International Journal of Smart Sensor and Adhoc Network

Volume 3
Issue 4 *Role of Emerging & Intelligent Technologies for the Society.*
Article 6

January 2023

MACHINE LEARNING ALGORITHMS FOR DETECTION OF CYBER THREATS USING LOGISTIC REGRESSION

Hari Gonaygunta
DEPARTMENT OF INFORMATION TECHNOLOGY, UNIVERSITY OF THE CUMBERLANDS,
hmriet@gmail.com

Recommended Citation
Gonaygunta, Hari (2023) "MACHINE LEARNING ALGORITHMS FOR DETECTION OF CYBER THREATS USING LOGISTIC REGRESSION," *International Journal of Smart Sensor and Adhoc Network*: Vol. 3: Iss. 4, Article 6.
DOI: 10.47893/IJSSAN.2023.1229
Available at: https://www.interscience.in/ijssan/vol3/iss4/6

MACHINE LEARNING ALGORITHMS FOR THE DETECTION OF CYBER THREATS USING LOGISTIC REGRESSION

[1]HariGonaygunta
[1]Dept. ofComputer and Information Sciences, University of the Cumberlands, KY, USA
[1]hmriet@gmail.com

Abstract-The threat of cyber attacks is expanding globally; thus, businesses are developing intelligent artificial intelligence systems that can analyze security and other infrastructure logs from their systems department and quickly and automatically identify cyber attacks. Security analytics based on machine learning the next big thing in cyber security is machine data, which aims to mine security data to show the high maintenance costs of static relationship rules and methods. But, choosing the appropriate machine learning technique for log analytics using ML continues to be a significant barrier to AI success in cyber security due to the possibility of a substantial number of false-positive detections in large-scale or global Security Operations Centre (SOC) settings, selecting the proper machine learning technique for security log analytics remains a substantial obstacle to AI success in cyber security. A machine learning technique for a cyber threat exposure system that can minimize false positives is required. Today's machine learning methods for identifying threats frequently use logistic regression. Logistic regression is the first of three machine learning subcategories—supervised, unsupervised, and reinforcement learning. Any machine learning enthusiast will encounter this supervised machine learning algorithm at the beginning of their machine learning career. It's an essential and often applied classification algorithm.

Keywords: SOC, Machine learning, Cyber threats, MLAW, Regression analysis

I. Introduction

Cyber security refers to the policies, defence mechanisms, technologies, and structures put in place to protect programs, data, networks, and computers against illegal access, harm, and cyber threats [1]. The computer network and its applications are one of the fastest-growing components of Information Communication Technology (ICT); due to this promise, cyber threats are also increasing and gaining ground in the cyber world [2]. Individuals, research institutes, companies, and governments have all suffered significant losses and harm due to cyber threats. Many efforts have been put in place by industries, research institutes, and governments to curb the activities of intruders. Still, all actions must be revised to handle the intruders [1].

Many enterprises are nearing physical limits when gathering, parsing, normalizing, searching, analyzing, visualizing, and investigating the vast amount of cyber-defence results collected by Security Information and Management (SIEM) Systems. In any SOC, conventional SIEM devices are meant to statically examine security incidents and offer warnings for possible cyber threats. Many SIEM systems rely on static threats to report security concerns and track incident reactions. Often, these static rules need more context and expected behaviours. More operational overhead is required to prevent them from providing real-time data analysis capabilities, which are essential for finding advanced and

complex attacks [3]. Timely detection and solving static issues are the solution for prompt cyber-attack resolution. The SOC is key in capturing incident response and proving patterns and results.

II. ML Analytic workflow (MLAW) for threat detection

Traditional monitoring, enhanced by practical and adaptable Machine Learning-based Threat Hunting, is increasingly regarded by modern organizations as an essential component of any Security Monitoring portfolio. Due to the noise in security log data, it is necessary to utilize basic ML models like classification, regression, and forecasting algorithms to find anomalies and novel patterns to address insider and cyber risks to identify sophisticated cyber-attacks. Machine learning-based safety analytics should use efficient workflow to enable effective data pre-treatment before using a suitably competent ML predictor or classifier for subsequent data analytics (see Fig. 1).

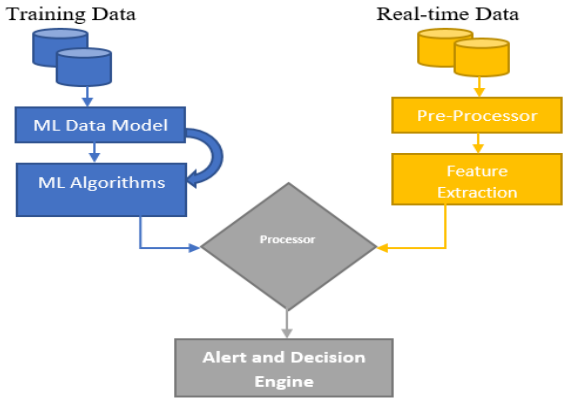

Fig 1. Machine Learning Analytic workflow for threat detection.

A workflow like the one described above may help alleviate these issues by compressing a massive stream of security events into some outliers and providing security analysis with possible pointers of spitefulbehavior to feed into cyber threat detection and hunting procedures. Analysts should select the algorithms that best represent the data while generating the fewest false positives since different types of security events respond well to different types of algorithms [3, 4]. The workflow for ML with the analytical ability to find cyber vulnerabilities is shown in the diagram above.

III. Case Taxonomy of Cyber Threats

Here, we will talk about a few machine learning-based analytics projects that MIT Research has completed successfully to identify cyber threats. The taxonomies developed by MITR Engineering, known as MITRE (ATT&CK, CAPEC, MAEC), serve as a standard reference for cyber threat instances [5].The MITRE ATTACK Framework is a curetted knowledge source that follows threat actors' cyber adversary strategies and approaches throughout the attack lifecycle. The framework is designed to better an organization's security posture rather than just a collection of facts. Some suitable projects like ATT&CK and CAPEC are referenced, discussed, and recommended by Broman deret al.[3].The former describes a wide range of post-compromise approaches, while the latter lists generic sorts of assault patterns throughout the whole cyber-attack life cycle. MAEC's scope is limited since it only includes an explanation of malware-specific standards [5]. Furthermore, Lockheed's guidelines for the pre-compromise and post-compromise phases do not distinguish between the methodology and techniques used at different periods of an assault.

Persistence	Privilege Escalation	Defense Evasion	Credential Access	Discovery
Lateral Movement	Executions	Collection	Exfiltration	Command & Control

Fig 2. MITRE ATT&CK Technique Matrix [5].

The below table (see Table 1) elaborates on a few attacks use cases and feature extraction. These use cases are built based on the MITRE framework and

are very helpful in understanding the real technicality before the models and their usage.

Techniques	Example Use Cases
Data processing and Execution	Anomalous data execution
Patten Discovery	Prediction of threats
Feature Extraction	Data rate analytics and visualization
Exploitation	Parent-child threat analytic

Table 1. Online attack use cases and feature extraction

IV. Logistic Regression Model(LRM) in Machine Learning

Logistic regression is anML approach used mainly for classification in machine learning that employs supervised learning to estimate the likelihood of a target variable. The following Logistic Regression models can be categorized based on their utilization.

1. *Binary Logistic Regression Model (BLRM)*

The most used Logistic Regression model is this one, often known as BLRM. It helps with data categorization into two categories and value prediction for fresh inputs corresponding to one of the two groups. For instance, a cyber attack might either be real or fake—but never both.

2. *Multinomial Logistic Regression Model (MLRM)*

This methodology aids in categorizing target groups into two different categories-independent of mathematical importance. One example is predicting the clothing preferences an individual is likely to order based on their ethnicity, culture, area of origin, and previous experiences.

2. *Ordinal Logistic Regression Model (OLRM)*

This OLRM model is used mainly to classify the target variable. For example, a student's performance on a test may be graded as insufficient, data insufficient, sound, or superb in a categorized manner. As a result, the data is divided into three unique groups, each with its level of relevance [5].

V. Use case of a Logistic Regression to detect cyber-attacks

As per some Chinese scientific reports, companies in China go down by 20 billion Yuan annually due to cybercrimes via malicious URLs [2]. URLs are the foundation of all such online activities. Machine learning (ML) algorithms have drawn the attention of academics in several industries, including healthcare, fintech, weather forecasting, stock market forecasting, bank predictions, anti-money laundering (AML), agriculture, etc. [6].

An ML logic-based algorithm may be instructed on vast amounts of data to discover patterns in data and predict/classify fresh data instances from the same source. In recent years, the features of machine learning algorithms have made them increasingly popular in academics for use in a range of cyber security applications such as intrusion detection, harmful URL classifications, anomaly detection, cyber-attack prediction, cyber crime true or false categorization, bonnet detection, and so on [6].

The proposed model classifies the target URL using Logistic Regression (LR). This method predicts the outcome by combining three essential modules: data exploration, feature extraction, and classification, as shown in Fig. 3.Logistic Regression (LR) is used to classify the target URL. As shown in Fig. 1, this approach consists of four key modules: (i) data gathering and exploration, (ii)feature

extraction,(iii) classification, and the predictionof the outcome.

A. Data Exploration

The public repositories on Git Hub, Kaggle, and Phish Tank may be used to acquire the example training data. The data may then be separated into two groups depending on tags and URLs. It is possible to remove client-specific highlights from the data. Several pieces of information, including the URL, IP address, country name, and URL length, can be used to detect whether a connection is HTTP or HTTPS.

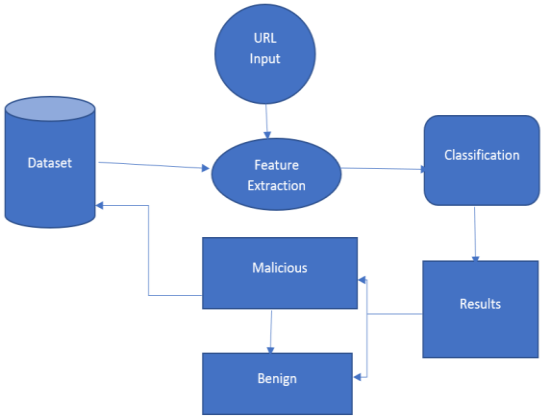

Fig3. Architecture for a Logistic Regression flow

B. Extraction of the Features

Based on their structure and placement, the suggested framework divides all features into three major kinds. Host, domain, and linguistic features are the three categories of features [7].

Host-based features include the country name, host sponsor, and site design and development technologies. Most Jakarta Server Pages (JSP), HTML, and PHP sites prefer to design and construct phishing sites to avoid suspicion. Rules are also based on continents and countries [13]. Governments will have other policies influencing how these sites are classified [7, 13, 14].

$$LR = log\left(\frac{P}{1-P}\right) \quad \text{------(1)}$$

Equation 1. Logistic regression equation where (p/1-p) is the odd ratio.

Hackers and criminal activists purchase hosts from neglectful patron nations with lax cybercrime policies to accomplish this. Researchers can model a non-linear link linearly by applying a logarithmic transformation to the outcome variable [7, 8]. The logistic Regression equation is presented below (see Equation 1).

The most common dynamic malware domain names are *.TK, *.CC, and *.COM. Most of the time, these URL owners let the domain expire [7].The main difference between malicious and benign URLs is that the server section size is more significant in malicious URLs. Paypal.login.itpay.com is one example of a URL that looks to be associated with Paypal and used for financial dealings. However, the top domain name, itpay.com, may not be affiliated with PayPal [7]. As a result, the server and domain length is critical in recognizing fraudulent URLs. Lexical features include a variety of characteristics, including the number of dots in the URL and keywords that indicate whether the URL is dangerous or benign. Popular websites issue warnings regarding such terms to safeguard users from recognized threats.

C. Training and Classification

When there are binary class labels, logistic regression is used. In general, regression is a technique for predicting regression class brands. Linear, multiple, polynomial, and non-parametric regression are more varieties of regression methods. A confusion matrix is used to demonstrate the results of RL, and it is nothing more than a tabular representation of Actual vs. Predicted values [9, 15, 16]. This allows us to determine the model's correctness and

avoid over fitting. Please refer to the figure below for the confusion matrix and how it is helping in cyber threat detection (see Figure 4).For accurate prediction, the confusion matrix provides precise true positives and negatives. The lower the number of false positive or false negative results in a confusion matrix, the better the investigation results. It gives confidence high on the outputs and procedure of the result techniques. The details are shown in the figure below.

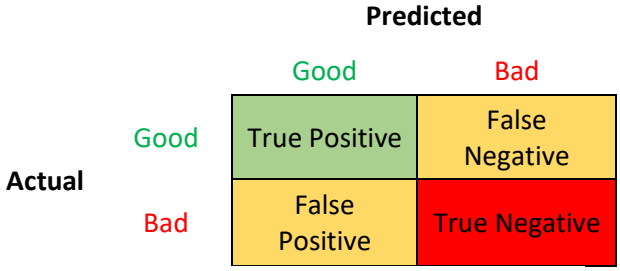

Fig 4. Confusion Matrix showing theaccuracy of the model on cyber threat detection

D. Model Evaluation

The input or interface for performance monitoring is how the data collected from the Logistic Regression (LR) model components and performance statistics are supplied to the data collector. Based on the output and model fit parameters, the ML model evaluator entity analyses the performance monitor (PM) data to decide when and where a new model re-training is required. A threshold-based policy (an update is necessary when PM data surpasses a particular value) or a pattern recognition study can be used to do this [13, 17].

E. Model Update

Keeping a training alert is essential to see whether your model is overfitting or underfitting with prescribed parameters. When a retraining alert is raised, the ML Model Evaluator tells the model orchestrator that a model update is necessary. The ML model trainer restarts the training preparation using an expanded instruction dataset that has been compressed with recently acquired data, depending on the circumstance [13, 14]. A completely different model design is also an option if the previous two options are insufficient. In any case, utilizing correct network data to train models using MLAW may lead to variations in model performance. Therefore, mild retraining on current world data is typically advised before using the suggested models in real-world network situations. The model update is very critical in periodically maintaining the model output better with new data and time.

VI. Recommendation for future work

As remote work becomes more prevalent in the work-from-home culture, cyber dangers are becoming a significant challenge for employers worldwide [10, 11]. Every organization should be concerned about cyber security. However, not computers or technology should be feared; it is the human element, frequently the non-trustworthy shakiest link in a security network. User-specific inaccuracies can occur due to unsafe practices, ignorance of digitized globalization, or dissent with security protocols [12]. Organizations cannot disregard innovation to survive in the global economy, but they must also protect corporate and employee privacy with more robust ML algorithms [14].Furthermore, the above study illustrates the skeleton of applying logistic regression as a case study. Before reaching this point, future researchers must invest critical effort in feature engineering to detect new cyber threats that are constantly changing. Many organizations utilize ML algorithms without understanding their core assumptions or cyber data models. However, now is the opportunity to explore and deploy a solid framework to combat the coming cyber war.

VII. Conclusion

This article gave examples of how Machine Learning analytics may be used to improve cyber security monitoring and research the best algorithms for typical cyber threat situations. ML-based analytics is a powerful tool for delivering context derived from learning security occurrences, resulting in a low probability of false-positive security alarms. Furthermore, machine learning analytics are well-suited to assessing massive quantities of security events and feeding deviations from conventional baselines as indicators or leads of possibly malicious activities into proactive threat-hunting operations. Malicious URL detection is a critical activity that must be thoroughly accomplished before processing cyber data over the Internet. It demonstrates how to use Logistic Regression to detect dangerous URLs. Identifying malicious URLs is a recursive process comprising data gathering, feature extraction, and model training. Finally, the system's performance is evaluated using well-established metrics and frameworks such as accuracy, precision, recall, and false-positive rate.

Reference

[1] Samuel, O. S. (2021). Cyber situation awareness perception model for a computer network. *International Journal of Advanced Computer Science and Applications*, *12*(1). https://doi.org/10.14569/ijacsa.2021.0120147

[2] Olofintuyi, S. S., &Omotehinwa, T. O. (2021). Performance evaluation of supervised ensemble cyber situation perception models for a computer network. *Advances in Multidisciplinary & Scientific Research Journal Publication*, *12*(1), 1–14.https://doi.org/10.22624/aims/cisdi/2021/v12n1p1

[3] Bromander, S., Jøsang, A., &Eian, M. (2016, November). Semantic Cyberthreat Modelling. In *STIDS* (pp. 74-78).

[4] Sharma, P., Dash, B., & Ansari, M. F. (2022). Anti-phishing techniques – a review of Cyber Defense Mechanisms. *IJARCCE*, *11*(7). https://doi.org/10.17148/ijarcce.2022.11728

[5]Farooq, H. M., &Otaibi, N. M. (2018, March). Optimal machine learning algorithms for cyber threat detection. In 2018 UKSim-AMSS 20th International Conference on Computer Modelling and Simulation (UKSim) (pp. 32-37). IEEE.

[6]Mavroeidis, V., &Bromander, S. (2017). Cyber Threat Intelligence Model: An evaluation of taxonomies, sharing standards, and ontologies within Cyber Threat Intelligence. *2017 European Intelligence and Security Informatics Conference (EISIC)*. https://doi.org/10.1109/eisic.2017.20

[7] *Logistic Regression for Machine Learning: A complete guide*. upGrad blog. (2022, April 18). Retrieved September 22, 2022, from https://www.upgrad.com/blog/logistic-regression-for-machine-learning/

[8] Chiramdasu, R., Srivastava, G., Bhattacharya, S., Reddy, P. K., & Reddy Gadekallu, T. (2021). Malicious URL detection using logistic regression. *2021 IEEE International Conference on Omni-Layer Intelligent Systems (COINS)*. https://doi.org/10.1109/coins51742.2021.9524269

[9] Dash, B., & Ansari, M. F. (2022). An Effective Cybersecurity Awareness Training Model: First Defense of an Organizational Security Strategy.

[10] Ansari, M. F., Sharma, P. K., & Dash, B. (2022). Prevention of Phishing Attacks Using AI-Based Cybersecurity Awareness Training. Prevention.

[11] Dash, B. (2022). Remote work and innovation during this covid-19 pandemic: An employers' challenge. *International Journal of Computer Science and Information Technology*, *14*(2), 13–18. https://doi.org/10.5121/ijcsit.2022.14202

[12] Ansari, M. F. (2022). A quantitative study of risk scores and the effectiveness of AI-based Cybersecurity Awareness Training Programs. *International Journal of Smart Sensor and Adhoc Network.*, 1–8. https://doi.org/10.47893/ijssan.2022.1212

[13] Singh, S. K., Singh, R., &Kumbhani, B. (2020, April). The evolution of radio access network towards open-RAN: Challenges and opportunities. In 2020 IEEE Wireless Communications and Networking Conference Workshops (WCNCW) (pp. 1-6). IEEE.

[14] Giannopoulos, A., Spantideas, S., Kapsalis, N., Gkonis, P., Sarakis, L., Capsalis, C., ... &Trakadas, P. (2022). Supporting Intelligence in Disaggregated Open Radio Access Networks: Architectural Principles, AI/ML Workflow, and Use Cases. IEEE Access, 10, 39580-39595.

[15] Gorski, E. G., Loures, E. D. F. R., Santos, E. A. P., Kondo, R. E., & Martins, G. R. D. N. (2022). Towards a smart workflow in CMMS/EAM systems: An approach based on ML and MCDM. Journal of Industrial Information Integration, 26, 100278.

[16] Chahal, D., Ojha, R., Choudhury, S. R., & Nambiar, M. (2020, April). Migrating a recommendation system to the cloud using my workflow. In Companion of the ACM/SPEC International Conference on Performance Engineering (pp. 1-4).

[17] Dash, B. (2021). A hybrid solution for extracting information from unstructured data using optical character recognition (OCR) with natural language processing (NLP).