International Journal of Research in Engineering Sciences and Management
Vol.2, No.1, Jan-June' 2025
E-ISSN: 3107-4383

CYBER THREAT PREDICTIVE ANALYTICS FOR IMPROVINGCYBER SUPPLY CHAIN SECURITY

A. Mithil[1*], K.Madhavan[1],S Pavithra[1],A.Mallikarjuna[2]

[1]Dept. of Computer Science, Chennai Institute of Technology, Chennai-6000035, India.

[2]Dept. of Physics, Sree Venkateswara college of Engineering, North Raju Palem,Nellore-524316, India.

[1]Emails: mithilyadav07@gmail.com, madhavank@citchennai.net,Pavithras@citchennai.net

[2]Emails: drmallikarjuna1979@gmail.com

ABSTRACT:

The Cyber Supply Chain (CSC) system is intricate, comprising various subsystems tasked with different functions. Securing this supply chain is challenging due to inherent vulnerabilities, which can be exploited anywhere within it, posing a significant risk to business continuity. Therefore, it's crucial to comprehend and anticipate potential threats to implement adequate security measures. Cyber Threat Intelligence (CTI) offers insights into identifying threats, utilizing factors such as threat actor skills, motivation, Tactics, Techniques, and Procedures (TTPs), and Indicators of Compromise (IoCs). This study aims to analyze and anticipate threats to enhance cyber supply chain security by leveraging CTI alongside Machine Learning (ML) techniques. By doing so, inherent vulnerabilities in the CSC can be pinpointed, enabling organizations to take appropriate control measures for overall cybersecurity enhancement.to validate our approach, CTI data was collected and several ML algorithms, including Logistic Regression (LG), Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), Cat Boost, and Gradient Boost, were employed using the Microsoft Malware Prediction dataset. The experiment focused on input parameters such as attack and TTP, with vulnerabilities and Indicators of Compromise (IoC) as output parameters. Results from the predictive analytics indicated that Spyware/Ransomware and spear phishing were the most foreseeable threats within the CSC. Additionally, we suggested relevant controls to mitigate these threats. We advocate the utilization of CTI data for ML-based predictive modeling to bolster overall CSC cybersecurity.

Keywords: Cyber Supply Chain (CSC), Cyber Threat Intelligence (CTI), Logistic Regression (LG), Random Forest (RF), Decision Tree (DT), Tactics, Techniques, and Procedures (TTPs).

1. INTRODUCTION

1.1 Motivation: The digital sphere is ever-changing, with cyber threats increasing in complexity and frequency. Recent notable cyberattacks targeting supply chain networks emphasize the critical necessity for preemptive security measures.

1.2 Problem Statement: The increasing complexity of the cyber supply chain presents a growing challenge, as malicious actors continually adapt their strategies. Traditional cybersecurity methods often fall short in identifying and thwarting these evolving threats. This initiative aims to tackle this pressing issue by creating advanced predictive analytics models that can anticipate cyberattacks in the supply chain, enabling proactive and timely security measures.

1.3 Objective of the Project: The primary goal of this initiative is to create a reliable predictive analytics platform for evaluating the probability of cyberattacks within the cyber supply chain. Through the utilization of advanced data analysis and machine learning methods, our objective is to forecast cyber threats, empowering the implementation of proactive security measures to protect crucial digital assets and infrastructure.

1.4 Scope: This project involves thoroughly examining past cyber threat information, creating predictive models, and setting up live monitoring systems throughout the cyber supply chain. It will address various attack methods like malware, phishing, and insider threats, aiming for a comprehensive cybersecurity strategy applicable to different sectors and businesses.

2. LITERATURE SURVEY

2.1 Related Work: Safeguarding the reliability of Industrial Internet of Things (IIoT) networks is paramount to prevent potential harm, given their critical role in industrial operations. However, conventional security measures fall short due to protocol disparities, limited upgrade avenues, and outdated operating systems common in industrial environments. To tackle this challenge, we propose bolstering the trustworthiness of IIoT networks, particularly within

Supervisory Control and Data Acquisition (SCADA) systems, by introducing an efficient cyber-attack detection framework. Our method integrates random subspace learning with random tree ensemble learning to scrutinize SCADA-based IIoT network traffic. This innovative model mitigates issues such as irrelevant features and overfitting, leading to superior detection rates compared to traditional approaches. Evaluation across 15 SCADA datasets illustrates the effectiveness of our model in fortifying IIoT platform security and reliability.

Detecting cyber-attacks using machine learning is vital given the increasing frequency and complexity of such incidents. It's crucial for understanding the situation and implementing effective defense strategies. While traditional methods like spam filters, firewalls, and IDS/IPS are common, adversaries are now using adversarial machine learning to exploit weaknesses. This research examines how machine learning can predict and classify malware attacks to improve cyber threat intelligence. We propose creating a classifier to automatically identify events as either "Detected" or "Not Detected", assessing the likelihood of malware infiltration and network manipulation. We focus on decision tree algorithms, using a dataset from the Microsoft Malware threat prediction Kaggle site to explore our approach's feasibility. By simulating cyber-attacks on smart grid systems, we show how machine learning can detect and predict threats. Essentially, our study aims to use machine learning to train classifiers and decision tree algorithms to identify potential cyber-attacks and their detection status.

Ensuring cybersecurity throughout a supply chain is essential for organizations to safely achieve their business goals. While technology integration has streamlined operations, it has also brought challenges like increased dependencies among stakeholders. These challenges include the lack of third-party audit mechanisms and the rise of interconnected cyber threats. These threats range from tampering with design specifications to changes during distribution. This study aims to understand and address these threats by examining cyber supply chain (CSC) attacks and how stakeholders report them. It considers concepts such as objectives, participants, attacks, tactics, techniques, and procedures (TTPs), and threat actors within the supply chain framework. The proposed model utilizes the STIX threat model and is validated through a case study involving a smart grid scenario. An algorithm is developed to simulate the attack, and a discrete probability method assesses its spread and cascading effects. The results show the efficacy of this approach

in threat analysis. Furthermore, a set of CSC controls is suggested to bolster the organization's overall security posture.

As cyber threats evolve, it's evident that no system is entirely safe. Thus, it's essential to evaluate and anticipate risk levels within systems to prevent potential cyber attacks. This is where Risk Teller comes in: a system designed to analyze logs of binary file occurrences on devices, forecasting which devices might be vulnerable to infection months beforehand. Risk prediction models are developed by creating thorough profiles for each device, capturing usage patterns, and then correlating these profiles with risk levels through various learning methods. The system underwent testing using a year-long dataset from 18 different enterprises, showcasing its ability to accurately foresee future infections based on device profiles. This proactive cybersecurity approach reflects the increasing trend of transitioning from reactive to proactive security measures, especially as individuals and businesses seek cyber insurance to mitigate potential losses stemming from inevitable cyber incidents.

Power system disruptions, arising from a range of natural and human activities, present significant difficulties in identifying their sources and appropriate responses. Operators shoulder the responsibility of deciphering these disruptions, but in instances of cyber-attacks, where deceit is prevalent, human judgment becomes less dependable. To assist decision-makers, we explore the potential of machine learning in distinguishing different types of power system disruptions, with a particular emphasis on detecting cyber-attacks marked by deception. By assessing various machine learning methods, our objective is to bolster existing power system frameworks by integrating automated disruption classification within a smart grid setup. This investigation sets initial standards for utilizing machine learning to bolster power system resilience against cyber threats.

Feature selection involves the process of choosing a subset of original features, aiming to enhance efficiency and accuracy by eliminating redundant and irrelevant ones. Widely utilized in machine learning, feature selection finds application across various domains. Our novel approach combines Affinity Propagation and SVM sensitivity analysis to produce a feature subset, further refining it through forward selection and backward elimination based on feature ranking. We apply this method to address human resource selection, utilizing data gathered through questionnaire surveys.

Simulation results demonstrate the effectiveness of our approach, reducing the number of human resource features while improving classification performance.

3. SYSTEM ANALYSIS

3.1 Existing System: The rising prominence of machine learning has led to a clear distinction between traditional and machine learning-based computer techniques. This segment explores previous studies on psychosocial instabilities using conventional approaches and highlights the superiority of machine learning methods over traditional ones. The current methodology employed in this project follows a specific workflow for model development. However, it faces challenges such as high memory requirements and inadequate accuracy in results.

3.2Disadvantages

- Low efficiency.
- Time consuming.
- High complexities.
- Resources consuming

3.3 Proposed System

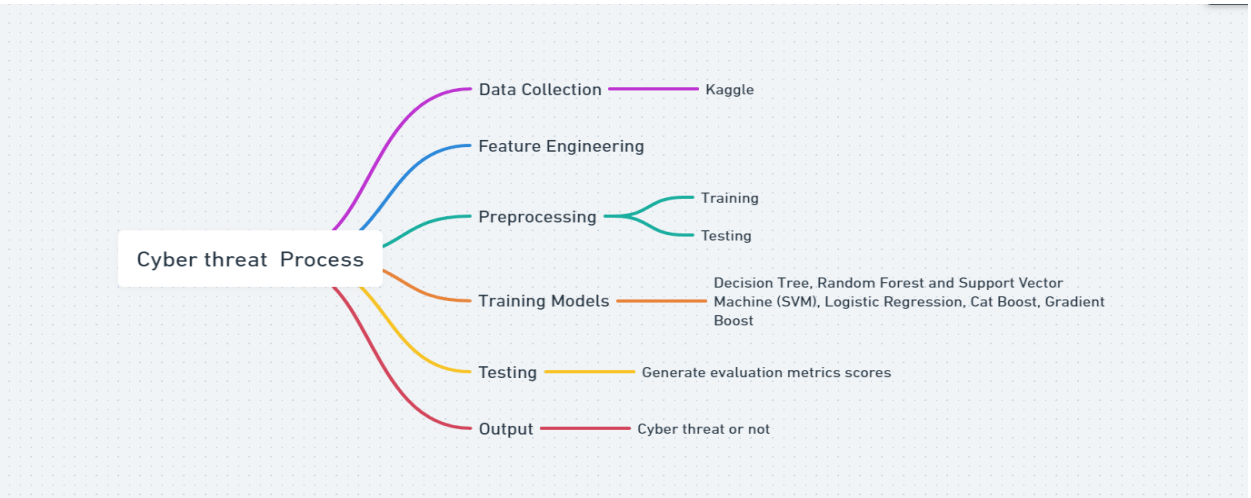
In our anticipated model, we evaluated ten features to enhance the distinctiveness of our comparison. Our algorithms were then assessed based on the results obtained and compared with previous studies to demonstrate the percentage of improvement. In one instance, a decrease in performance was observed with CatBoost. The most significant increase was seen in comparison to previous works, particularly in the realm of cyber threat assessment. We utilized machine learning techniques such as Decision Trees, Random Forests, Support Vector Machines (SVM), Logistic Regression, CatBoost, and Gradient Boosting to discern whether an event constitutes a cyber-attack or not.

3.4 Advantages

- High efficiency.

- Time Saving.
- Inexpensive.
- Low complexities.

3.5 work Flow of Proposed system



4. REQUIREMENT ANALYSIS

4.1 Functional Requirements:

1. Data Gathering and Fusion: The system must effectively gather and fuse Cyber Threat Intelligence (CTI) data from diverse sources.

2. Threat Examination: The system should analyze collected data to detect potential cyber threats within the Cyber Supply Chain (CSC), considering factors like threat actor capabilities, motivations, Tactics, Techniques, and Procedures (TTPs), and Indicators of Compromise (IoCs).

3. Integration of Machine Learning: The system needs to incorporate Machine Learning (ML) methods to analyze CTI data and anticipate potential cyber threats within the CSC.

4. Model Training and Assessment: The system must train and assess ML models using past CTI data to enhance prediction accuracy.

5. Threat Forecasting: The system should forecast potential cyber threats within the CSC based on factors such as attack methods, TTPs, vulnerabilities, and IoCs.

6. Recommendation of Control Measures: The system should suggest appropriate control measures to mitigate identified cyber threats within the CSC.

Non-functional Requirements:

1. Performance: The system should efficiently process and analyze large amounts of CTI data to offer timely threat forecasts.

2. Accuracy: ML models utilized should possess high accuracy in predicting CSC cyber threats to minimize erroneous results.

3. Security: The system must guarantee the security and confidentiality of CTI data during collection, processing, and storage to prevent unauthorized access or breaches.

4. Usability: The system interface should be intuitive, enabling cybersecurity professionals to input data easily, interpret results, and implement suggested control measures.

5. Compatibility: The system should seamlessly integrate and operate with existing cybersecurity infrastructure and tools.

Hardware Requirements

4	Operating system	:	Windows 7 or 7+
5	RAM	:	8 GB
6	Hard disc or SSD	:	More than 500 GB
7	Processor	:	Intel 3rd generation or high or Ryzen with 8 GB Ram

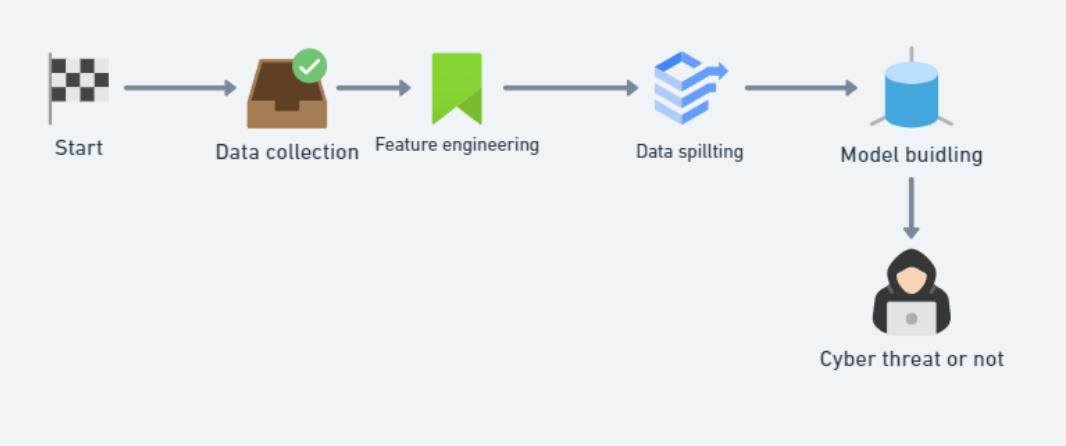
Software Requirements:

8 Software's : Python 3.6 or high version

9 IDE : PyCharm.

10 Framework : Flask

1.4 Architecture:



5.METHODOLOGY

5.11. Logistic Regression (LG):

Overview: Logistic Regression is a statistical approach employed in binary classification problems. It predicts the probability of a binary outcome based on one or more predictor variables.

Mechanism: Logistic Regression computes the likelihood that a given input belongs to a specific class by applying the logistic function to a linear combination of predictor variables. The logistic function transforms real-valued inputs into probabilities between 0 and 1, indicating the likelihood of the input belonging to the positive class. Through optimization techniques, the model is trained to determine coefficients that best fit the data.

Suitability for the Project: Logistic Regression is well-suited for this project due to its ability to provide probabilities for outcomes, aiding in understanding the likelihood of various cyber threats within the Cyber Supply Chain (CSC). Moreover, its computational efficiency and

interpretability make it suitable for analyzing and interpreting relationships between predictor variables and cyber threats.

2. Support Vector Machine (SVM):

Overview: Support Vector Machine is a supervised learning algorithm utilized for classification and regression tasks. It seeks to find the optimal hyperplane that effectively separates data points into different classes.

Mechanism: SVM operates by identifying the hyperplane that maximizes the margin between the closest data points of distinct classes, termed support vectors. It transforms input data into a higher-dimensional space using a kernel function to achieve separability of classes. Subsequently, SVM identifies the hyperplane with the maximum margin in this transformed space.

Suitability for the Project: SVM is appropriate for this project due to its capability to handle high-dimensional data and effectively deal with linearly and non-linearly separable data. Given the complex and non-linear relationships often involved in cyber threat prediction, SVM's ability to ascertain complex decision boundaries makes it a viable choice.

3. Random Forest (RF):

Overview: Random Forest is an ensemble learning technique that constructs multiple decision trees and amalgamates their predictions to enhance acscuracy and resilience.

Mechanism: Random Forest builds a forest of decision trees during training, each utilizing a random subset of training data and features. During prediction, each tree independently predicts the class, and the final prediction is determined by majority vote or averaging across all individual tree predictions.

Suitability for the Project: Random Forest is well-suited for this project owing to its resilience to overfitting, adeptness in handling high-dimensional data, and capacity to capture intricate relationships between predictor variables and cyber threats. Additionally, its provision of feature importance measures aids in identifying the most significant factors contributing to cyber threat prediction within the CSC.

4. Decision Tree (DT):

Description: The Decision Tree algorithm is a straightforward supervised learning method used for classification and regression tasks. It constructs a tree-like structure of decisions based on the values of input variables.

Operation: Decision Tree divides the data into subsets recursively, guided by the values of chosen input variables. Each split aims to maximize the similarity of subsets concerning the target variable. This process iterates until a stopping condition is met, like reaching a specified depth or no significant improvement in performance.

Suitability for the Project: Decision Tree is ideal for this project due to its simplicity, making it easy to interpret and visualize. It's valuable for comprehending the factors influencing cyber threats in the CSC. Moreover, Decision Trees can grasp non-linear relationships, enhancing predictive modeling for this project.

5. Cat Boost:

Description: Cat Boost stands as a gradient boosting algorithm tailored for handling categorical features efficiently in machine learning tasks. It's an extension of gradient boosting, alleviating the need for preprocessing categorical variables like one-hot encoding.

Operation: Cat Boost constructs an ensemble of decision trees sequentially, where each subsequent tree rectifies errors from prior ones. Notably, it employs a unique approach to handle categorical variables, considering their statistical properties during tree building. Additionally, Cat Boost integrates regularization techniques to curb overfitting.

Suitability for the Project: Cat Boost is well-suited for this project due to its natural handling of categorical variables, crucial for analyzing cyber threat data. Its regularization methods also enhance generalization, crucial for effective modeling in cybersecurity applications.

6. Gradient Boost:

Description: Gradient Boosting is an ensemble learning method that amalgamates multiple weak learners, usually decision trees, to forge a robust predictive model. It develops the model incrementally, where each new model rectifies errors from its predecessors.

Operation: Gradient Boosting fits a sequence of decision trees to the data, with each tree predicting the residuals from previous ones. The final prediction aggregates the contributions of all trees in the ensemble.

Suitability for the Project: Gradient Boosting suits this project well due to its ability to capture intricate relationships and interactions between predictor variables and cyber threats. Its robustness to overfitting and resilience with noisy data make it a reliable choice for predictive modeling in cybersecurity contexts.

6. SYSTEM DESIGN:

Input Design:

In an information system, the defined output is exposed to as input. Developers must consider input devices such as Windows, OCR, and Wrong . the penal during planning phase.

Output Design:

The most important tip towards every system is the performance design. During system analysis, developers determine the level of areas to improve, used and the distributes power and low fidelity report layouts.

Implementation and Results

1.User:

1.1 Home Page Viewing:

Users can access the Cyber Threat application's home page to get started.

1.2 About Page Exploration:

The about page provides insights into the Cyber Threat platform, offering users an opportunity to learn more.

1.3 Data Loading:

In the load_data page, users can import datasets essential for modeling.

1.4 Model Input:

Users input specific values into designated fields to initiate model processing.

1.5 Results Viewing:

Users can observe the outcomes generated by the model.

1.6 Accuracy Assessment:

Users have the capability to view the accuracy score represented as a percentage.

Graph:

Graphical representation illustrating the comparison of accuracy across various models.

2.System

2.1 Data Verification:

The system checks for data availability and proceeds to load it into CSV files.

2.2 Data Preprocessing:

Preprocessing procedures are applied to enhance model accuracy and provide comprehensive data insights.

2.3 Data Training:

The dataset is divided into training and testing sets before undergoing algorithm training.

2.4 Model Construction:

This module assists in building models optimized for predicting datasets accurately.

2.5 Performance Evaluation:

Users can access the accuracy score, depicted as a percentage.

2.6 Result Generation:

Machine learning algorithms are trained to generate predictions efficiently.

CONCLUSION

The implementation of various machine learning classifiers, including Random Forest Classifier, Decision Tree Classifier, Gradient Boosting Classifier, Logistic Regression, Cat Boost Classifier, Ada Boost Classifier, and Extra Tree Classifier, in the context of "Cyber Threat Predictive Analytics for Improving Cyber Supply Chain Security" underscores the importance of a multifaceted approach to cybersecurity within supply chains. These algorithms collectively contribute to a robust cybersecurity strategy. RandomForest and ExtraTreeClassifier excel at handling complex data, while Decision Tree Classifier provides transparency. Gradient Boosting Classifier, Cat Boost Classifier, and Ada Boost Classifier offer ensemble-based adaptability. LogisticRegression provides a strong baseline, and CatBoostClassifier stands out for managing categorical data efficiently. The project's comprehensive approach leverages the strengths of each algorithm to enhance threat detection, risk assessment, and proactive defense in the dynamic and interconnected realm of cyber supply chains. In doing so, it contributes to the long-term resilience and security of supply chain networks while aligning with evolving cybersecurity regulations and standards. The literature survey reinforces the significance of these strategies in addressing the evolving cyber threat landscape, emphasizing the need for ongoing research and adaptation to stay ahead of emerging threats.

Future Enhancement

Future enhancements in the realm of cyber threat predictive analytics for improving cyber supply chain security will focus on continuous innovation and adaptation. These developments will encompass advanced machine learning algorithms, leveraging artificial intelligence and deep learning to better identify emerging threats. Enhanced data integration and real-time analysis will enable quicker threat detection and response. Furthermore, there will be a greater emphasis on collaboration and information sharing among organizations within supply chains, creating a collective defense against cyber threats. Improved visualization techniques will make complex threat data more accessible to decision-makers. Additionally, the integration of blockchain technology for secure data sharing and authentication will likely play a role in future enhancements. As threats evolve, so too will the tools and strategies used to protect supply

chains, emphasizing agility and proactive measures to safeguard against an ever-changing cyber landscape.

References

[1] National Cyber Security Centre. (2018). Example of Supply Chain Attacks. [Online] Available: https://www.ncsc.gov.uk/collection/supply chain-security/supply-chain-attack-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. Buchman, ''Supply chain in the software era,'' in Scowcroft Centre 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-opportunities and-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: ttps://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-Secur. Digit. Forensics, vol. 8, no. 1, pp. 43–57, 2019.