International Journal On Engineering Technology and Sciences – IJETS™
ISSN (P):2349-3968, ISSN (O): 2349-3976
Volume X, Issue III, March - 2023

MACHINE LEARNING FOR ENHANCED CYBER SECURITY

Lakshmikanth S
Computer Science and Engineering
Francis Xavier Engineering College,
Tirunelveli – Tamil Nadu - India
lakshmikanths.ug19.cs@francisxavier.ac.in

Maria Jabestine M
Computer Science and Engineering
Francis Xavier Engineering College,
Tirunelveli – Tamil Nadu - India
mariajabestinem.ug19.cs@francisxavier.ac.in

Kishorekumar S
Computer Science and Engineering Francis Xavier Engineering College
Tirunelveli– Tamil Nadu - India
kishorekumars.ug19.cs@francisxavier.ac.in

Dr. R. Ravi
Professor / Dept. of Computer Science and Engineering,
Francis Xavier Engineering College
Tirunelveli – Tamil Nadu – India
dr.r.ravi@francisxavier.ac.in

Abstract:

Cyber security is a big issue in current society since exploiting computer network vulnerabilities has become simple thanks to technological advances and human talents. Currently, several types of assaults are occurring, such as DOS attacks, probing, R2U, R2L viruses, port scanning, buffer overflow, CGI attacks, and floods, among others. We require a foundation on which to build a system for detecting and preventing these threats. The majority of the most recent ways for implementing IDS for computer security are covered in this article. Intrusion Detection Systems are the best answer for cyber-attacks. In a continuously changing environment, machine learning-based intrusion detection systems exhibit excellent accuracy. This study also addresses the least accurate ML approach and investigates potential research areas for researchers.

Keywords: Cyber security, DOS attacks, probing, R2U, R2L viruses

Introduction:

Malicious hackers have a significant edge in the cyberwarfare since, out of numerous attempts, the attacker only requires one successful effort, while security personnel need a success rate of 100%. Edwin Raja S and Ravi R (2020) proposed to use the DMLCA approach to increase the detection accuracy utilizing a variety of factors, including detection accuracy based on true positive ratio, precision, and recall [1].and it can identify many types of attacks. However, the Internet environment is increasing network complexity, structure, and diversity, while assailants are also updating attack technology, making traditional IDS challenging to meet security needs. We need an advanced IDE for network attack detection and prevention. Machine Learning methods are one of the notable approaches used for identifying network assaults. There are several branches of artificial intelligence. Machine learning is an example of them; it may self-learn based on prior data and enhance systems autonomously without even being computer vision ML approaches rely on mathematical models to make decisions after analyzing patterns in datasets, and IDS then predicts the outcome for fresh inputted data. Machine Learning has various application and spread over a large domain. Among them are ecommerce, where ML is used to propose products to customers based on their behavior, and health care, where ML applications are used to recommend products to

patients based on their symptoms. There are three kinds of machine learning algorithms.

Supervised Learning:

The primary goal of supervised learning is to learn from pairs of inputs and outputs, a function that maps inputs to outputs. It takes labeled data and outputs a prediction about a function. Artificial neural networks, regression, Bayesian statistics, and the Gaussian distribution are all examples of supervised learning methods. Bayesian statistics, Practitioner, Distribution function, Decision Tree, and Support Vector Machine Naive Bayes, K-nearest neighbor, and random forest models

Unsupervised Learning:

Learn without being watched for their methods, scientists often make use of unlabeled data samples. This method employs clustering. Common unsupervised learning approaches include cluster analysis, apriori algorithms, the eclat algorithm, and outlier identification

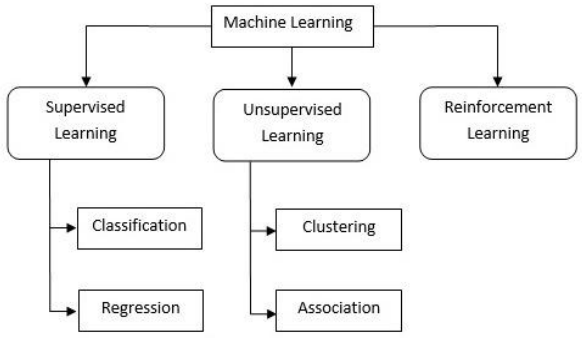
Reinforcement Learning:

In this method, the system is programmed to carry out actions in response to environmental cues. In this method, the label for an unlabeled instance is set by the user.

Figure 1: Types of Machine Learning

Cyber Assaults:

Machine Learning



ransomware, spyware, malware installations, phishing, SQL injection, and denial of service attacks. In the previous several decades, hackers have targeted a wide variety of businesses, exposing the personal information of hundreds of thousands of customers. The vulnerable companies' markets eventually collapsed. Here are some common names for forms of cybercrime.

1. Probing Attack: Probing is an attack method used to gather information or identify potential weak spots in computer networks. There are various services and computer paths open to assaults in the network. Some people utilize social engineering in probing tactics; this is a strategy that has been widely discussed and can be applied with just a little amount of training.

2. Denial of Service: Buffer overflow assaults, ICMP floods, and SYN floods are all examples of DOS attacks, which are designed to overload the resources of the system being attacked. This allows the authorized user access to the system. Bugs in the implementation or in the system's miss-configuration are the primary targets of attacks.

3. User to root: Third-party attackers may use user-to-root assaults, in which they aim to gain full control of a local system. And use that unlawful access to their advantage.

Remote to User: The fourth kind of assault is known as a "remote to user attack," and it involves an attacker sending packets to a distant system via a network in an effort to obtain access to that machine. Although R2L is employed in many contexts, it is most often seen in social engineering. The challenge now is figuring out how to prevent hackers from gaining access to sensitive user information while also keeping them out of the network.

Related Works:

According to M. Chandru, S. Kasi Rajesh, S. Eeben, A. Mano Pandiyan, and R. Ravi (2021) utilize the basic troubleshooting commands to determine the state of the router. Configure the communication server. Use the Cisco Discovery Protocol (formerly known as CDP) to learn the fundamentals of a network's topology [3].
In the researchers conducted a thorough analysis of previous research as well as their own tests on real-world, high-traffic networks and major companies. The authors divide Machine Learning into two groups, Shallow

Learning and Deep Learning, which in turn are separated into supervised and unsupervised algorithmic approaches.

In [9], anti-phishing technique "Scamming Detection by Leaning on Characteristics of Email Received" (PILFER) was created and tested on a dataset including 860 spam emails and 695 ham instances, with C4.5 and decision trees also being used. Several elements were identified as potential indicators of whether an email was a phishing attempt or not. These included Internet Protocol addresses, message headers, HTML, the amount of internal links, and even Java script. Therefore, the authors indicated that by combining all 10 characteristics from the classifier besides " Spam filter output " [6], PILFER might enhance message clustering. Nonetheless, Logistic Regression was investigated for this study.

R. Mallika@pandeeswari, G. Rajakumar, and R. Ravi (2020) discussed the learning of functional representations and the development of deep metric awareness of new loss functions and provide in-depth data analysis, produce analysis on current datasets [2].
In the authors contrasted several phishing detection models in terms of their content and characteristics; the experimental part shows that the knowledge-based strategy given by Ridor and eDRI methods seems to be suitable to fight phishing for two reasons.

For one, the resulting models have very competitive classification accuracy.

So that even inexperienced users can make informed judgments, the models provide information that is both accessible and simple to grasp.

The detection rates of the Bayesian Net and SVM decision trees were high.

Using the CTU-13 dataset, the authors of [8] decided to apply four distinct deep learning models, analyses, including a Convolutional Neural Network (CNN), a Long Short-Term Memory (LSTM), a hybrid CNN-LSTM, and a Multi-layer Perspective (MLP) for bot identification and simulation. The research shows 100% precision via using the CNN model for Scenario 3, which contains only 24 botnet flows. All measures of accuracy, sensitivity, specificity, precision, and F1 Score were shown to be perfect when the same authors used CNN-LSTM and MLP. Logistic Regression, a method of supervised machine learning, will be applied to the same dataset.

Misclassification's hidden costs are a major issue in the cybersecurity industry. Security operators are frustrated by false positives in malware categorization and intrusion detection, and they are hampered in their efforts to fix true infections [10]. The method described in [9] is one we examine while analyzing malware. In this work, we take the "Logistic Regression" supervised machine learning method and do a thorough study and comparison of its many parameters which have been training and testing on the same dataset.

Methodology:

The literature review has led us to start looking for studies that go further into the topic of Bot activity and Malicious traffic classification using Machine Learning. We wanted models that can be used in a variety of security scenarios depending on the requirements of the business at hand, such as a model with extreme accuracy, between many F1-score and Recall values, that can be implemented in highly sensitive solutions, or a prototype with delivering The highest quality to capture as many events as reasonably possible of the model's precision, or a model that combines appropriate accuracy and recall detection. Authors in [9] have conducted experiments with several Machine Learning (ML) techniques. But we could expand the study and go further into evaluating many more factors. We split our tests between a robust physical server and Amazon Web Service (AWS) instances in the cloud. Accuracy, Recall, and F1-score values were gathered during testing and are shown in Figure (1).

Dataset:

IDS developers need a dataset, which is a compiled set of assaults, to evaluate the effectiveness of their product. Lots of datasets are now readily accessible. Just to provide one illustration :NSL-KDD

KDD cup99

CIC-IDS 17

Selection of features

This is a necessary process for picking characteristics from of the retrieved ones. The dimensionality of the input pattern matrix is decreased by using feature selection methods. Different methods are used for this purpose, including filtering features with Pearson Correlation, wrapper methods with Backward Process of separation, embedded methods inside the Random Forest Classifier, the Principal Component Analysis (PCA), and the t-distributed Stochastic Neighbour Embedding (t-SNE).

Locating Botnets

We put our model through its paces on the CTU-13 database [16] to determine whether it can accurately identify botnet activity. Everywhere you look, there are signs of bot traffic, or software-generated traffic, being used for a particular reason. It is critical to sort out the good bot traffic from the bad. The F1 score is used to evaluate the effectiveness of botnet detection systems as a whole.

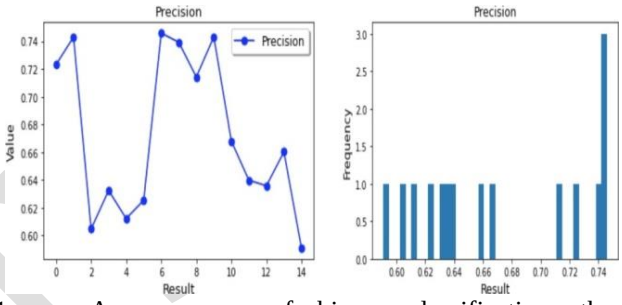
Classification:

1. An Intrusion Detection System (IDS) is a proactive piece of equipment that monitors and analyzes network traffic in order to identify malicious threats like hackers, spammers, and viruses. Machine learning methods and intrusion detection systems are discussed in A. Shakeela Joy and R. Ravi (2017) introduced the IRE Scheme, which uses ECC for encryption to recognize iris patterns [9].
2. To the extent that there are any, trigger the alarm. The components of IDS might be either software, hardware, or a hybrid of the two. The only purpose of IDS is to stop an attacker in the act, preventing any permanent harm to data or information. Threats to a network are thwarted by IDS. An IDS provides three crucial functions: monitoring, detection, and the generation of a signal, making it an essential weapon in the arsenal against network security threats. Deep Learning, Evolutionary Algorithms, Support Vector Machines Rajastephi, S., and Priskilla, J. Angel Rani and Ravi (2014) suggested that nodes may work together. By collaboratively sending each other's data, nodes become partners. It stops heterogeneous networks from interfering with one another [6].

, Logistic Regression, and Artificial Neural Networks are all examples of intelligent IDS methodologies. The algorithms can't be put through their paces without a sample data and intrusion detection systems to run them on. Due to the lengthy nature of data collection from computer networks, developers often test their IDS using preexisting datasets. All conceivable types of tests are included in these datasets. Eighty percent of the information we use to train is connected to attackers.

Result and Discussion:

Logistic Regression:



1. As a means of binary classification, the "Logistic regression" supervised learning model is used. The origin of the term may be traced back to the study of statistics. Logistic functions, a useful sigmoid curve with applications in many disciplines, including neural networks, form the basis of this approach. For issues of classification with two potential outcomes, logistic regression provides a statistical model. Malicious network traffic may be identified with the use of logistic regression. After being motivated by the work in According to S. Raja Ratna and R. Ravi (2015) the suggested method considerably identifies suspect routes with a higher detection rate and a reduced false positive probability; it also achieves higher throughput and less delay [4].

we trained and extensively tested a logistic regression model.

Precision values were determined for 15 separate tests, each with a distinct Width Window and Stride, and are shown in Figure (2). Experiment 6 yielded the best result, a "0.745" when Width was set to 3 minutes and Stride to 30 seconds. Using the parameters "Width" and "Stride," the smallest value was "0.591," which was found in experiment number 14.

The classifier found a grand total of one True Positive and 14,951 True Negatives. There were 10 erroneous positive predictions made by the classifier, but none erroneous negative ones. In these simulations, the proportion of bot, malware, and human visitors is unpredictable, just as it would be in the real world. To View a Figure (3).

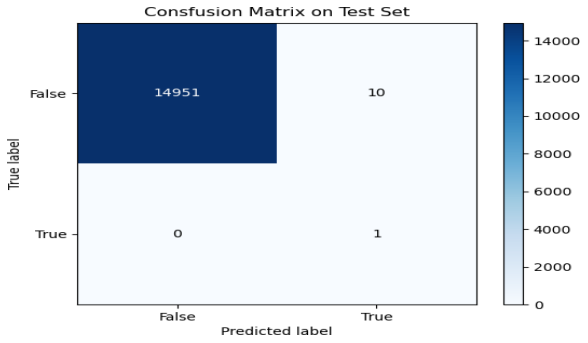


Figure 3 : Confusion Matrix

Conclusions:

1. Various forms of cyber-attacks and applications of machine learning are discussed throughout this article. An overview of the Machine Learning techniques currently in use for cyber-attack detection is also included. An explanation of what IDS is and how it operates is also provided. Finally, a table-formatted collection of machine learning algorithms with the highest accuracy against a subset of cyber-attacks is provided, as is a description of the most recent cyber-attacks and the means by which they might be avoided. There are benefits and drawbacks to each strategy. According to M. Chandru, S. Kasi Rajesh, S. Eeben, A. Mano Pandiyan, and R. Ravi (2021) utilize the basic troubleshooting commands to determine the state of the router. Configure the communication server. Use the Cisco Discovery Protocol (formerly known as CDP) to learn the fundamentals of a network's topology [5].

Finding new cyberattacks and improving our defense strategies is an active field of study. To ensure that our classifier is the most effective for detecting Bot traffic, we have tested it with the same 189 second Width and 129 second Stride values in different situations. 55.3 percent (695 flows) are Botnet flows, 3.6 percent (4,679 flows) are Normal flows, 11.5 percent (206 flows) are Command & Control flows, and 94.7 percent (124,252 flows) are Background flows in Scenario #5. The classifier found a grand total of one True Positive and 14,951 True Negatives. There were 10 erroneous positive predictions made by the classifier, but none erroneous negative ones. In these simulations, the proportion of bot, malware, and human visitors is unpredictable, just as it would be in the real world. To View a Figure (3).

References: Edwin Raja S and Ravi R, "A performance analysis of Software Defined Network based prevention on phishing attack in cyberspace using a deep machine learning ith CANTINA approach(DMLCA)", Computer Communications, vol. 152, pp.0-6, 2020.

[1] D.Priyadharshini,R. malliga@pandeeswari, S. shargunam, and R. Ravi, "Cyber security: a comprehensive survey and perspective on recent works", Francis Xavier Journal of Science Engineering and Management, vol. 1, no.1, pp.1-3, 2020.

[2] M.Chandru, S.Kasi Rajesh , S. Eeben, A. Mano Pandiyan, and R. Ravi, "Carrier Ethernet And Edge Networking", International Journal of Advanced Research in Management, Architecture, Technology and Engineering, vol. 7, no. 4, pp.108-115, 2021.

[3] S. Raja Ratna et al. (2015) suggested identifying bad nodes and removing them from the network. Through simulation tests, it has been found that the suggested approach increases throughput and packet delivery ratio while reducing delay [17].

[4] According to M. Chandru, S. Kasi Rajesh, S. Eeben, A. Mano Pandiyan, and R. Ravi (2021) utilize the basic troubleshooting commands to determine the state of the router. Configure the communication server. Use the Cisco Discovery Protocol (formerly known as CDP) to learn the fundamentals of a network's topology

[5] Rajastephi, S., and Priskilla, J. Angel Rani and Ravi (2014) suggested that nodes may work together. By collaboratively sending each other's data, nodes become partners. It stops heterogeneous networks from interfering with one another [96]doi:10.1109/DEST.2010.5610590.

[6] R. Ravi and S. Radhakrishnan, "Provisioning QoS in Virtual Private Network using Dynamic Scheduling", Journal of Computer Science, vol. 4, no. 1, pp. 1-5, 2008.

[7] R.Binisha, M.Anisha Vergin, and R.Ravi, "Controlling the occurance of mobbing for preventing the packet loss in data centric network", International Journal On Engineering Technology and Sciences, vol.8, no. 9, pp. 15-18, 2021.

[8] M. Masthan ,and R. Ravi, "Preventing Zero Day Malware Attack Outbreaks in a Network Using Cyber Resilience Recovery Model", International Journal on Recent Researches in Science, Engineering and Technology, vol.4, no 6, pp. 1-20, 2016.

[9] According to A. Shakeela Joy and R. Ravi (2017) an enhanced endorsement method using elliptic curve cryptography offers higher security, confidentiality, and privacy. The technique is vulnerable to offline password guessing attacks including spidering, stolen-verifier, and keystroke dynamics