Study of Intrusion Detection System using Machine Learning Approach

[1]Sachin Ahirwar, [2]Prof. Sarvesh Site

M. Tech. Scholar, Department of Computer Science and Engineering, All Saints' College of Technology, Bhopal[1]

Guide, Department of Computer Science and Engineering, All Saints' College of Technology, Bhopal[2]

Abstract: Due to the expansion and development of modern networks, the volume and destructiveness of cyber-attacks are continuously increasing. Intrusion Detection Systems (IDSs) are essential techniques for maintaining and enhancing network security. IDS-ML is an open-source code repository written in Python for developing IDSs from public network traffic datasets using traditional and advanced Machine Learning (ML) algorithms. The accuracy and timely detection should be ensured by Network Intrusion Detection System (NIDS). For intrusion detection in balance and imbalance network traffic, machine learning and deep learning methods can be used. In this paper a survey of different intrusion detection systems based on machine learning and deep learning methods is performed. The proposed system adds on ensemble learning approach to improve accuracy. A review on various intrusion detection system (IDS) using the techniques in machine learning is been put forwarded.

Keywords: - Intrusion Detection System, Machine Learning, Network Intrusion Detection System

I. INTRODUCTION

An intrusion detection system (IDS) is a network security tool that monitors network traffic and devices for known malicious activity, suspicious activity or security policy violations. An IDS can help accelerate and automate network threat detection by alerting security administrators to known or potential threats, or by sending alerts to a centralized security tool. A centralized security tool such as a security information and event management (SIEM) system can combine data from other sources to help security teams identify and respond to cyber threats that might slip by other security measures. IDSs can also support compliance efforts. Certain regulations, such as the Payment Card Industry Data Security Standard (PCI-DSS), require organizations to implement intrusion detection measures. An IDS cannot stop security threats on its own. Today IDS capabilities are typically integrated with—or incorporated into—intrusion prevention systems (IPSs), which can detect security threats and automatically act to prevent them.

One of the resources that is used the most is the Internet. People are carrying out digitized transactions as a result of the huge increase in internet usage. The expansion in number of web clients is 4% to 6% consistently. It has been observed that the growth is much higher in developing nations like India. A huge amount of data is being generated online as a result of high levels of computer and internet technology adoption [1, 2]. It is difficult to find a person who does not have an online presence because these technologies have also become essential components of human life. Expansion in on the web presence has likewise leaded to sharing or support of individual data on the web. Albeit this acquires enormous comfort terms of access, they are additionally helpless against assaults or interruptions. These intrusions may result in significant financial losses as well as the disclosure of private information to unintentional users [3].

II. LITERATURE REVIEW

Pierre-Francois Marteau, "Sequence Covering for Efficient Host-Based Intrusion Detection", IEEE Transactions On Information Forensics And Security, Vol. 14, No. 4, April 2019, pp 944-1006 [1]. Inthis paper, author introduces a new similarity measure, the covering similarity, which they formally define for evaluating the similarity between a symbolic sequence and a set of symbolic sequences. A pair wise similarity can also be directly derived from the covering similarity to compare two symbolic sequences. An efficient implementation to compute the covering similarity is proposed which uses a suffix-tree data structure, but other implementations, based on suffix array for instance, are possible and are possibly necessary for handling very large scale problems. They have used this similarity to isolate attack sequences from normal sequences in the scope of host- based intrusion detection. They have assessed the covering similarity on two well-known benchmarks in the field.

Rashidah Funke Olanrewaju, Burhan Ul Islam Khan, Athaur Rahman Najeeb, Ku Nor Afiza Ku Zahir and Sabahat Hussain, "Snort-Based Smart and Swift Intrusion Detection System", Indian Journal of Science and Technology, Vol 11(4), DOl: 10.17485/IJST/2018/v11i4/120917, January 2018[2]. In this paper, author introduces a smart Intrusion Detection System(IDS) has been proposed that detects network attacks in less time after monitoring incoming traffic thus maintaining better performance. **Methods/Statistical**

Analysis: The features are extracted using back-propogation algorithm. Then, only these relevant features are trained with the help of multi- layer perceptron supervised neural network.

Ashima Chawla, Brian Lee, Sheila Fallon, and Paul Jacob, "Host Based Intrusion Detection System with Combined CNN/RNN Model", Springer Nature Switzerland August 2019, pp 149-158 [3]. In this paper, author describes a computational efficient anomaly based intrusion detection system based on Recurrent Neural Networks. Using Gated Recurrent Units rather than the normal LSTM networks it is possible to obtain a set of comparable results with reduced training times. The incorporation of stacked CNNs with GRUs leads to improved anomaly IDS. Intrusion Detection is based on determining the probability of a particular call sequence occurring from a language model trained on normal call sequences from the ADFA Data set of system call traces. Sequences with a low probability of occurring are classified as an anomaly.

Hebatallah Mostafa Anwer, Mohamed, Farouk, Ayman Abdel-Hamid, "A Framework for Efficient Network Anomaly Intrusion Detection with Features Selection", IEEE 2018, pp 157- 162 [4]. In this paper, author presents a features selection framework for efficient network anomaly detection using different machine learning classifiers. The framework applies different strategies by using filter and wrapper features selection methodologies. The aim of this framework is to select the minimum number of features that achieve the highest accuracy. UNSW-NB15 dataset is used in the experimental results to evaluate the proposed framework. The results show that by using 18 features from one of the filter ranking methods and applying J48 as a classifier, an accuracy of 88% is achieved.

Rana Aamir Raza Ashfaq , Xi-Zhao Wang , Joshua Zhexue Huang , Haider Abbas , Yu-Lin He, "Fuzziness based semi-supervised learning approach for intrusion detection system", Elsevier 2016, pp 484-497 [6]. This paper proposes a novel fuzziness based semi-supervised learning approach by utilizing unlabeled samples assisted with supervised learning algorithm to improve the classifier's performance for the IDSs. A single hidden layer feed-forward neural network (SLFN) is trained to output a fuzzy membership vector, and the sample categorization (low, mid, and high fuzziness categories) on unlabeled samples is performed using the fuzzy quantity. The classifier is retrained after incorporating each category separately into the original training set. The experimental results using this technique of intrusion detection on the NSL-KDD dataset show that unlabeled samples belonging to low and high fuzziness groups make major contributions to improve the classifier's performance compared to existing classifiers e.g., naive bayes, support vector machine, random forests, etc.

2.1 Problem Formulation

An intrusion identification is to monitor irregular behavior and misuse in the network. Intrusion recognition was presented in 1980's after the development of internet with surveillance to monitor the risk presents in the network. The reputation and incorporation of security infrastructures are suddenly increased. From that point onward, a few activities in IDS innovation have advanced detection of network interruption in its present state. One of the focused areas to resolve cyber- attacks quickly is to detect the attack process early [1] from the network using NIDS. Network intrusion detection systems (NIDS) are designed to detect malicious activities including virus, worm, DDOS attacks.

III. MACHINE LEARNING

The main property of an ML is its capability to learn. Learning or preparing is a procedure by methods for which a neural system adjusts to a boost by making legitimate parameter modifications, bringing about the generation of wanted reaction. Learning in an ML is chiefly ordered into two classes as [9].

- Supervised learning
- Unsupervised learning

Supervised Learning

Regulated learning is two stage forms, in the initial step: a model is fabricated depicting a foreordained arrangement of information classes or ideas. The model developed by investigating database tuples portrayed by traits. Each tuple is expected to have a place with a predefined class, as dictated by one of the qualities, called to have a place with a reclassified class, as controlled by one of the traits called the class name characteristic. The information tuple are dissected to fabricate the model all things considered from the preparation dataset.

Unsupervised learning

It is the kind of learning in which the class mark of each preparation test isn't knows, and the number or set of classes to be scholarly may not be known ahead of time. The prerequisite for having a named reaction variable in preparing information from the administered learning system may not be fulfilled in a few circumstances.

Data mining field is a highly efficient techniques like association rule learning. Data mining performs the interesting machine-learning algorithms like inductive-rule learning with the construction of decision trees to development of large databases process. Data mining techniques are employed in large interesting organizations and data investigations. Many data mining approaches use classification related methods for identification of useful information from continuous data streams.

Nearest Neighbors Algorithm

The Nearest Neighbor (NN) rule differentiates the classification of unknown data point because of closest neighbor whose class is known. The nearest neighbor is calculated based on estimation of k that represents how many nearest neighbors are taken to characterize the data point class. It utilizes more than one closest neighbor to find out the class where the given data point belong termed as KNN. The data samples are required in memory at run time called as memory-based technique. The training points are allocated weights based on their distances from the sample data point. However, the computational complexity and memory requirements remained key issue. For addressing the memory utilization problem, size of data gets minimized. The repeated patterns without additional data are removed from the training data set.

Naive Bayes Classifier

Naive Bayes Classifier technique is functioned based on Bayesian theorem. The designed technique is used when dimensionality of input is high. Bayesian Classifier is used for computing the possible output depending on the input. It is feasible to add new raw data at runtime. A Naive Bayes classifier represents presence (or absence) of a feature (attribute) of class that is unrelated to presence (or absence) of any other feature when class variable is known. Naïve Bayesian Classification Algorithm was introduced by Shinde S.B and Amrit Priyadarshi (2015) that denotes statistical method and supervised learning method for classification. Naive Bayesian Algorithm is used to predict the heart disease. Raw hospital dataset is employed. After that, the data gets preprocessed and transformed. Finally by using the designed data mining algorithm, heart disease was predicted and accuracy was computed.

Support Vector Machine

SVM are used in many applications like medical, military for classification purpose. SVM are employed for classification, regression or ranking function. SVM depends on statistical learning theory and structural risk minimization principal. SVM determines the location of decision boundaries called hyper plane for optimal separation of classes as described in figure 1.4. Margin maximization through creating largest distance between separating hyper plane and instances on either side are employed to minimize upper bound on expected generalization error.

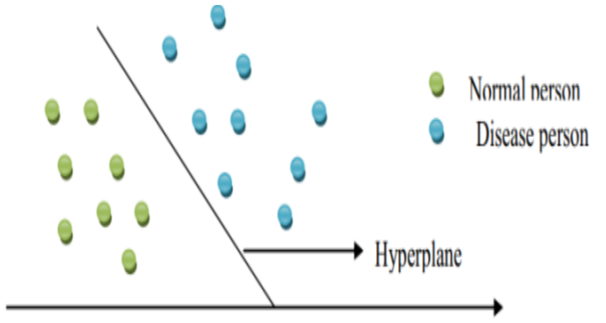

Figure 1: Support Vector Classification

Classification accuracy of SVM not depends on dimension of classified entities. The data analysis in SVM is based on convex quadratic programming. It is expensive as quadratic programming methods need large matrix operations and time consuming numerical computations.

IV. INTRUSION DETECTION SYSTEM

Like other security measures like antivirus software, firewalls, and access control plans, Intrusion Detection Systems (IDS) are designed to improve the security of information and Internet of Things communication systems. The firewall's primary function is to sort packets according to allow/deny rules based on information in the header fields. The filtering of packets that pass through particular hosts or network ports, which are typically open on the majority of computer systems, is the firewall's primary function. It doesn't do deep analysis, which is like finding malicious code in a packet, and it treats each packet as a separate thing. An anti-virus program is a running process that, rather than monitoring network traffic, examines executables, worms, and viruses in the memory of protected computer/network systems [6].

While IDS requires more embedded intelligence than other security products like antivirus programs, it analyzes the information it collects and derives useful results [7]. This is the difference between IDS and other security products like antivirus programs. DARPA established the CIDF (Common Intrusion Detection Framework) working group in 1998 with the primary goal of coordinating and defining a common framework in the IDS field. This group has produced noteworthy work [8]. A general IDS architecture based on the consideration of the four kinds of functional modules depicted in Figure 1 was developed by the group, which was incorporated into the IETF in the year 2000 and adopted the brand-new acronym IDWG ('Intrusion Detection Working Group').

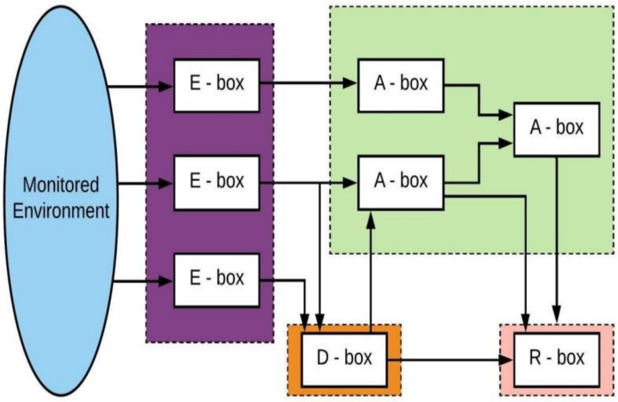

Figure 2: General CIDF architecture for IDS

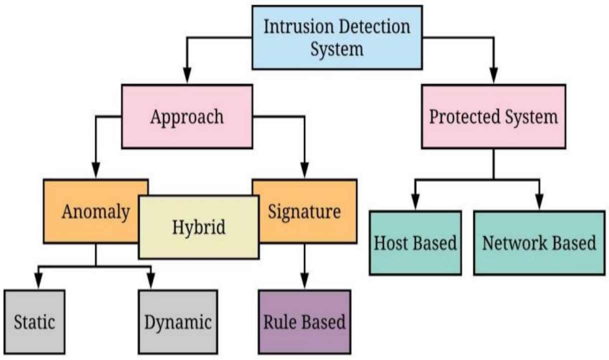

Figure 3: IDS Classifications

Contingent upon the sort of examination did, interruption location frameworks are delegated by the same token signature-based or abnormality based displayed in Figure 2. Signature-based plans (additionally indicated as abuse based) look for characterized examples, or marks, inside the dissected information. A signature database that corresponds to known attacks is specified a priori for this purpose.

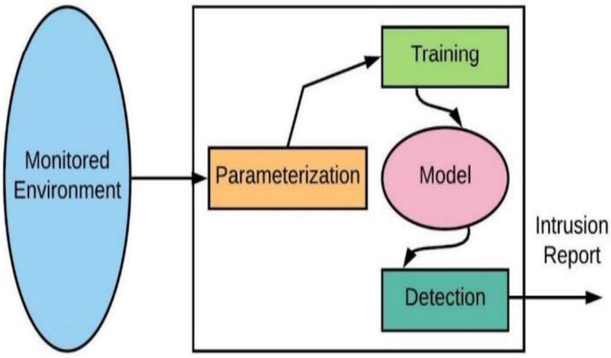

Figure 4: Generic Anomaly based IDS Functional Architecture

Anomaly-based detectors, on the other hand, attempt to estimate the "normal" behavior of the system that needs to be protected and issue an anomaly alarm whenever the difference between a specific observation and the normal behavior exceeds a predetermined threshold. Modeling the system's "abnormal" behavior and sending an alert when the difference between what is seen and what is expected falls below a certain threshold is another option. For specific, well-known attacks, signature-based schemes provide excellent detection results. Even if they are designed as minimal variants of attacks that are already known, they are unable to detect new, unknown intrusions. Contrarily, the main advantage of anomaly-based detection methods [5] is that they can pick up on intrusions that haven't been seen before. Anomaly-based Intrusion Detection Systems (A-IDS) are currently the primary focus of intrusion detection research and development due to their promising capabilities. Numerous novel plans are being considered, and numerous new systems with A-IDS capabilities are becoming available. Although there are a variety of A-IDS approaches, the fundamental modules or stages depicted in Figure 3 are common to all of them.

V. CONCLUSION

Intrusion detection systems are an important part of today's information technology-based enterprises' security. It's a difficult task to provide an efficient and high-performance IDS approach to deal with a wide range of security assaults. Deep learning approaches have recently been shown to be effective at solving intrusion detection challenges, and various deep learning-based IDS strategies have been published. Deep learning is a subset of machine learning techniques that employ multiple layers to do nonlinear processing and learn multiple levels of data representation. From this experiment found that deep learning is better than machine learning techniques. Malicious cyber-attacks can lurk in enormous amounts of legitimate data in unbalanced network traffic.

REFERENCES

[1] Pierre-Francois Marteau, “Sequence Covering for Efficient Host-Based Intrusion Detection”, IEEE Transactions on Information Forensics and Security, Vol. 14, No. 4, APRIL 2019, pp 944-1006.

[2] Rashidah Funke Olanrewaju, Burhan Ul Islam Khan, Athaur Rahman Najeeb, Ku Nor Afiza Ku Zahir and Sabahat Hussain, “Snort-Based Smart and Swift Intrusion Detection System”, Indian Journal of Science and Technology, VOL 11(4), DOI: 10.17485/ijst/2018/v11i4/120917, January 2018.

[3] Ashima Chawla, Brian Lee, Sheila Fallon, and Paul Jacob, “Host Based Intrusion Detection System with Combined CNN/RNN Model”, Springer Nature Switzerland August 2019, pp 149-158.

[4] Md. Zahangir Alom, Venkata Ramesh Bontupalli, and Tarek M. Taha, “Intrusion Detection using Deep Belief Networks”, IEEE 2015, pp 339-344.

[5] Hebatallah Mostafa Anwer, Mohamed, Farouk, Ayman Abdel-Hamid, “A Framework for Efficient Network Anomaly Intrusion Detection with Features Selection”, IEEE 2018, pp 157-162.

[6] Nutan Farah Haq, Musharrat Rafni, Abdur Rahman Onik, “Application of Machine Learning Approaches in Intrusion Detection System: A Survey”, (IJARAI) International Journal of Advanced Research in Artificial Intelligence, Vol. 4, No.3, 2015, pp 9-18.

[7] Rana Aamir Raza Ashfaq , Xi-Zhao Wang , Joshua Zhexue Huang , Haider Abbas , Yu-Lin He , “Fuzziness based semi-supervised learning approach for intrusion detection system”, Elsevier 2016, pp 484-497.

[8] Anna L. Buczak and Erhan Guven, “A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection”, IEEE COMMUNICATIONS SURVEYS & TUTORIALS, VOL. 18, NO. 2, SECOND QUARTER 2016, pp 1153-1176.

[9] JABEZ J, Dr.B.MUTHUKUMAR, “Intrusion Detection System (IDS): Anomaly Detection using Outlier Detection Approach”, Procedia Computer Science 48 2015, pp 338 – 346.

[10] M Firoj kabir,Sven Hartman, “Cyber Security:Challenges An efficient Intrusion Detection

System Design",IEEE 2018, pp 19-24.

[11] Poonam Sinai Kenkre, Anusha Pai, and Louella Colaco, "Real Time Intrusion Detection and Prevention System", Springer International Publishing Switzerland 2015, pp 405-411.

[12] Pierre-Francois Marteau, "Sequence Covering for Efficient Host-Based Intrusion Detection", JOURNAL OF LATEX CLASS FILES, VOL. 14, NO. X, FEBRUARY 2018, pp 1-14.

[13] G. Yedukondalu, J. Anand Chandulal and M. Srinivasa Rao, "Host-Based Intrusion Detection System Using File Signature Technique", Springer Nature Singapore Pte Ltd. 2017, pp 225-232.

[14] Okan CAN, Ozgur Koray SAHINGOZ, "A Survey of Intrusion Detection Systems in Wireless Sensor Networks",IEEE 2015, pp 1-6.

[15] Shijoe Jose, D.Malathi, Bharath Reddy, Dorathi Jayaseeli, "A Survey on Anomaly Based Host Intrusion Detection System", NCMTA 2018, pp 1-11.