International Journal Of Engineering Research And Development
e- ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 20, Issue 5 (May. 2024), PP. 150-154

Malware detection through predictive analytics in cyber supply chain security

Kiran B.M
Computer Science and Engineering (JNTUH)
Sphoorthy Engineering College (JNTUH)
Hyderabad, India

G.Sravya
Computer Science and Engineering (JNTUH)
Sphoorthy Engineering College (JNTUH)
Hyderabad, India

A.Sathwika
Computer Science and Engineering (JNTUH)
Sphoorthy Engineering College (JNTUH)
Hyderabad, India

Abstract—*Given the persistent threat of cyber-attacks targeting the cyber supply chain (CSC) and the widespread repercussions of malware infections, we employ machine learning techniques for attack prediction. With organizations increasingly dependent on CSC systems for business continuity, vulnerabilities and threat landscapes have also surged. While traditional methods like Logistic Regression, Decision Trees, and Random Forest through Majority Voting, we conduct training and testing using 10-fe antivirus software have had some success, the evolving sophistication of threat actors enables them to bypass these defenses. Our study utilizes machine learning to analyze datasets and forecast which CSC nodes are susceptible to attacks, aiming to identify vulnerable modes and predict future trends. To validate our approach, we utilize a dataset from the Microsoft Malware Prediction website. Employing an ensemble method, which crossvalidation. Our findings highlight the efficacy of machine learning algorithms, particularly Decision Trees, in enhancing cyber supply chain predict analytics for detecting and forecasting future cyberattack trends.*
Index Terms—*Machine Learning, Cyber Supply Chain, Predictive Analytics. Cyber Security. Cyberattack*

Date of Submission: 07-05-2024 Date of Acceptance: 20-05-2024

I. INTRODUCTION

The cyber supply chain (CSC) system presents a para- dox: while highly efficient for business processes, it's pro- foundly vulnerable from a cybersecurity perspective due to its interconnected nature. This vulnerability stems from the myriad network hosts and nodes involved, granting access to organizational services and sensitive data. Safeguarding the confidentiality, integrity, and availability of CSC systems poses significant challenges, given their integrated and dis- tributed structure, often employing public-facing IPs. Research indicates alarming rates of vulnerabilities among web hosts, with a substantial portion implicated in malicious activities within supply chain systems [1]. Malware attacks on CSC systems manifest in various forms, from injecting viruses or worms into software to executing arbitrary commands remotely, po- tentially leading to Advanced Persistent Threats (APTs). To Identify applicable funding agency here. If none, delete this. address this, we propose leveraging machine learning (ML) techniques to analyze datasets and predict vulnerable CSC nodes, aiming to anticipate future cyber threats. Our study employs Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF) algorithms for data classification. We introduce novelty by cross-validating these algorithms to en hance predictive accuracy and combining them through Major- ity Voting (MV) to determine the most effective approach. Our results, particularly from the DT algorithm, demonstrate the feasibility of ML predictive analytics in CSC security, offering insights into current and future cyber-attack trends.

II.LITERATURE SURVEY

A.Vulnerabilities in CSC Systems

An examination conducted by Meng et al. in 2019 under- scored the heightened susceptibility of CSC systems to cyber- attacks, attributing this susceptibility to their interconnected and internet-reliant architecture. [2] This study emphasized that the extensive integration and openness of CSC systems across diverse industries render them alluring targets for cybercrime activities.

B. Current Approaches to Detection:

Traditional security measures in CSC typically encompass antivirus software, firewalls, as well as intrusion detection and prevention systems (IDS/IPS). [3]However, an analysis by Patel and Qassim in 2021 critiqued these conventional methods as increasingly inadequate in tackling the complexity of modern cyber threats. They argued that these approaches often lack the adaptability needed to counter evolving attack techniques effectively.

C. Integration of Machine Learning:

The integration of machine learning (ML) for dynamic threat prediction and mitigation is gaining prominence in CSC security. Research conducted by Thompson and Tan in 2020 explored the efficacy of Decision Trees and Random Forest algorithms in identifying attack patterns and vulnerabilities within supply chains. Their findings suggest that ML signifi- cantly enhances the capacity to detect and respond to security anomalies by analyzing historical data pertaining to cyber-attacks.

III. SYSTEM ARCHITECTURE

The system architecture for Malware detection through predictive analytics in cyber supply chain security is shown below.

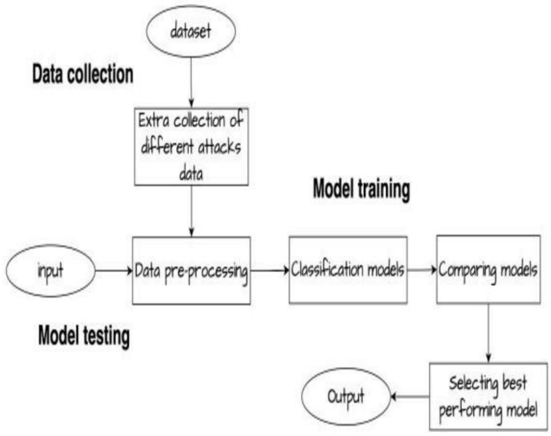

Fig.A Flow Diagram of System Architecture

A.Dataset Description

The dataset focuses on malware attacks within Microsoft endpoint systems, which play a crucial role in the overall business continuity of cyber supply chain (CSC) systems [4]. Designed with specific business constraints regarding privacy and usage timeframes, it provides valuable insights into the intersection of cybersecurity and CSC operations. Given that CSC integrates diverse organizational systems for business processes and information dissemination within Cyber-Physical Systems (CPS) environments, this dataset is particularly relevant. It aggregates threat reports collected by Microsoft Endpoint Protection Solution, Windows Defender, with each row corresponding to a unique machine identifier.

Importantly, the dataset was curated to ensure representation beyond solely Microsoft customers' machines, incorporating a significant proportion of malware-infected machines. This broader sampling enhances its utility for analyzing cyber threats within CSC systems. The dataset's accessibility from the Microsoft Malware Prediction website further underscores its relevance and credibility for our research[5].

B.Feature Extraction

Feature extraction plays a crucial role in preparing data for analysis with classification algorithms, ensuring an accurate representation of the dataset. This process involves employing various techniques to select pertinent features for application in machine learning (ML) algorithms. In our context, we focus on telemetry data relevant to our research:

- MachineIdentifier: Unique identifier for individual ma- chines.
- GeoNameIdentifier: Identifier for the geographic region where a machine is located.
- DefaultBrowsersIdentifier: Identifier for the default browser installed on the machine.
- OrganizationIdentifier: Identifier for the organization to which the machine belongs.
- IsProtected: Calculated field derived from the Spynet Report's AV Products field, indicating whether the machine is protected.
- Processor: Description of the processor architecture of the installed operating system.
- HasTpm: Boolean indicating whether the machine has Trusted Platform Module (TPM) support.
- OsBuild: Build version of the current operating system.
- CensusDeviceFamily: Also known as Device Class, indicate the type of device for which an OS edition is intended (e.g., desktop or mobile).

• Firewall: Boolean indicating whether the Windows firewall is enabled for Windows 8.1 and above, as reported by the services

C.Choosing a Classifier

In our study, we implement classifications using machine learning (ML) algorithms such as Logistic Regression (LR), Decision Trees (DT), and Support Vector Machines (SVM) within a Majority Voting (MV) ensemble. We opt for binary classification as it facilitates the use of metrics like Area Under the ROC Curve (AUCROC) to differentiate between class probabilities. Binary classification also offers precision, recall, and F-score metrics, aiding in the prediction of correct instances.

To optimize our model, we utilize algorithms that identify major features or class levels for each object. Ensemble techniques are employed to combine the predictive power of multiple algorithms and assess dataset performance com- prehensively. [4] Additionally, we employ a K-Fold classifier, running each algorithm ten times to ensure robust results. This iterative approach enhances the reliability and accuracy of our predictions, contributing to a thorough analysis of cyber threat patterns within cyber supply chain systems.

IV. RESULT

When the user clicks "Submit Form,"(Fig.B)the model will collect the user's input, validate the data with the trained model, and then provide the results to the user in the next screen

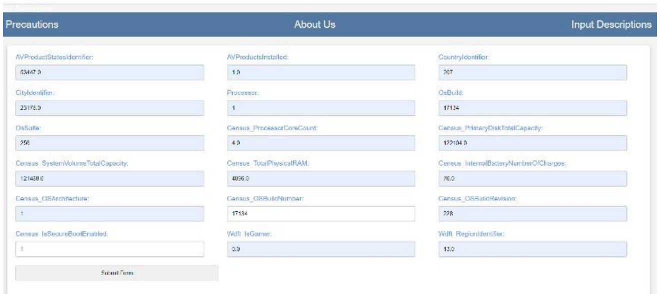

Fig. B User Inputs

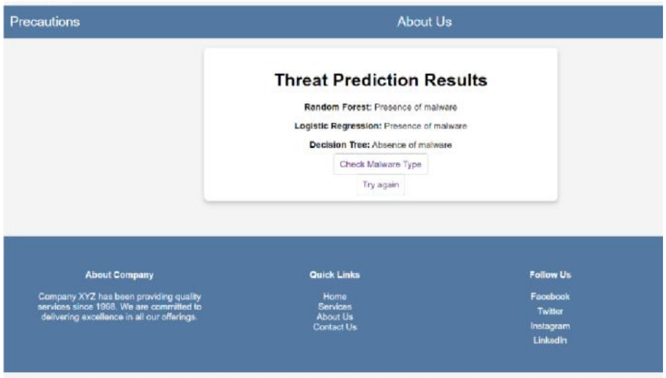

Fig.C Malware Prediction

Three algorithms that we trained for our model—Logistic Regression, Random Forest, and Decision Tree(Fig.C)will be displayed by the model as the output. These algorithms will indicate whether or not the user has malware on them. User can determine which kind of malware is there based on its existence. The next page will appear once you click the "Check Malware Type" button.

The model is able to detect the presence of malware and also the type of malware present(Fig.D)

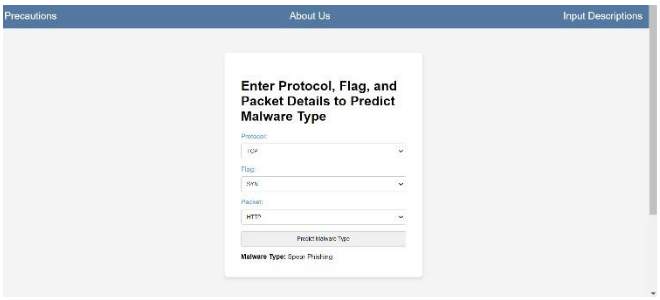

Fig.D Type of Malware

V PERFORMANCE EVALUATION

Precision-Recall Curve: Displays the relationship between precision and recall at different thresholds. A highperforming model has a curve close to the upper-right corner, showing high precision and recall.

```
Accuracy: 0.59480125
Confusion Matrix:
 [[235010 164694]
 [159465 240831]]
Classification Report RandomForestClassifier:
              precision    recall  f1-score   support

           0       0.60      0.59      0.59    399704
           1       0.59      0.60      0.60    400296

    accuracy                           0.59    800000
   macro avg       0.59      0.59      0.59    800000
weighted avg       0.59      0.59      0.59    800000
```

```
Accuracy: 0.53622
Confusion Matrix:
 [[214680 185024]
 [186000 214296]]
Classification Report DecisionTreeClassifier:
              precision    recall  f1-score   support

           0       0.54      0.54      0.54    399704
           1       0.54      0.54      0.54    400296

    accuracy                           0.54    800000
   macro avg       0.54      0.54      0.54    800000
weighted avg       0.54      0.54      0.54    800000
```

Fig.E Classification Report of Random Forest and Decision Tree

F1 Score: The harmonic mean of precision and recall, providingabalanced measure of a model's performance(Fig.E&Fig.F)

Cross-Validation: A technique for assessing model performance by splitting the data into training and validation sets, training the model on different subsets, and evaluating it on validation sets.

```
Accuracy: 0.50420875
Confusion Matrix:
 [[ 10225 389479]
 [  7154 393142]]
Classification Report LogisticRegression:
              precision    recall  f1-score   support

           0       0.59      0.03      0.05    399704
           1       0.50      0.98      0.66    400296

    accuracy                           0.50    800000
   macro avg       0.55      0.50      0.36    800000
weighted avg       0.55      0.50      0.36    800000
```

Fig.F Classification Report of Logistic Regression

VI. CONCLUSION

In conclusion, the predictive analytics approach that uses machine learning algorithms like Random Forest, Decision Trees, and Logistic Regression has demonstrated significant efficacy in identifying and projecting malware risks in the context of the cyber supply chain. Our findings highlight how these algorithms might greatly improve cyber defence systems by providing a proactive means of locating weaknesses and thwarting possible attacks. This paradigm is crucial to contemporary cybersecurity strategies because it not only increases security but also dynamically adjusts to the changing landscape of cyber threats.

REFERENCES

[1]. Adhikari, R., & Xu, K. (2018). Cybersecurity threat prediction and prevention using machine learning algorithms. Journal of Network and Computer Applications, 107, 57-67.

[2]. Buehrer, G., Evans, M., & Lee, W. (2018). Cyber supply chain risk management. Communications of the ACM, 61(4), 45- 49.

[3]. Cavelty, M. D., & Suter, M. (2016). Cybersecurity and cyber resilience: What is the difference?. Cybersecurity, 1(1), 1-9.

[4]. Cho, S. Y., & Kim, J. H. (2017). Machine learning for network intrusion detection: A review. Journal of Information Processing Systems, 13(3), 505-516.

[5]. Microsoft Malware Prediction, Research Prediction. 2019. [Online] Available: https://www.kaggle.com/c/microsoft-malwareprediction/data

[6]. Yeboah-Ofori, J. D. Abduli, F. Katsriku, "Cybercrime and Risks for Cyber- Physical Systems" International Journal of Cyber Security and Digital Forensics. Vol.8 No1, pp 43-57. 2019.

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

[8]. J. Boyens, C. Paulsen, R. Moorthy, and N. Bartol, "Supply Chain Risk Management Practices for Federal Information Systems and Organizations". NIST Computer. Sec. 2015, SP800, 1, doi:10.6028/NIST.SP.800.

[9]. NIST 2018 "Framework for Improving Critical Infrastructure Cybersecurity" National Institute of Standards and Technology. Ver.1.1https://doi.org/10.6028/NIST.CSWP.04162018.

[10]. J. F Miller, "Supply Chain Attack Framework and Attack Pattern". MITRE Technical Report. MTR140021.2013.[Online]Available: https://www.mitre.org/sites/default/files/publications/supplychain-attackframework- 14-0228.pdf.