pandas is an open-source Python Library providing high-performance data manipulation and analysis tool using its powerful data structures. Python was majorly used for data munging and preparation. It had very little contribution towards data analysis. Pandas solved this problem. Using Pandas, we can accomplish five typical steps in the processing and analysis of data, regardless of the origin of data load, prepare, manipulate, model, and analyze. Python with Pandas is used in a wide range of fields including academic and commercial domains including finance, economics, Statistics, analytics, etc.

d). Matplotlib:

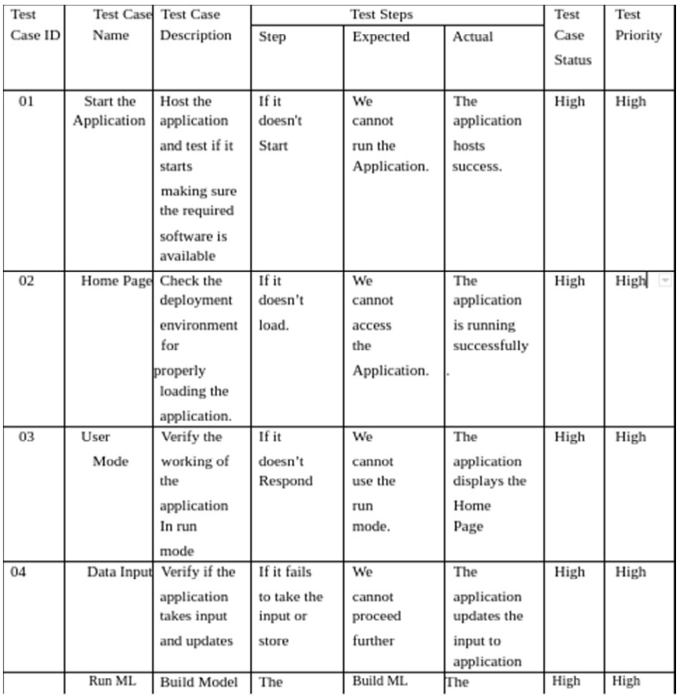
- Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shells, the Jupyter Notebook, web application servers, and four graphical user interface toolkits. Matplotlib tries to make easy things easy and hard things possible. You can generate plots, histograms, power spectra, bar charts, error charts, scatter plots, etc., with just a few lines of code. For examples, see the sample plots and thumbnail gallery.
- For simple plotting the pyplot module provides a MATLAB-like interface, particularly when combined with IPython. For the power user, you have full control of line styles, font properties, axes properties, etc, via an object oriented interface or via a set of functions familiar to MATLAB users.

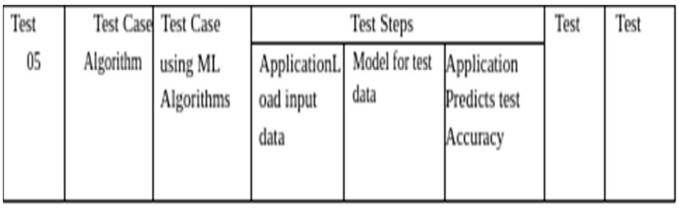
e). Scikit – learn: Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python. It is licensed under a permissive simplified BSD license and is distributed under many Linux distributions, encouraging academic and commercial use.

B. Methodology

Testing is a process of executing a program with the aim off inding error. To make our software perform well it should be error free. If testing is done successfully it will remove all the errors from the software.

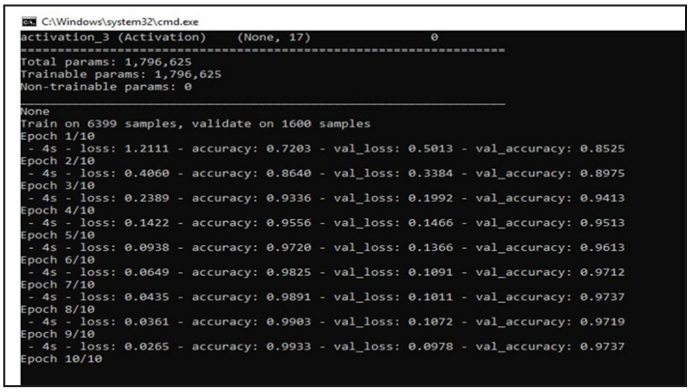
1. Unit Testing: Software verification and validation method in which a programmer tests if individual unitsof source code are fit for use. It is usually conducted by the development team.
2. Integration Testing:The phase in software testing in which individual software modules are combined and tested as a group. It is usually conducted by testing teams.
3. Alpha Testing: Type of testing a software product or system conducted at the developer's site. Usually it is performed by the end users.
4. Beta Testing: Final testing before releasing application for commercial purpose. It is typically done by end- users or others.
5. Performance Testing: Functional testing conducted to evaluate the compliance of a system or component with specified performance requirements. It is usually conducted by the performance engineer.



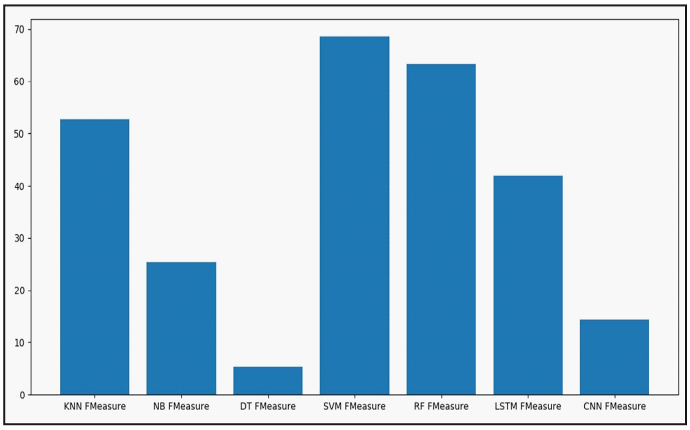


Test Case ID	Test Case Name	Test Case Description	Test Steps			Test Case Status	Test Priority
			Step	Expected	Actual		
01	Start the Application	Host the application and test if it starts making sure the required software is available	If it doesn't Start	We cannot run the Application.	The application hosts success.	High	High
02	Home Page	Check the deployment environment for properly loading the application.	If it doesn't load.	We cannot access the Application.	The application is running successfully .	High	High
03	User Mode	Verify the working of the application In run mode	If it doesn't Respond	We cannot use the run mode.	The application displays the Home Page	High	High
04	Data Input	Verify if the application takes input and updates	If it fails to take the input or store	We cannot proceed further	The application updates the input to application	High	High
	Run ML	Build Model	The	Build ML	The	High	High
Test	Test Case	Test Case	Test Steps			Test	Test
05	Algorithm	using ML Algorithms	ApplicationLoad input data	Model for test data	Application Predicts test Accuracy		

VII. Results



```
activation_3 (Activation)    (None, 17)              0
=================================================================
Total params: 1,796,625
Trainable params: 1,796,625
Non-trainable params: 0
_________________________________________________________________
None
Train on 6399 samples, validate on 1600 samples
Epoch 1/10
 - 4s - loss: 1.2111 - accuracy: 0.7203 - val_loss: 0.5013 - val_accuracy: 0.8525
Epoch 2/10
 - 4s - loss: 0.4060 - accuracy: 0.8640 - val_loss: 0.3384 - val_accuracy: 0.8975
Epoch 3/10
 - 4s - loss: 0.2389 - accuracy: 0.9336 - val_loss: 0.1992 - val_accuracy: 0.9413
Epoch 4/10
 - 4s - loss: 0.1422 - accuracy: 0.9556 - val_loss: 0.1466 - val_accuracy: 0.9513
Epoch 5/10
 - 4s - loss: 0.0938 - accuracy: 0.9720 - val_loss: 0.1366 - val_accuracy: 0.9613
Epoch 6/10
 - 4s - loss: 0.0649 - accuracy: 0.9825 - val_loss: 0.1091 - val_accuracy: 0.9712
Epoch 7/10
 - 4s - loss: 0.0435 - accuracy: 0.9891 - val_loss: 0.1011 - val_accuracy: 0.9737
Epoch 8/10
 - 4s - loss: 0.0361 - accuracy: 0.9903 - val_loss: 0.1072 - val_accuracy: 0.9719
Epoch 9/10
 - 4s - loss: 0.0265 - accuracy: 0.9933 - val_loss: 0.0978 - val_accuracy: 0.9737
Epoch 10/10
```



VIII. Acknowledgment

We express our sincere gratitude to our guide, Assistant Professor Mr. G. L AnandBabu for suggestion and support during every stage of this work. We also convey our deep sense of gratitude to Professor Dr. K. S. Reddy, Head of Information Technology department.

References

[1] National Cyber Security Center. (2018). Example of Supply Chain Attacks. [Online] Available:https://www.ncsc.gov.uk/collection/supplychainsecurity/supply-chain-atta ck-examples

[2] A. Yeboah-Ofori and S. Islam, ‘‘Cyber security threat modelling for supply chain organizational environments,’’ MDPI. Future Internet, vol. 11, no. 3, p. 63, Mar. 2019. [Online]. Available: https://www.mdpi.com/1999-5903/11/3/63

[3] B. Woods and A. Bochman, ‘‘Supply chain in the software era,’’ in the Scowcroft Center for Strategic and Security. Washington, DC, USA: Atlantic Council, May 2018.

[4] Exploring the Opportunities and Limitations of Current Threat Intelligence Platforms, Version 1, ENISA, Dec. 2017. [Online]. Available: https://www.enisa.europa.eu/publications/exploring-the-opportunitiesand-limitations -of-current- threat-intelligence-platforms

[5] C. Doerr, TU Delft CTI Labs. (2018). Cyber Threat Intelligences Standards—A High Level Overview. [Online]. Available: https://www.enisa.europa.eu/events/2018-cti-eu-event/cti-eu-2018-presentations/ cyber-threat-intelligence-standardization.pdf

[6] Research Prediction. (2019). Microsoft Malware Prediction. [Online]. Available: https://www.kaggle.com/c/microsoft-malware-prediction/data

[7] A. Yeboah-Ofori and F. Katsriku, ‘‘Cybercrime and risks for cyber physical systems,’’ Int. J. Cyber-Security. Digit. Forensics, vol. 8, no. 1, pp. 43–57, 2019.

[8] CAPEC-437, Supply Chain. (Oct. 2018). Common Attack Pattern Enumeration and Classification: Domain of Attack. [Online]. Available: https://capec.mitre.org/data/definitions/437.html

[9] Open Web Application Security Project (OWASP). (2017). The Ten Most Critical Application Security Risks, Creative Commons Attribution-Share Alike 4.0 International License. [Online] Available: https://owasp.org/ www-pdf-archive/OWASP_Top_10-2017_%28en%29.pdf.pdf

[10] US-Cert. (2020). Building Security in Software & Supply Chain Assurance. [Online]. Available: https://www.us-cert.gov/bsi/articles/