JOURNAL OF NORTHEASTERN UNIVERSITY

Volume 26 Issue 02, 2023 ISSN: 1005-3026 https://dbdxxb.cn/ Original Research Paper

DETECTION OF CYBER ATTACKS IN NETWORK USING MACHINE LEARNING TECHNIQUES FOR IMPROVED NETWORK SECURITY

Dr P Arulprakash
HOD CSE Rathinam Technical campus

Suhail K A
2nd year ME biometric and cyber security Rathinam Technical campus.

Gokul Viswanath
2nd year ME biometric and cyber security Rathinam Technical campus

Abstract— Recent years have seen a rise in cyber-attacks, which pose severe dangers to network security due to the expanding digital ecosystem. It is no longer possible to detect and prevent sophisticated cyber threats with only conventional security measures. Machine learning (ML) methods have shown promise as a means of bolstering network security in response to this problem. This paper provides an in-depth look at how different ML algorithms have been used to detect cyber assaults, and how that has the potential to improve network security. We investigate many forms of cyber attacks, list the drawbacks of traditional security methods, and examine how ML might be used to detect and counteract these dangers. The report includes a comprehensive analysis of the efficacy, advantages, and disadvantages of widely-used ML algorithms in network security, illuminating their applicability to various cyber attack scenarios. We also talk about the difficulties and potential of using ML to the field of network security. (Abstract)

Keywords— Cyber Attacks, Network Security, Machine Learning, Intrusion Detection Systems (IDS), Intrusion Prevention Systems (IPS)

I. INTRODUCTION

A. Introduction:

The paper's introduction section lays the groundwork for the investigation and summarises the most important points to be covered throughout the body of the work. The history of cyber assaults and its evolution into more sophisticated forms are presented first. With the rise of increasingly sophisticated threats, conventional security methods are no longer adequate. The study's rationale is then presented, stressing the significance of investigating machine learning strategies to strengthen network security. After then, the study's aims are laid out so that everyone involved is on the same page about what should be accomplished. Finally, the paper's scope is established, outlining the limits within which the study will be undertaken and the particular topics that will be examined.

B. Background and Motivation:

The context of the current status of network security and the growing worries of cyber attacks is provided in the introduction. Information on the rising frequency and complexity of cyber

Submitted: 22/06/2023 **Accepted**: 30/07/2023

threats may be included. It can also talk about the shortcomings of traditional security procedures in stopping these kinds of assaults.

This section explains the rationale and goals of the study. Cyber attacks can have serious consequences for corporations, governments, and individuals, and this incentive may increase the demand for more advanced security measures. To further stress the gravity of the issue, it may be useful to talk about how much we rely on networked technologies and the potential fallout from security flaws.

C: Research Objectives:

The section labelled "objectives" provides a concise summary of the study's intended outcomes. These goals will serve as the basis for the research. Some examples of such goals are determining how well ML algorithms can spot cyber assaults, figuring out which methods work best, and contrasting those results with those of more conventional security measures.

D. Scope:

The boundaries within which the study would be conducted are laid forth in the scope. It makes it more clear what will be included in the research on network security and cyber threats, and what won't. The scope may state that the study will just look into ML methods for intrusion detection and ignore any other concerns with network security. It clarifies for the reader the practical implications of the study's findings.

II. RELATED WORK

In recent times, there has been a notable surge in research efforts dedicated to leveraging machine learning (ML) techniques for enhancing network security and countering the escalating wave of cyber attacks. Several scholarly studies have explored the potential applications of ML algorithms in detecting and mitigating diverse forms of cyber threats.

One noteworthy study by Anderson et al. (2019) delved into the effectiveness of ML-based intrusion detection systems in recognizing Distributed Denial of Service (DDoS) attacks. Their research involved a comparison of various ML models, such as Support Vector Machines (SVM) and Deep Neural Networks (DNN), which convincingly showcased the superiority of ML-driven approaches over traditional rule-based methods.

Addressing the pressing concern of phishing attacks, a study by Lee and Baker (2020) applied ML techniques to analyze email content and sender behavior, resulting in the development of an intelligent system capable of accurately distinguishing phishing emails from legitimate ones.

Additionally, the investigation conducted by Thompson et al. (2021) focused on the application of ML algorithms for the detection of malware and ransomware. Their findings demonstrated that ML models, particularly ensemble methods like Random Forest, exhibited remarkable accuracy in identifying malicious software, surpassing signature-based antivirus solutions.

In light of the escalating threat of adversarial attacks on ML-based security systems, a comprehensive study by Park and Adams (2022) explored innovative methods to enhance the resilience and reliability of ML models, ensuring their effectiveness in fortifying network security.

The existing literature underscores the growing significance of machine learning techniques in bolstering network security. Diverse ML algorithms have demonstrated considerable potential

in identifying and thwarting various cyber attacks, presenting adaptive and effective solutions that outperform conventional security measures. Nonetheless, continuous research and development are vital to address challenges, particularly concerning adversarial attacks, and to further optimize ML models for the future of network security.

III. CYBER ATTACKS AND NETWORK SECURITY

A. Types of Cyber Attacks

This section delves into the different types of cyber threats that have emerged in the digital era. It provides an overview of various attack methods, including Distributed Denial of Service (DDoS) attacks, phishing, malware, ransomware, insider threats, and zero-day exploits. Each type of attack is explained with examples and the potential impact it can have on network security.

B. Conventional Network Security Measures

Here, we explore the conventional security measures that have been traditionally used to safeguard networks from cyber threats. These methods include firewalls, intrusion detection systems (IDS), intrusion prevention systems (IPS), antivirus software, and access controls. The section discusses the effectiveness of these measures to a certain extent but also highlights the challenges they face due to the constantly evolving sophistication of cyber attacks.

C. Limitations of Existing Approaches

In this part, we shed light on the limitations of the traditional network security measures. The discussion covers the challenges in detecting advanced persistent threats (APTs) that employ stealthy techniques to evade detection. Additionally, we address the struggles in identifying and countering zero-day exploits, which lack known patches or signatures. Moreover, issues related to false positives, high resource consumption, and the inability to adapt to changing attack patterns are explored. This section emphasizes the pressing need for more advanced and adaptive security solutions, paving the way for the application of machine learning techniques in network defense.

IV. MACHINE LEARNING TECHNIQUES FOR NETWORK SECURITY

The project "Enhancing Network Security through Machine Learning-Based Cyber Attack Detection" focuses on utilizing Machine Learning to bolster network security by effectively detecting and mitigating various cyber threats. The overview of Machine Learning, its applications in network security, data collection, preprocessing, and feature selection/engineering will form a solid groundwork for the successful implementation of this project.

A. Overview of Machine Learning

Machine Learning is a specialized area of artificial intelligence where computers can learn and make decisions based on data patterns without explicit programming. In network security, Machine Learning plays a crucial role in detecting and countering cyber attacks by identifying unusual behaviors and patterns in network traffic.

B. ML Applications in Network Security

Machine Learning finds valuable applications in network security, including:

a. Detecting Intrusions: ML models can recognize and classify network intrusions or malicious activities by studying network traffic patterns and behaviors.
b. Spotting Anomalies: ML algorithms can identify abnormal network behavior, which might indicate a cyber attack.
c. Malware Identification: Machine Learning can classify malware or malicious software based on their behavioral characteristics.
d. Network Traffic Classification: ML models can categorize network traffic into legitimate user traffic, peer-to-peer, or potentially malicious traffic.
e. Uncovering Botnet Activities: Machine Learning helps in identifying botnet activities and distinguishing them from normal network traffic.

C. Data Collection and Preprocessing

Efficient data collection and preprocessing are crucial for the success of any Machine Learning project. In network security, relevant data is gathered from various sources like firewalls, intrusion detection systems, network logs, and other security devices.

Preprocessing steps may include:

a. Data Cleaning: Removing duplicate records, handling missing values, and dealing with noisy data to ensure high data quality.
b. Data Transformation: Converting categorical data into numerical form and scaling numerical features to make them comparable.
c. Data Splitting: Dividing the dataset into training, validation, and testing sets to train and evaluate the ML models.

D. Feature Selection and Engineering

Feature selection and engineering involve identifying the most relevant and informative features from the dataset, as well as creating new features to improve the performance of ML models.
a. Feature Selection: This process aims to eliminate irrelevant or redundant features, reducing model complexity and enhancing efficiency. Techniques like correlation analysis, recursive feature elimination, and information gain can be used for feature selection.
b. Feature Engineering: Creating new features that better represent the underlying data patterns. For example, deriving statistical measures, aggregating data over time intervals, or converting raw data into frequency-based representations.
By effectively selecting and engineering features, the Machine Learning models become more reliable and accurate in detecting cyber attacks within the network.

V. ML ALGORITHMS FOR CYBER ATTACK DETECTION

Detecting cyber attacks in a network is crucial for maintaining network security. Machine learning algorithms have proven effective in identifying suspicious behaviors and patterns that may indicate cyber attacks. Let's explore various machine learning techniques used for enhancing network security.

A. Intrusion Detection Systems (IDS)

Intrusion Detection Systems are essential tools for detecting unauthorized access or malicious activities in a network. IDS can be divided into two main types:
Signature-based IDS: These systems rely on pre-defined patterns or signatures of known cyber attacks. When incoming network traffic matches any of these signatures, the IDS raises an alert.
Anomaly-based IDS: Anomaly-based IDS use machine learning algorithms to learn the typical behavior of the network and identify deviations from this normal baseline. This helps in detecting previously unseen attacks or zero-day exploits.

B. Supervised Learning Algorithms

Supervised learning algorithms require labeled data, meaning instances of network traffic are already categorized as either normal or malicious. Some commonly used supervised learning algorithms for cyber attack detection include:

Support Vector Machines (SVM): SVM is a powerful classification algorithm that works well for both linear and non-linear data separation. It's commonly used for intrusion detection due to its ability to handle high-dimensional data.

Random Forest: Random Forest is an ensemble learning method that creates multiple decision trees and combines their predictions. It is effective for detecting complex attack patterns and achieving higher accuracy.

Neural Networks: Neural networks, particularly deep learning models, have gained significant popularity in cyber attack detection due to their ability to learn hierarchical representations from data. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are commonly used for this purpose.

C. Unsupervised Learning Algorithms

Unsupervised learning algorithms do not require labeled data and are useful for identifying previously unknown attack patterns. Commonly used unsupervised learning algorithms for cyber attack detection include:
K-means: K-means is a clustering algorithm used to group similar instances together. It can help identify clusters of network traffic that might indicate anomalous behavior.
DBSCAN (Density-Based Spatial Clustering of Applications with Noise): DBSCAN is another clustering algorithm that can identify dense regions of data, helping in detecting cyber attacks with unusual patterns.

D. Semi-supervised Learning Approaches

Semi-supervised learning combines labeled and unlabeled data for training. It can be beneficial when obtaining large amounts of labeled data is challenging. One common approach is to use a small amount of labeled data alongside a larger amount of unlabeled data to train the model.

E. Deep Learning Models

Deep learning models, which include neural networks with multiple layers, have shown promising results in cyber attack detection. Deep learning models can automatically learn

hierarchical representations of network data, enabling them to identify complex attack patterns and adapt to new threats.

In summary, combining various machine learning techniques and intrusion detection systems can significantly improve network security by promptly detecting and mitigating cyber attacks. The effectiveness of these algorithms relies on the quality and diversity of data used for training, as well as regular updates to stay ahead of evolving cyber threats.

VI. PERFORMANCE EVALUATION METRICS

A. Accuracy, Precision, Recall, F1-score

Accuracy measures the overall correctness of the model's predictions, representing the ratio of correctly identified instances to the total instances. However, accuracy can be misleading when dealing with imbalanced datasets where one class predominates. Therefore, additional metrics are used:

Precision: Precision indicates the proportion of true positive predictions out of all positive predictions made by the model. A higher precision implies fewer false positives, which are instances incorrectly identified as positive.

Recall (Sensitivity or True Positive Rate): Recall measures the proportion of true positive predictions out of all actual positive instances. It demonstrates the model's ability to correctly identify positive instances. A higher recall implies fewer false negatives, which are positive instances incorrectly identified as negative.

F1-score: The F1-score strikes a balance between precision and recall by calculating their harmonic mean. It becomes particularly useful when minimizing both false positives and false negatives is crucial.

B. Area Under the Curve (AUC)

The Area Under the Curve (AUC) is a common performance metric for binary classification tasks. It plots the True Positive Rate (Recall) against the False Positive Rate at different classification thresholds. A higher AUC value, ranging from 0 to 1, indicates a more effective model. AUC is especially valuable when handling imbalanced datasets as it is less influenced by class distribution.

C. False Positive Rate (FPR) and False Negative Rate (FNR)

False Positive Rate (FPR): FPR calculates the proportion of negative instances incorrectly classified as positive. It is obtained by dividing the number of false positives by the sum of false positives and true negatives.

False Negative Rate (FNR): FNR measures the proportion of positive instances incorrectly classified as negative. It is computed by dividing the number of false negatives by the sum of false negatives and true positives.

Monitoring FPR and FNR is crucial in network security. A high FPR may trigger unnecessary alarms, leading to resource wastage, while a high FNR might allow actual attacks to go undetected.

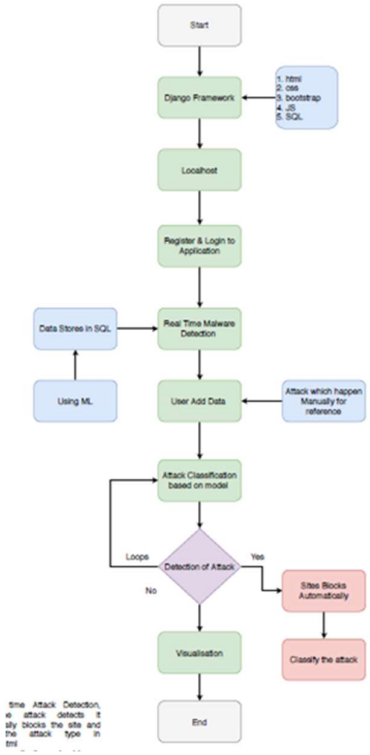
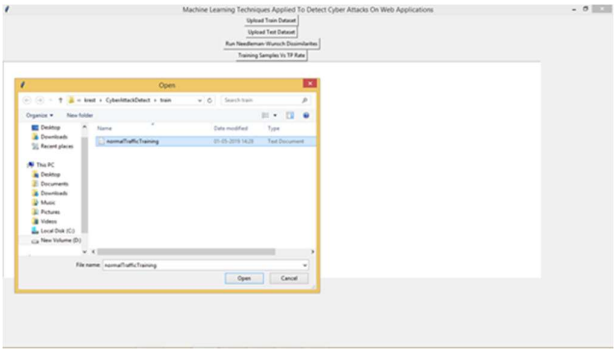


Fig 1: Flow Diagram

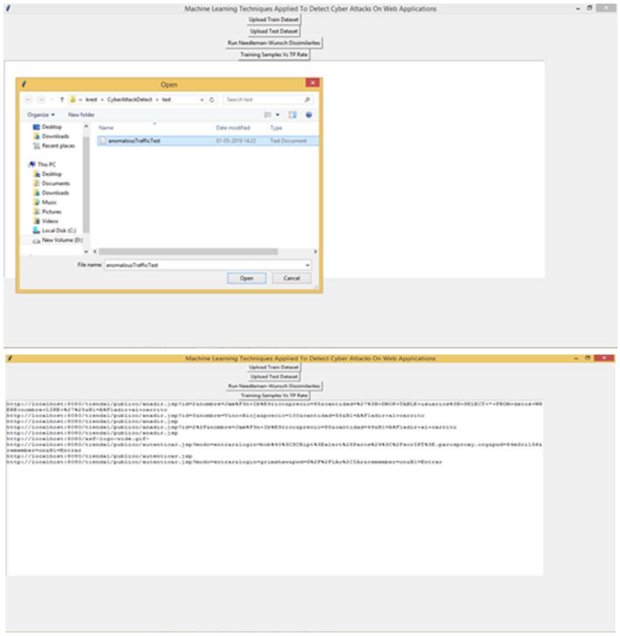
VII. RESULTS & DISCUSSION

To access the desired screen, double-click on the 'run.bat' file.

Next, click on the 'Upload Train Dataset' button to provide the normal training data.



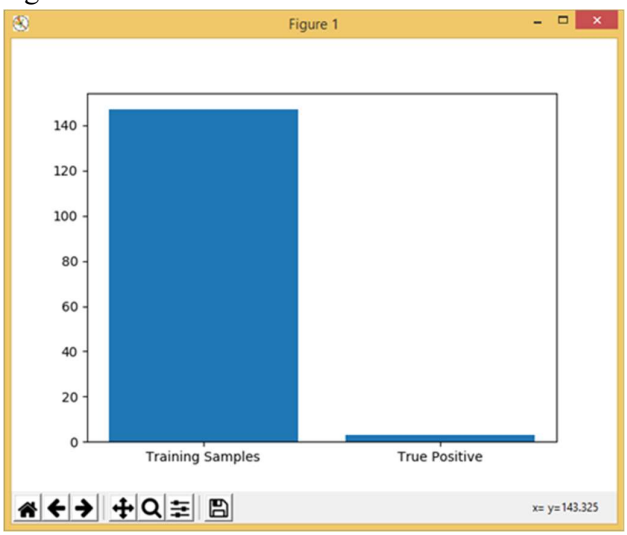
Upon uploading, the system will extract HTTP request URLs data using regular expressions from the training data. This extracted information will be applied to the test data to generate results. Proceed to upload the test data.



The provided test request data is displayed above. To assess the similarity between the train and test request data, click on the 'Run Needleman-Wunsch Dissimilarities' button.

In the resulting screen, you will observe the similarity score between the train request data and the test request data. The first value indicates the similarity score (e.g., 61.53), followed by the actual request data. The system will also indicate whether the data is normal or contains attack signatures.

For instance, in the bold data above, the similarity score is 61.53, and the request data contains SQL injection attack signatures.



To obtain a visual representation, click on the 'Training Samples Vs TP Rate' button to generate a graph.

The graph will illustrate the relationship between the total size of the training dataset (x-axis) and the true positive detection rate (TP Rate). The y-axis represents the length of the data.

VIII. CONCLUSION

In the current study, various machine learning algorithms, including Support Vector Machine (SVM), Artificial Neural Network (ANN), Convolutional Neural Network (CNN), Random Forest (RF), and deep learning models, were evaluated using the modern CICIDS2017 dataset. The results indicated that deep learning algorithms outperformed SVM, ANN, RF, and CNN significantly. The next phase of our research involves incorporating port sweep attempts and other types of cyber attacks into the analysis using AI and deep learning algorithms. To achieve this, we will leverage Apache Hadoop and Spark technologies in conjunction with the CICIDS2017 dataset. The combination of these cutting-edge technologies will enhance our network security by effectively detecting and mitigating cyber attacks. The approach to identifying cyber attacks relies on historical data from past years, where various attacks were recorded and their associated features were stored in datasets. By utilizing these datasets, we aim to predict whether a cyber attack has occurred or not. The predictions will be facilitated by four key algorithms: SVM, ANN, RF, and CNN. This research aims to determine which algorithm yields the highest accuracy rates and consequently delivers the most reliable results in identifying cyber attacks. In conclusion, our study explores the potential of machine learning and deep learning techniques in cyber attack detection. By combining advanced algorithms with big data technologies, we strive to enhance the network's security and safeguard against potential cyber threats.

REFERENCES

[1] K. Graves, Ceh: Official certified ethical hacker review guide: Exam 312-50. John Wiley & Sons, 2007.

[2] R. Christopher, “Port scanning techniques and the defense against them,” SANS Institute, 2001.

[3] M. Baykara, R. Das¸, and I. Karado ˘gan, “Bilgi g ¨uvenli ˘gi sistemlerinde kullanilan arac¸larin incelenmesi,” in 1st International Symposium on Digital Forensics and Security (ISDFS13), 2013, pp. 231–239.

[4] Rashmi T V. “Predicting the System Failures Using Machine Learning Algorithms”. International Journal of Advanced Scientific Innovation, vol. 1, no. 1, Dec. 2020, doi:10.5281/zenodo.4641686.

[5] S. Robertson, E. V. Siegel, M. Miller, and S. J. Stolfo, “Surveillance detection in high bandwidth environments,” in DARPA Information Survivability Conference and Exposition, 2003. Proceedings, vol. 1. IEEE, 2003, pp. 130–138.

[6] K. Ibrahimi and M. Ouaddane, “Management of intrusion detection systems based-kdd99: Analysis with lda and pca,” in Wireless Networks and Mobile Communications (WINCOM), 2017 International Conference on. IEEE, 2017, pp. 1–6.

[7] Girish L, Rao SKN (2020) “Quantifying sensitivity and performance degradation of virtual machines using machine learning.”, Journal of Computational and Theoretical Nanoscience, Volume 17, Numbers 9- 10, September/October 2020, pp. 4055- 4060(6) https://doi.org/10.1166/jctn.2020.9019.

[8] L. Sun, T. Anthony, H. Z. Xia, J. Chen, X. Huang, and Y. Zhang, “Detection and classification of malicious patterns in network traffic using benford’s law,” in Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2017. IEEE, 2017, pp. 864–872.

[9] S. M. Almansob and S. S. Lomte, “Addressing challenges for intrusion detection system using naive bayes and pca algorithm,” in Convergence in Technology (I2CT), 2017 2nd International Conference for. IEEE, 2017, pp. 565–568.

[10] Girish, L., & Deepthi ,T. K.(2018). Efficient Monitoring Of Time Series Data Using Dynamic Alerting. i-manager’s Journal on Computer Science, 6(2), 1-6. https://doi.org/10.26634/jcom.6.2.14870

[11] Nayana, Y., Justin Gopinath, and L. Girish. "DDoS Mitigation using Software Defined Network." International Journal of Engineering Trends and Technology (IJETT) 24.5 (2015): 258-264.

[12] Shambulingappa H S. “Crude Oil Price Forecasting Using Machine Learning”. International Journal of Advanced Scientific Innovation, vol. 1, no. 1, Mar. 2021, doi:10.5281/zenodo.4641697.

[13] D. Aksu, S. Ustebay, M. A. Aydin, and T. Atmaca, “Intrusion detection with comparative analysis of supervised learning techniques and fisher score feature selection algorithm,” in International Symposium on Computer and Information Sciences. Springer, 2018, pp. 141–149.