iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** I **Month of publication:** January 2024

DOI: https://doi.org/10.22214/ijraset.2024.57786

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue I Jan 2024- Available at www.ijraset.com

Role of Machine Learning in Cybersecurity: Techniques and Challenges

Prateek Rana[1], Himanshu Tiwari[2], Nikhil Kumar Singh[3]
[1, 3]*Student, Dept of Computer Science, GNIOT, Greater Noida, India*
[2]*Asst. Professor, Dept of Computer Science, GNIOT, Greater Noida, India*

Abstract: ***Machine learning (ML) plays a crucial role in cybersecurity across various aspects of threat detection, prevention, and response. Here are specific roles and applications of machine learning in cybersecurity. Machine Learning (ML) represents a pivotal technology for current and future information systems, and many domains already leverage the capabilities of ML. However, deployment of ML in cybersecurity is still at an early stage, revealing a significant discrepancy between research and practice. Such a discrepancy has its root cause in the current state of the art, which does not allow us to identify the role of ML in cybersecurity. The full potential of ML will never be unleashed unless its pros and cons are understood by a broad audience. In the computer world, data science is the force behind the recent dramatic changes in cybersecurity's operations and technologies. The secret to making a security system automated and intelligent is to extract patterns or insights related to security incidents from cybersecurity data and construct appropriate data-driven models. Data science, also known as diverse scientific approaches, machine learning techniques, processes, and systems, is the study of actual occurrences via the use of data. Recent breakthroughs in Machine Learning (ML) methods promise new solutions to each of these infamous diversification and asymmetric information problems throughout the constantly increasing vulnerability reporting data-bases. Due to their varied methodologies, those procedures themselves display varying levels of performance. The authors provide a method for cognitive cybersecurity that enhances human cognitive capacity in two ways. To create trustworthy data sets, initially reconcile competing vulnerability reports and then pre-process advanced embedded indicators***

Keywords: Security, Machine Learning, Survey, Machine Learning, Intrusion Detection, Spam Cybersecurity.

I. INTRODUCTION

In the current era of computing devices, most of the devices that we are using are connected to the Internet in an Internet of Things (IoT) environment. These devices share and transmit their data through the insecure (open) communication medium, also called as the Internet. Most of the time this data is sensitive in nature (i.e., healthcare data, banking data, insurance data, other finance related data, and social security numbers). The malicious entities, such as the online attackers (hackers) are always in search of that, where they play with the things (for example, they can launch attacks, like replay, man-in the-middle, impersonation, credential guessing, session key computation, malware injection and data modification) [1,2]. Therefore, from time-to-time several researchers propose different security protocols to mitigate these attacks protocols or cyber security protocols can be divided into different categories: “authentication protocols”, “access control protocols”, “intrusion detection protocols”, “key management protocols”, and “blockchain enabled security protocols”. The advent of technologies ranging from smartphones to large-scale communication systems has resulted in an exceptionally digital interconnected society and humongous usage of the internet. It is estimated that there are more than 6 billion smart devices and 3.5 billion internet users in the world as of today. This cyber connectivity is widely being used in a diverse set of applications, such as online banking and shopping, email, documents or critical information sharing, video chatting, and gaming, to name a few. Consequently, lots of data, in terabytes per second, are being created, processed, exchanged, and stored by different applications as well as the Internet of Things (IoT). In fact, it is believed that 92% of the data in the world today has been generated in the last two years alone.

Although cyber-attacks do not use any physical weapons, they are the most dangerous and harmful weapons that may cause revelation of the topmost classified information of government organizations through espionage or sensitive personal information through to phishing. According to cybersecurity experts, just in 2018 cyber-attacks might have caused US$5 billion worth of damage and will grow in the future. The development of modern technology makes it possible to communicate effectively across every field; specifically, the Cyber Physical System (CPS) is a cutting-edge system that provides a more efficient environment for data sharing and transmission from one endpoint to another via various proper communication channels

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue I Jan 2024- Available at www.ijraset.com

II. MACHINE LEARNING IN CYBER SECURITY

Machine learning techniques are playing a major role in fighting against cybersecurity threats and attacks such as intrusion detection system, malware detection, phishing detection, spam detection, and fraud detection to name a few more. We will focus on malware detection, intrusion detection system, and spam classification for this review. Malware is a set of instructions that are designed for malicious intent to disrupt the normal flow of computer activities. Malicious code runs on a targeted machine with the intent to harm and compromise the integrity, confidentiality and availability of computer resources and services. Saad et al. in discussed the main critical problems in applying machine learning techniques for malware detection. Sad et al. argued that machine-learning techniques have the ability to detect polymorphic and new attacks. Machine learning techniques will lead to all other conventional detection methods in the future. The training methods for malware detections should be cost-effective. The malware analysts should also be able to keep with the understanding of ML malware detection methods up to an expert level. One of the critical downsides of the security system is that the security reliability level of the computing resources is generally determined by the ordinary user, who does not possess technical knowledge about security.

The use of machine learning techniques has grown in value, allowing for more efficient threat detection and response. The approaches used to use machine learning in cybersecurity are thoroughly reviewed in this article, with emphasis on their advantages, disadvantages, and practical applications

A. Data Collection and Preprocessing

Data collection and preprocessing are fundamental steps in the machine learning pipeline. The quality and suitability of the data you use greatly impact the performance and generalization of your machine learning models. Here's an overview of these two crucial stages:

1) *Data Collection:* Identify the sources from which you will collect data. This could be databases, APIs, files, sensors, or any other relevant sources**.** Gather the data from the identified sources. This may involve web scraping, querying databases, downloading datasets, or setting up data collection systems. Check for data quality issues such as missing values, outliers, and inaccuracies. Cleaning the data at this stage is crucial for accurate model training.
2) *Data Preprocessing:* Decide on a strategy for handling missing data, whether through imputation, deletion of rows/columns, or other methods. Identify and handle outliers appropriately. This may involve removing them, transforming them, or treating them separately. Convert data into a suitable format for modelling. This may include scaling numerical features, encoding categorical variables, or transforming data distributions.

B. Model Selection

Model selection and evaluation are critical steps in the machine learning process, helping you choose the most suitable algorithm for your problem and assess its performance.

1) Clearly define the problem you are trying to solve. Is it a classification, regression, clustering, or another type of problem.
2) Based on the nature of your problem, select a few candidate machine learning algorithms. Consider factors such as the size of your dataset, the type of data, and the interpretability of the model.
3) Implement and train the selected models using a portion of your dataset. Start with default hyperparameters.
4) Assess the initial performance of each model using appropriate metrics. For classification, this could include accuracy, precision, recall, F1 score, etc. For regression, metrics like mean squared error or mean absolute error are common.
5) Compare the performance of different models and select the one that shows the most promising results. Keep in mind that model selection is an iterative process.

C. Model Evaluation

1) Use cross-validation techniques (e.g., k-fold cross-validation) to get a more robust estimate of your model's performance. This helps ensure that your model generalizes well to new, unseen data.
2) Choose appropriate evaluation metrics based on the nature of your problem. For classification, you might use precision, recall, and ROC-AUC, while for regression, metrics like R-squared and mean absolute error may be more relevant.
3) Check for signs of overfitting, where the model performs well on the training data but poorly on new data. Adjust the model complexity or use regularization techniques to mitigate overfitting.
4) Once satisfied with the model's performance on the validation set, evaluate it on a separate test set that the model has not seen during training or validation. This provides a final measure of how well your model generalizes to new data.

D. Deployment and Integration

Deployment and integration are crucial steps in bringing a machine learning model into practical use within real-world applications.

1) Serialize your trained machine learning model into a format that can be easily stored or transmitted. Common formats include pickle for Python-based models or ONNX (Open Neural Network Exchange) for interoperability across different frameworks.
2) Select the deployment environment based on your application requirements. This could be on-premises servers, cloud platforms (e.g., AWS, Azure, Google Cloud), or edge devices.
3) Consider containerizing your model using technologies like Docker. This makes it easier to package the model, its dependencies, and the inference code into a single, reproducible container.
4) Deploy the containerized model to a server or a cloud-based service. Many cloud platforms provide specific services for deploying machine learning models, such as AWS Sage Maker, Azure ML, or Google AI Platform

III. DIFFERENT BLOCKCHAIN ENABLED SECURITY PROTOCOLS

1) *Authentication Protocols:* Authentication is a process of checking the genuineness (authenticity) of someone of some device. It can be performed through some credentials or factors (i.e., username, password, smartcard, biometrics), which are closely associated with the users or device. We can have user to user authentication, user to device authentication or device to authentication. On the basis of available factors, user authentication protocols can be again divided into three categories, i.e., one-factor user authentication protocol, two-factor user authentication protocol and three-factor user authentication protocol.
2) *Access Control Protocols:* Access control is a process of putting restrictions on the unauthorized access of someone or some device(s). Users or devices can access the other users or devices in a secure way after the completion of all steps of a user/device access control protocol.
3) *Intrusion Detection Protocols:* Signature-based intrusion detection relies on predefined patterns or signatures of known attacks. These signatures are created based on the characteristics of known malicious activities. Protocol/Tool: Snort is a widely used open-source intrusion detection system that employs signature-based detection. Anomaly-based intrusion detection involves establishing a baseline of normal network or system behaviour and then flagging any deviations from this baseline as potential intrusions. Honeypots are decoy systems or services designed to attract and detect attackers. They can be used to study attack techniques and gather information about potential threats.
4) *Blockchain Enabled Security Protocols:* Blockchain is one of the emerging technologies of the era. Blockchain maintains data in the form of certain blocks, which are chained together with some hash values. In blockchain data is maintained in the form of distributed ledger, which is named as distributed ledger technology (DLT). All the genuine parties (sometimes miner) of the network have access to the DLT. The data that we store over the blockchain safe and secured against the various possible cyber-attacks.
5) *Key Management Protocols:* Key management protocols are used for secure key management among the various entities, such as some devices (for example, smart Internet of Things (IoT) devices and smart vehicles) and some users (smart home user, doctor, traffic inspector). Usually, a trusted registration authority does the registration of all entities of the communication system and then stores the secret credentials (i.e., secret keys) in their memory. We need a key management process for the fresh keys' generation and their storing in the devices, key establishment and key revocation purposes.

IV. ADVANTAGES OF UNITING CYBER SECURITY AND MACHINE LEARNING

Both cyber security and machine learning are essential for each other and can improve their mutual performances. Some of the advantages of their uniting are as follows.

1) *Full Proof Security of ML Models:* As discussed earlier, the ML models are vulnerable to various attacks. The occurrence of these attacks may affect the working, performance and predictions of the ML models. However, these unwanted incidences can be secured through the deployment of certain cyber security mechanisms. Under the deployment of cyber security mechanisms, the working and performance and the input datasets of the ML models become secured and we get the correct predictions and results.
2) *Improved Performance of Cyber Security Techniques:* When we use the ML algorithms in the cyber security schemes (i.e., intrusion detection systems) that improve their performances (i.e., improved accuracy and detection rate with less false positive rate). ML techniques, like supervised learning, unsupervised learning, reinforcement learning and deep learning algorithms can be used as per the communication environment and the associated systems.

3) *Effective Detection of zero-day Attacks:* The cyber security methods, which detect the intrusion through the ML models seem very effective for the detect of zero-day attacks (i.e., unknown malware attacks). It happens because they perform the detection with the help of some deployed ML models. The ML models work through collection and matching of certain features, if the features of a program match with the malicious program's features, then that can be considered as the malicious program. This detection task can be performed by the ML models automatically. Thus, detection of zero-day attacks can be performed effectively with the uniting of cyber security and machine learning.
4) *Quick Scanning and Mitigation:* The ML based intrusion detection systems work very efficiently to detect the presence of the attacks because they work through certain ML algorithms. Therefore, uniting of machine learning with the cyber security systems performs the scanning of intrusions very fast and also provide fast response in case of any sign of intrusion. The only thing that we need to take care is the suitable ML algorithm selection.

V. ISSUES AND CHALLENGES OF UNITING OF CYBER SECURITY AND MACHINE LEARNING

1) *Compatibility Issues:* The uniting of cyber security and machine learning contains different types of security techniques (i.e., encryption algorithms, signature generation and verification algorithms, hashing algorithms) and machine learning algorithms (clustering, classification, convolutional neural networks (CNNs)). Moreover, the data, which is the main input for analysis process comes from the different sources i.e., IoT devices. These IoT devices are operated through different communication techniques. During the amalgamation of these many algorithms, there may be the issues related to the compatibility.
2) *Overloading:* In uniting cyber security and machine learning, we use various algorithms as discussed earlier. For the execution of such algorithms, we need the resources in extra amount. Otherwise, the system will not work properly. Therefore, the amalgamation and use of various algorithms may cause the overloading to the system that may further affect the actual working of the system. For example, we cannot allocate entire resources of the system for the security related processes. We also need some resources for the execution of ML-related tasks. Hence, we should select the algorithms wisely and as per the resources of the communication environment.
3) *Accuracy:* In the uniting of cyber security and machine learning, we use various ML mechanisms i.e., machine learning (ML) models to predict about some physical phenomena (i.e., chances of roadside accident in the intelligent transportation system). The ML models work with the help of certain datasets, if we have some error in the dataset or in the settings of the ML model then this can give big trouble.
4) *Flaws in Security Mechanisms:* In the uniting of cyber security and ML, we may use various cyber security mechanisms. If these mechanisms have some flaws, it may then cause the trouble to the security to the system. Most of the time, the hackers try to search for the zero-day vulnerabilities and then exploit them.

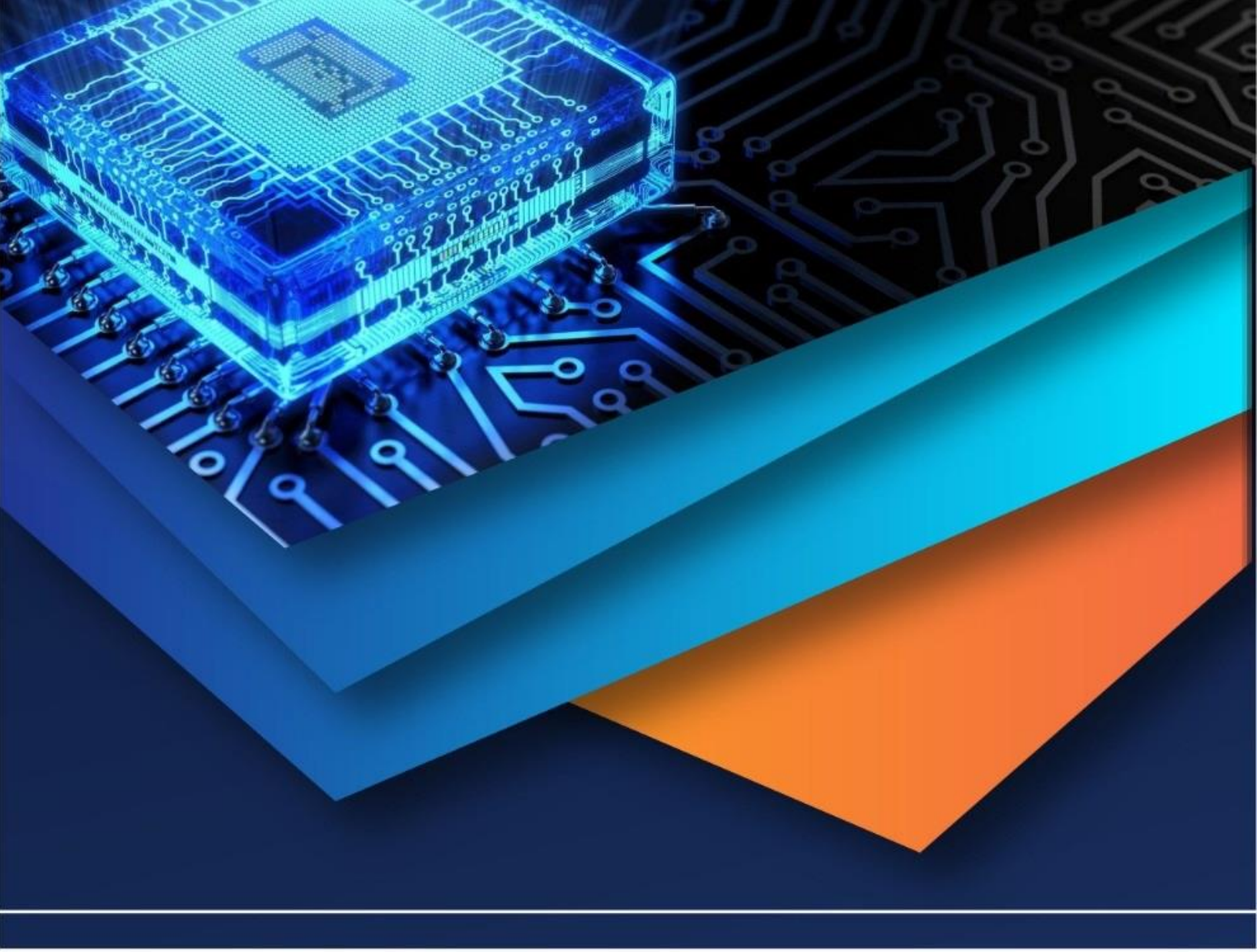
VI. CONCLUSION

We presented the details of two different concepts by uniting of cyber security and machine leaning: "machine learning in cyber security" and "cyber security in machine learning". We then discussed the advantages, issues and challenges of uniting of cyber security and ML. Further, we highlighted the different attacks and also provided a comparative study of various techniques in two different considered categories. Finally, some future research directions are provided. machine learning techniques are becoming quite useful in the cybersecurity industry. Traditional detection techniques have shown to be insufficient in addressing the developing nature of cybercrimes, given the rapid increase of cyber threats and attacks. By creating automated and intelligent systems that can analyse massive amounts of data, spot patterns, and spot potential security breaches in real-time, machine learning provides a solution. This article has covered a number of machine learning applications in cybersecurity, such as spam classification, malware detection, intrusion detection, and more. These software programmes make use of machine learning methods to improve threat detection and reaction times. Machine learning algorithms can learn to distinguish between legitimate and harmful activity by being trained on labelled datasets, making it possible to identify cyber threats and attacks. Yet, there are difficulties in applying machine learning to cybersecurity.

REFERENCES

[1] Butun, P. Osterberg, H. Song, Security of the internet of things: Vulnerabilities, attacks, and countermeasures, IEEE Commun. Surv. Tutor. 22 (1) (2020) 616–644, http://dx.doi.org/10.1109/COMST.2019. 2953364.
[2] Z. Lv, L. Qiao, J. Li, H. Song, Deep-learning-enabled security issues in the internet of things, IEEE Internet Things J. 8 (12) (2021) 9531–9538.
[3] Y. Wang, J. Yu, B. Yan, G. Wang, Z. Shan, BSV-PAGS: Blockchainbased special vehicles priority access guarantee scheme, Comput. Commun. 161 (2020) 28–40.

[4] N. Magaia, R. Fonseca, K. Muhammad, A.H.F.N. Segundo, A.V. Lira Neto, V.H.C. de Albuquerque, Industrial internet-of-things security enhanced with deep learning approaches for smart cities, IEEE Internet Things J. 8 (8) (2021) 6393–6405.
[5] S.A. Parah, J.A. Kaw, P. Bellavista, N.A. Loan, G.M. Bhat, K. Muhammad, V.H.C. de Albuquerque, Efficient security and authentication for edge-based internet of medical things, IEEE Internet Things J. 8 (21) (2021) 15652–15662.
[6] Y. Sun, A.K. Bashir, U. Tariq, F. Xiao, Effective malware detection scheme based on classified behaviour graph in IIoT, Ad Hoc Netw. 120 (2021) 102558.
[7] J. Yang, Z. Bian, J. Liu, B. Jiang, W. Lu, X. Gao, H. Song, No reference quality assessment for screen content images using visual edge model and Ada Boosting neural network, IEEE Trans. Image Process. 30 (2021) 6801–6814.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY