Machine Learning for Human Intelligence
2024, Vol. 2, Issue 1

Research Article

Model for Predicting Cyber Threats Utilizing Advanced Data Science Techniques

Sania Amin[1,⋆] and Aleena Nawaz[2]

[1,2]Department of Computer Science, Emerson University Multan, Pakistan; Email: saniaamin2356@gmail.com , aleenanawazem@gmail.com
⋆Corresponding author: Sania Amin (saniaamin2356@gmail.com)

Article History

Academic Editor:
Dr. Azhar Imran
Submitted: December 19, 2023
Revised: January 27, 2024
Accepted: March 1, 2024

Keywords:
Machine Learning; Network Security; Cyber Threats; Data Science; Predictive Model

Abstract

In the technological age and amidst the pervasive utilization of the internet, the data and personal information of internet users are increasingly vulnerable. Among numerous cyber-attacks, DDoS is one of the most perilous, employing single or multiple targets to render resources unavailable on both small and large scales. The frequency and severity of cyber-attacks are progressively rising alongside the growing use of the internet. Defensive measures are developed over time to safeguard a network and its devices from various breaches and attacks perpetrated by cyber terrorists. Data science enhances the prediction and detection of cyber-attacks, beyond conventional protection measures. This work suggested a data science-driven predictive model utilizing a significant dataset, CICDDOS2019. This research employs various Machine Learning models, including Decision Tree, Random Forest, SVM, and Naïve Bayes, after the dataset's cleansing and the selection of optimal relevant characteristics to achieve maximum accuracy in detecting and predicting cyber risks.

1 Introduction

As internet-connected computer systems become increasingly prevalent, the International Telecommunication Union (ITU) reported a significant number of global internet users who consistently utilize online services. For instance, e-commerce, e-banking, entertainment, and education have expanded by billions. The risk of data breaches has increased due to the proliferation of virtual networks and internet-connected gadgets. Cyber terrorists are highly engaged in the theft and exploitation of data, information, and credentials from users of active internet-based devices. Substantial amounts of sensitive user data are susceptible to numerous attacks, both internal and external. Cyberattacks have become increasingly sophisticated as algorithms have advanced due to technological progress [1]. The escalating frequency of cyber-attacks surpasses the development of efficient defenses, necessitating firms to augment their expenditure in cybersecurity. Recent methodologies for defensive cyber-attacks encompass machine learning strategies, policy-driven frameworks, and dynamic, rule-based systems [2]. The phrase "cybersecurity" refers to the protection of data and information transmitted via the internet.

Cybersecurity is a suite of technological advancements aimed at thwarting cyberattacks, mitigating damage, and preventing unwanted access to computers, networks, apps, and data. Cybersecurity has had significant advancements in recent years, both in technology and its operational dynamics inside the computing environment. Data science is crucial in a novel scientific prototype [3, 4] that employs machine learning to transform cybersecurity. Conventional firewall solutions are insufficient to guarantee the security of open ports in any internet-based communication system. An Intrusion Detection System (IDS) is typically designated to categorize and identify diverse cyber-attacks.

2 Motivation

Numerous cyber attacks consistently jeopardize important information security. As technology advances, cyber terrorists are increasingly active, utilizing diverse innovative techniques to compromise network security and exploit confidential and sensitive information [14]. Security experts propose several methods to counter contemporary cyber-attack methodologies; yet, these measures are insufficient due to the exponential increase of data and cyber threats. Additional resources and contemporary techniques must be utilized and executed to evade cyber-attacks. Cybersecurity researchers are actively monitoring to anticipate and address emerging threats. Research from 2016 indicates that roughly 95 percent of cyber intrusions target the government, retail, and technology sectors. These management strategies are directed towards two primary objectives [5, 6].

An earnest effort has been devoted to investigating IDS systems that rely on a machine learning methodology for dataset evolution. The classification of events as either harmful or benign is the fundamental requirement for this study, relying on the labeled datasets [7]. The imbalance of datasets is a critical issue during the dataset creation phase. The bias is attributable to the data employed in training the machine learning model. Despite the enormous nature of machine learning research, a mere fraction of publications scrutinize the particulars of the data employed in their studies. For academics, prioritizing the development of authentic and intricate models supersedes the observation of patterns within datasets. Nonetheless, it remains accurate that nearly all substantial datasets generated using machine learning algorithms exhibit bias [8]. Researchers are increasingly acknowledging the necessity for proportionate datasets in machine learning to mitigate bias. When dangerous samples are fewer than benign samples, there exists a greater diversity of classes; an imbalanced dataset leads to low encounter rates of less prevalent classes, where the dataset contains more instances related to the benign class than to the malicious class. Data destruction and overfitting are two instances of how this diminishes performance. All courses will be accurately identified when learners opt for the bulk class [9, 10]. Therefore, the objective of this work is to devise methods that mitigate the impact of bias in datasets while evaluating efficiency.

3 Problem Statement

Consequently, it necessitates regular updates to incorporate new signatures into the product database. Consequently, zero-day attacks are undetectable and unavoidable. Moreover, traditional methods are overly binary and offer limited advantages compared to predictive models that can estimate the probability of assaults or dangerous behaviors based on data analysis techniques. Access to extensive data facilitates the resolution of complex security issues. In the context of big data and data mining, an increase in data collection correlates with enhanced precision and accuracy in analysis [11]. Data science, in its most comprehensive definition, is employing a scientific methodology to derive insights from data and uncover new knowledge. By utilizing novel technological advancements in storage, processing, and behavioral analytics, data science may assist in the development of innovative cybersecurity solutions [12]. Thus, Data Science is essential to cybersecurity as it depends on data and high-performance computation to counteract cybercrime and protect consumers. An effective data science project necessitates a proficient approach that addresses all challenges and allocates resources appropriately within budgetary constraints [13].

4 Related Work

In recent years, researchers have devised innovative techniques for identifying assaults. They tackled their specific study deficiencies by employing diverse data sources, detecting techniques, and methodologies. This section will examine the most recent advancements in this domain [20, 21, 22]. As attack detection methodologies have evolved in response to the complexity of modern attack generation.

4.1 Approaches Based on DDoS Attacks

Radial Basis Function (RBF) neural networks represent an innovative category of neural networks developed to identify DDoS attacks based on packet attributes. This strategy can be applied to routers located at the periphery of a target network. Seven feature vectors were employed to activate an RBF neural network at each subsequent time step [23, 24, 25]. The RBF neural network categorizes inputs into two classifications: standard and attack types [26]. The Filtering and Attack Alarm Modules obtain the source IP addresses of attack packets when the incoming traffic is identified as malicious. If the traffic is deemed normal, it will be directed to its destination. [27, 28] delineates a data-mining methodology for detecting DDoS attacks. The authors employed an FCM clustering technique and an a priori association strategy to extract models of network traffic and network packet protocol status, thereby establishing the threshold for the detection model. The authors employed decision trees and gray relational analysis [29] to detect DDoS attacks. Fifteen criteria can be employed to categorize an attack, including the assessment of incoming and outgoing packet and byte rates, as well as the compilation of TCP SYN and ACK flag rates to characterize traffic flows. The decision tree technique was employed to assess the detection of normal traffic flow through these features.

Various DDoS attack methodologies have been introduced in the literature during the past decade. The protocol level of a DDoS flooding attack has been examined in greater detail and categorized into two types [30]. TCP, UDP, ICMP, DNS, and ICMP protocol packets are frequently employed to execute DDoS flooding attacks on a network or transport layer. DDoS floods assault the application layer by exhausting resources like sockets, RAM, CPU, and bandwidth, hence disrupting legitimate user services. Detecting application-level incursions is more difficult than identifying volumetric ones, as they resemble normal traffic. The prompt mitigation of assaults at their origin is a critical concern in combating DDoS attacks; nevertheless, a comprehensive solution that addresses these needs has not yet been realized [15, 30, 31]. Information concealing seeks to obscure a clandestine message within innocuous media, referred to as cover media, allowing only the authorized recipient to retrieve the hidden message using the specified secret key. Information concealment can be broadly categorized into two types: steganography and watermarking. The former seeks to incorporate extensive information into a cover signal, whereas the latter prioritizes the resilience of the embedded information, sacrificing embedding capacity.

A data sampling-based flood attack detection method for web servers was developed utilizing a Hypertext Transfer Protocol (HTTP) methodology [32]. The quantity of requests originating from the application layer and the overall count of packets devoid of payload were utilized to evaluate whether the scrutinized traffic was normal or a target of a DDoS attack. The research findings indicate that a 20% sample rate had a detection rate ranging from 80% to 88%. Nevertheless, despite considerable progress, the proposed approach remains unsuitable for implementation in automated detection systems at this time.

D-FACE is a formidable collaborative system that employs metrics from GE and GID (FEs) in the context of DDoS attacks and Flash Events. Although the research achieves notable progress in the field, the validation depended on obsolete datasets. The proposed collaboration model restricts the practical applicability of the solution due to the requirement for extensive ISP involvement [18].

SkyShield uses the divergence between two Sketches to detect anomalies generated by attackers during the identification step. The mitigation phase safeguards users by filtering, whitelisting, blacklisting, and CAPTCHA. The system was evaluated using bespoke datasets.

Umbrella [33] establishes a stratified defensive framework designed to alleviate a wide spectrum of DDoS attacks. The authors introduced a method focused exclusively on victim detection and protection. The system was evaluated regarding traffic management utilizing the constructed testbed. The authors claim that the system can safeguard itself against large-scale attacks. However, this technique is widely utilized in commercial settings and proves ineffective against substantial DDoS attacks.

A recently developed semi-supervised machine learning technique classifies DDoS attacks. The CICIDS2017 dataset was employed to evaluate the system's performance metrics in this approach [34]. Although the work addresses contemporary DoS vectors, the method's online efficacy remains untested.

5 Proposed Architecture

This section delineates the entire framework for the proposed architecture, which is segmented into the following stages in accordance with the data science process.

5.1 Decision Tree

Decision tree analysis is a predictive modeling technique applicable in diverse scenarios. Decision trees can be constructed by an algorithmic approach that categorizes data based on multiple criteria. Decision trees are the most potent algorithms in the domain of supervised learning. They are appropriate for both classification and regression tasks. This research utilized 80% of the data to train the Decision Tree, employing a standardized dataset for the classification and regression of various DDoS attacks [17, 35].

5.2 Random Forest

A Random Forest encompasses nearly identical hyperparameters to a Decision Tree, in addition to those of a bagging classifier that governs the ensemble of trees. During node partitioning, a random selection of features aims to identify the optimal feature instead of determining the most significant one [16, 35].

5.3 Support Vector Machine

Support vector machines (SVMs) are a robust and adaptable supervised machine learning technique. They serve purposes in both classification and regression. Nevertheless, they are frequently utilized in categorization matters. Support Vector Machines (SVMs) were introduced in the 1960s and underwent enhancements in 1990. Support Vector Machines are executed distinctively compared to other machine learning methods. They have gained popularity in recent years due to their ability to manage numerous continuous and categorical variables.

5.4 Naive Bayes

The core principle of Naive Bayes is that each feature must be either independent or uncorrelated with commensurate contribution to the result. Naive Bayes is employed for supervised classification utilizing accessible features; hence, given the presence of many features in this research, the formula incorporating multiple feature variables will be as follows:

6 Implementation

The aforementioned model stages are delineated below to elucidate the execution of those processes for the proposed model.

6.1 Dataset Analysis

The downloaded dataset comprises ten CSV files that encompass diverse assault types, with a specific proportion from each file incorporated into this model.

The amalgamation of all dataset files resulted in the formation of a substantial dataset for this model, comprising 1,678,441 instances and 88 features. The utilized dataset comprises the specified quantities of various attacks and benign traffic as detailed. It is among the most often utilized instruments for exploratory data analysis and predictive modeling. It is a method for eliminating dominant patterns from a dataset by reducing variances.

6.2 Manual Feature Reduction

In this phase of manual feature reduction, features or columns from the CICDDOS2019 dataset are systematically eliminated to identify the most relevant attributes of the dataset. This is conducted based on the following parameters:

- Eliminate the columns that contain exclusively zero or null entries.
- Eliminate columns with above 70% null data.
- Eliminate the columns with negative or infinite values.

6.3 Data Splitting for Training and Testing

The data is divided into training and testing components at this phase. The data is divided in an 80:20 ratio. Eighty percent of the data is allocated for model training, whereas twenty percent is designated for model testing.

7 Results and Evaluation

All findings in this study are predicated on three parameters: results following chi-square analysis, results excluding chi-square analysis, and outcomes subsequent to manual reduction.

7.1 System Specifications

7.2 Accuracy Metrics

The precision of each method is detailed below in Table 1.

Table 1: Precision Results

Models	**After Chi-square**	**Without Chi-square**	**Manual Reduction**
Random Forest	99.327%	99.829%	99.871%
Decision Tree	91.464%	99.807%	99.844%
SVM	92.395%	74.792%	66.046%
Naive Bayes	66.842%	88.779%	87.705%

8 Discussion

Following the aforementioned preprocessing procedures of the dataset, it is evident that the accuracy of the various models employed in this research is significantly influenced by feature selection and the quality of the data. Manual reduction of dataset features based on null and negative values enhances model accuracy; nevertheless, this process is labor-intensive and complicates the identification of useful and informative features. Thus, there exists a tradeoff: enhancing accuracy necessitates additional time and effort.

Various contemporary feature selection and extraction techniques facilitate the selection and extraction of features or columns; nevertheless, this may impair the model's accuracy. The aforementioned accuracy table reveals observations. This research presently concentrates on a particular type of cyber-attack and develops predictive models. The future focus will be on expanding the scope of cyber-attack classifications. A model must be developed for predicting various sorts of cyber-attacks. The extraction and selection of features will occur automatically; nonetheless, the model's accuracy must remain intact.

9 Conclusion

The research concludes that a model utilizing Data Science and Machine Learning is significantly more effective and suitable for cyber security, especially in predicting cyberattacks. This prediction is more precise and automated, significantly enhancing the field of information security. This analysis reveals that data pre-processing is the essential and most critical element in the development of any model utilizing data science approaches. Open-source datasets comprise substantial volumes of data and encompass much raw information, hence it is straightforward to employ any dataset once it has been pre-processed to align with any machine learning model, such as the one utilized in our study. Consequently, the research concludes in this chapter. It presents an analysis of the findings and recommendations for additional research and development.

References

[1] Vinayakumar, R., Alazab, M., Soman, K., Poornachandran, P., Al-Nemrat, A., & Venkatraman, S. (2019). A deep learning methodology for an intelligent intrusion detection system. *IEEE Access*, 7, 41525-41550.

[2] Liu, M., Xue, Z., Xu, X., Zhong, C., & Chen, J. (2018). A review of host-based intrusion detection systems utilizing system calls and prospective trends. *ACM Computing Surveys (CSUR)*, 51(5), 1-36.

[3] Tansley, S., & Tolle, K. M. (Eds.). (2009). *The fourth paradigm: Data-centric scientific exploration* (Vol. 1). Microsoft Research.

[4] Cukier, K. (2013). Data is ubiquitous: A comprehensive analysis of information management. *The Economist.*

[5] Juniper Research. (2018). Cybersecurity breaches are projected to rise by nearly 70% over the next five years. Retrieved from https://www.juniperresearch.com/press/business-losses-cybercrime-data-breaches

[6] CybInt Solutions. (2018). 15 disturbing cybersecurity facts and statistics. Retrieved from https://www.cybintsolutions.com/cyber-security-facts-stats/

[7] Laskov, P., Düssel, P., Schäfer, C., & Rieck, K. (2005, September). Learning intrusion detection: Supervised or unsupervised? In *International Conference on Image Analysis and Processing* (pp. 50-57). Springer.

[8] Krishnamurthy, P. (2018). Understanding data bias in data science. Retrieved from https://towardsdatascience.com/survey-d4f168791e57

[9] Karatas, G., Demir, O., & Sahingoz, O. K. (2020). Enhancing the efficacy of machine learning-based intrusion detection systems on an imbalanced and current dataset. *IEEE Access*, 8, 32150-32162.

[10] Subba, B., Biswas, S., & Karmakar, S. (2016, March). A neural network-based system for intrusion detection and attack classification. In *Twenty-Second National Conference on Communication (NCC)* (pp. 1-6). IEEE.

[11] Mastrogiacomo, R. (2017). The conflict between data science and cybersecurity. Retrieved from https://www.information-management.com/opinion/the-conflict-between-data-science-and-cybersecurity

[12] Pegna, D. L. (2016). Developing cognitive cybersecurity. Retrieved from https://www.computerworld.com/article/2881551/creating-cyber-security-that-thinks.html

[13] Foroughi, F., & Luksch, P. (2018). Methodology of data science for cybersecurity initiatives. *arXiv preprint arXiv:1803.04219.*

[14] Jang-Jaccard, J., & Nepal, S. (2014). An analysis of nascent dangers in cybersecurity. *Journal of Computer and System Sciences*, 80(5), 973-993.

[15] Mukkamala, S., Sung, A., & Abraham, A. (2005). Cybersecurity challenges: Developing effective intrusion detection systems and antivirus software. In V. R. Vemuri (Ed.), *Enhancing Computer Security with Smart Technology* (pp. 125-163). Auerbach.

[16] Champion, A. C. (2016). Threats and attacks. In *CSE 4471: Information Security.*

[17] Fischer, E. (2016). Cybersecurity issues and challenges: A brief overview. Congressional Research Service Report. Retrieved from https://sgp.fas.org/crs/misc

[18] Sun, N., Zhang, J., Rimba, P., Gao, S., Zhang, L. Y., & Xiang, Y. (2018). Data-driven cybersecurity incident prediction: A survey. *IEEE Communications Surveys & Tutorials*, 21(2), 1744-1772.

[19] Jafarian, J. H., Al-Shaer, E., & Duan, Q. (2015). A proficient address mutation strategy for thwarting reconnaissance attacks. *IEEE Transactions on Information Forensics and Security*, 10(12), 2562-2577.

[20] Haider, W., Hu, J., Slay, J., Turnbull, B. P., & Xie, Y. (2017). Generating realistic intrusion detection system dataset based on fuzzy qualitative modeling. *Journal of Network and Computer Applications*, 87, 185-192.

[21] Creech, G., & Hu, J. (2013). Generation of a new IDS test dataset: Time to retire the KDD collection. In *2013 IEEE Wireless Communications and Networking Conference (WCNC)* (pp. 4487-4492). IEEE.

[22] Zargar, G. R., & Kabiri, P. (2009, July). Identification of effective network features for probing attack detection. In *2009 First International Conference on Networked Digital Technologies* (pp. 392-397). IEEE.

[23] Ram, P., & Ratley, D. (1996). Satan: A double-edged sword. *Computer*, 29(4), 19-20.

[24] Gadge, J., & Patil, A. A. (2008, December). Port scan detection. In *2008 16th IEEE International Conference on Networks* (pp. 1-6). IEEE.

[25] Hartley, D., & Clarke, J. (2009). What is SQL injection? In *SQL Injection Attacks and Defense.* O'Reilly Media.

[26] Su, Z., & Wassermann, G. (2006). The fundamental nature of command injection attacks in web applications. *ACM SIGPLAN Notices*, 41(1), 372-382.

[27] Kapetanovic, D., Zheng, G., & Rusek, F. (2015). An overview of physical layer security in massive MIMO, focusing on passive eavesdropping and active attacks. *IEEE Communications Magazine*, 53(6), 21-27.

[28] Salem, M. B., & Stolfo, S. J. (2011). Detection of masquerade attacks utilizing search behavior. In *Recent Advances in Intrusion Detection* (pp. 181-200). Springer.

[29] Nikiforakis, N., Meert, W., Younan, Y., Johns, M., & Joosen, W. (2011, February). SessionShield: Minimalist defense against session hijacking. In *International Symposium on Engineering Secure Software and Systems* (pp. 87-100). Springer.

[30] McIntosh, T., Jang-Jaccard, J., Watters, P., & Susnjak, T. (2019, December). The insufficiency of entropy-based ransomware detection. In *International Conference on Neural Information Processing* (pp. 181-189). Springer.

[31] Jang-Jaccard, J., & Nepal, S. (2014). An examination of nascent threats in cybersecurity. *Journal of Computer and System Sciences*, 80(5), 973-993.

[32] Alazab, M., Venkatraman, S., Watters, P., & Alazab, M. (2010). Detection of zero-day malware with supervised learning algorithms for API call signatures. In *Proceedings of the Australasian Data Mining Conference.*

[33] Bilge, L., & Dumitraş, T. (2012, October). An empirical investigation of real-world zero-day attacks. In *Proceedings of the 2012 ACM Conference on Computer and Communications Security* (pp. 833-844).

[34] Alam, M. F., Singla, P., & Phursule, R. N. (2022, April). Development of a detection system utilizing deep learning algorithms for network attack identification. In *2022 IEEE 7th International Conference on Convergence in Technology (I2CT)* (pp. 1-5). IEEE.

[35] Masood, R., & Anwar, Z. (2011, December). Analyzed the Stuxnet worm using Metasploit. In *2011 Frontiers of Information Technology* (pp. 142-147). IEEE.