**Intrusion Detection System with Machine Learning Techniques for Network Data Security**

Mr. Bhushan H. Talan

Student of Master of Engineering in (CSE),

Sipna College of Engineering and Technology, Amravati, India

b.talan@rediffmail.com

Prof. Harsha Vyawahare

Associate professor

Sipna College of Engineering and Technology, Amravati, India

harsha\_vyawahare@rediffmail.com

**Abstract:**

The rapid growth in the use of computer networks, connected devices, web applications and cloud computing, highlight the need for accurate and efficient network security. So to protect confidentiality, integrity and availability of data against the numerous threats and cyber-attacks some powerful security mechanism needs to be developed. Intrusion detection system (IDS) represents an unavoidable tool to secure our network. It is considered as a second defense line against the different form of attacks. The principal limits of the current IDSs are their inability to combine the detection of the new form of attacks with high detection rate and low false alarm rate. So, to avoid this it is necessary to develop the IDS based on machine learning techniques. This paper performs the survey of IDS and different machine learning techniques used to develop some advance intrusion detection and prevention system.

**Keywords:** Attacks, Datasets, Intrusion Detection System (IDS), Machine Learning (ML), Virtual Machines.

**I. INTRODUCTION**

The rapid evolution of technology implies the increased usage of computing devices with network and internet. With the tremendous grow in the usage of computers and Internet for information sharing, success of applications running on various platforms causes a serious risk to the security policy. It also imposes new cyber-security challenges. Complex behavior of malwares are also increase, the mechanism to catch malwares also needs improvement. The challenges grow towards the network security due to the continuous introduction of new attacks [1]. To tackle this growing trend in computer attacks and respond threats, industry professionals and academics are joining forces in order to build advance Intrusion Detection Systems (IDS) that combine high accuracy with low complexity and time efficiency.

Recent exploits also suggest that the more sensitive the information that is held is, the higher the probability of being a target. Earlier Study shows that, several retailers, banks, public utilities and organizations have lost millions of customer data to attackers, losing money and damaging their brand image [2]. In the second quarter of 2014, Code Spaces one of the source code company was forced out of business after attackers deleted its client databases and backups. JP Morgan, Americas‟ largest bank, suffered a cyber-attack in 2014 that impacted 76 million members [3]. In 2014, Benesse, A Japanese Education Company for children suffered a major breach whereby a disgruntled former employee of a third-party partner disclosed up to 28 million customer accounts to advertisers [4]. Most notably the “Sony Pictures hack” best displayed how significant a company’s‟ losses are in the aftermath of a security breach. The network servers were temporarily shut down due to the hack [4]. Cyber security experts estimate that Sony lost up to $100 million. To tackle this growing trend in computer attacks and respond threat, industry professionals and academics are joining forces in a bid to develop systems that monitor network traffic activity raising alerts for unpermitted activities. These systems are best described as Intrusion Detection Systems.

The main objective of this paper is to perform study for different techniques or algorithms of an intrusion detection that is based on the analysis of usage data coming from multiple partners in order to reduce the number of false alarms. On the other hand, when a new security flaws has been found on a system, the hackers will want to use it in as many information systems as possible. Thus a new anomaly that occurs on two or more information systems is probably not a new kind of usage, but rather an intrusion attempt [5]. Based on the analysis of the usage data coming from the different partners, we have to detect the common outliers they share. Such common outliers are likely to be true attacks and will trigger an alarm. This paper studied several different methods for reducing false alarm rate.

**Motivation:**

The recent advancement of network technologies has opened tremendous scope for new and improved types of vulnerabilities for the systems used all over the world. Enterprises and individuals struggle with their valuable, confidential and personal information due to this unbounded connectivity among hundreds of thousands computers. Although internet has brought significant improvements in our daily lives, it has added several cautions as well. If we want to look deep into the infrastructure of contemporary business enterprises or companies, the information technology would appear as the most prominent components which include computers, networking devices, multimedia devices, complex servers etc. In most cases organizations are highly dependent on these technology driven processes, for example to maintain employee records, to keep financial transaction records, to keep client information, and to store organization specific confidential data etc. Our day starts with checking emails, Facebook posts, tweets from twitter, checking online accounts, shopping online through credit cards, chatting with friends etc. All of these communication requires individual information exploitation over the internet and very vulnerable to be captured by other malicious users. Thus we need strong surveillance systems in the form of real time intrusion detection systems to protect our valuable information from these attacks.

**II. INTRUSION DETECTION SYSTEM (IDS)**

Intrusion detection system (IDS) is a combination of software and hardware that monitor and examines the actions occurring in network system to distinguish intrusions. Intrusion detection using data-mining plays a Key role of analyzing, processing and sorting the data to a systematic and organized manner without any errors. Traditional IDS search for possible malicious events on network traffics and occasionally finds real security attacks and abnormalities. But, many times fails to distinguish malicious actions or identify normal traffic as attack type in the network. To enhance the capability of intrusion detection system, Data mining methods are applied on network data because of it can process great volume of data and User’s subjective evaluation is not necessary, and it is more suitable to discover the ignored and hidden information.

**Taxonomy of IDS:**

A general definition of Taxonomy is the practice or principle of classification [6]. Taxonomy may serve several purposes in design. Firstly, it can describe the current global situation, assisting in refining complex situations and presenting it in a clearer description. Additionally using taxonomy to classify a number of objects, enables identification of missing objects early in design, which allows users to exploit the predictive qualities of a good taxonomy (Prediction). Lastly, a good taxonomy presents users with ideas, further explaining observed current occurrences (Explanation).

The identified taxonomies can be used in order to illustrate general relationships in IDSs‟. Figure 1 which is illustrated below displays the revised version of the taxonomy previously proposed, this version features additional criteria for classifications [6].



Fig. 1. Updated IDS taxonomy

**Types of Anomaly Detection:**

*A. Supervised Anomaly Detection*

When both normal and anomaly data is labeled for training and to make a predictive model, we call it supervised anomaly detection model. In this case, any unseen data instances are compared against the model to decide which class it belongs to. But, there are some issues found related to the imbalance of normal and anomaly dataset. This was solved using machine learning and data mining techniques. Another issue is to find the optimized labels for the anomaly class which is really challenging. It was addressed in [7] using artificial anomaly injection in the normal dataset to produce labeled training data set.

*B. Semi Supervised Anomaly Detection*

On the other hand, semi supervised anomaly detection techniques use only one labeled class for normal data for the training purpose. In most cases normal class is modeled due to the complexity of modeling anomaly class. For example, in road accident detection, an anomaly scenario would add the occurrence of an accident, which is complex to model. There are several systems as well those rely on using the training data set as anomaly set only. These systems lack predictability because it is not easy to create some anomalous dataset that will cover every possible abnormal behavior within the system.

*C. Unsupervised Anomaly Detection*

Unsupervised anomaly detection techniques do not require any training data. These techniques assume that the frequency of normal data is far greater than the frequency of anomalous data. Thus, the techniques can end up providing high false positive rate if the primary assumption violates.

**Intrusion Detection Methods:**

Following are main intrusion detection methods: that mostly comes under misuse and anomaly detection

*A. Misuse Detection*: This technique identifies intrusions based on known patterns for the malicious activity, known patterns are referred to as signatures that represents specific threat [8]. But has limitation to detect new threats whose signatures are not yet identified.

*B. Anomaly Detection:* This technique identifies intrusions based on deviations from established normal behavior. Anomaly detection builds models of normal network events and detects the events that deviate from these models. This method can detect new types of attack events because it only relies on known normal events [8], but has high false positive rate. There are many hybrid approaches which overcomes the disadvantages of misuse and anomaly methods.

**Intrusion Detection Systems:**

*A. Host IDS (HIDS):* Performs intrusion detection from within host it is monitoring have poor visibility of the internal state of the host machine Difficult for malicious code (malware) to bypass the HIDS. Disadvantage: Susceptible to attacks by malware.

*B. Network IDS (NIDS)* Performs intrusion detection through network connections and outside the host machine are more resistant to attacks by malware. Difficulty is poor visibility of the internal state of the host machine Easier for malware to bypass the NIDS.

*C. VMM-based IDS* (VMM-IDS): Performs intrusion detection for a virtual machine through the Virtual Machine

Monitor (VMM) perform better visibility of the internal state of the host machine, compared to an NIDS harder for malware to bypass the IDS less susceptible to attacks by malware [9]. Virtual Machine Monitor (VMM) or hypervisor is the software layer providing the virtualization which allows the multiplexing of the underlying physical machine between different virtual machines, each running its own operating system.

**III. RELATED WORK**

In this work, to improve the performance of the classifiers and identifying the intrusions in dataset the hybrid data mining approach based on ensemble classifier is utilized. The WEKA machine learning tool is utilizing for implementing the hybrid model. The relevant features are selected based on their information gain value. The effects of feature selection are elevated using the hybrid classification method based on random forest, random tree and Bayes net. The results shows that the hybrid approach achieved better classification accuracy and gives better Precision, Recall and F-measure for R2L and U2R attack types. This preferred approach gives a hybrid classifier which improves the overall detection rate of the system. Preferred approach gives more accuracy and decrease the false positive rate. Hybrid classification approach is best for Root-kit attack, where single random forest is not identifying this type of attacks [8].

In paper [10] authors introduces a technique for IDS based on fuzzy clustering and ANN approach. This method could be applicable to overcome the issues of weak stability detection as well as low precision detection. The restore point in this method was employed for registry keys, system files roll back, project database and installed programs. Fuzzy clustering will generate different subsets for training in order to reduce the amount of subset size and complexity. Then each subset is trained with different type of artificial neural network and finally processed to obtain significant results.

Paper [11] suggested a novel back propagation model for intrusion detection. This method makes training pair with a combination of input and equivalent target were generated and implemented into the network. Performance success can be measured by false alarm and detection rate. Detection rate was proven to be less than 80% for U2R, R2L, DoS and Probe attacks. However, the major issue of the method was found to be much inefficient to detect hidden attackers present in the system.

In this paper [12], author presented the details of a new approach called outlier detection where, the anomaly dataset is measured by the Neighborhood Outlier Factor (NOF). The training model consists of big datasets with distributed environment that improves the performance of Intrusion detection system. This approach is also been tested with the KDD datasets that are received from real world. The machine learning approaches detect the intrusion in the computer network with huge execution time and storage to predict the when compared to the proposed IDS system which takes less execution time and storage to test the dataset .Here in this study, the performance of proposed IDS is better than that of other existing machine learning approaches and can significantly detect almost all anomaly data in the computer network. Author said that, in future the proposed work can be possibly used for various distance computation function between the trained model and testing data.

Author in paper [13], proposed a two stage classifier based on the REPTree (Reduced Error Pruning Tree) algorithm and protocols subset for network intrusion detection system. The NSL-KDD dataset and UNSW-NB15 dataset were used to evaluate the performance of our novel detection algorithm. Network traffic if firstly divided into different classes according to the different network protocol. In the first stage the incoming network traffic is classify into normal or attack classes. In case of attack traffic, the second classifier identifies the type of the attack for providing the best necessary response. Extensive evaluation and comparison results showed that the proposed two stage classifier model yields better results in terms of speed of detection and prediction accuracy rate. But the attacks classification experiments on both NSL-KDD and UNSW-NB15 are still not perfect especially for UDP and Other protocols. So it is necessary to improve the detection accuracy in these protocols.

The author in this paper [14] proposed method that is used for application recognition of flows resources with the help of SDN and data mining techniques based on machine learning. Applying traffic classification techniques to the network makes the network be application-aware, and enables the network to know flow’s requirements. Due to maintaining a global view of the network, the controller could dynamically allocate bandwidth to flows on demand and thus improve their QoS and the analysis and prediction of traffic patterns in network make the controller further optimize resource allocation. This method mainly focused on minimizing controllers’ processing overhead and network traffic overhead for network traffic classification. This method does not impose any processing overhead to the controller because unlike the base method, packets contents are not checked. But they develop method is work of single format so it is necessary to include implementations on different device platforms and detection of flows belonging to a new application which is not part of the trained classifier.

**IV. DIFFERENT MACHINE LEARNING TECHNIQUES**

Machine learning, a branch of artificial intelligence, is a scientific discipline concerned with the design and development of algorithms that allow computers to evolve behaviors based on empirical data, such as from sensor data or databases. A major focus of machine learning research is to automatically learn to recognize complex patterns and make intelligent decisions based on data. Machine Learning has lots of applications, it is the starting point for much of the work using ML techniques in Internet traffic classification. Following are some of the data mining machine learning algorithmic techniques that are used of IDS.

**A. Naïve Bayes**

The naïve Bayes model is a heavily simplified Bayesian probability model [15]. In this model, consider the probability of an end result given several related evidence variables. The naïve Bayes classifier operates on a strong independence assumption [15]. This means that the probability of one attribute does not affect the probability of the other. Given a series of n attributes as in case of different IDS, the naïve Bayes classifier makes 2n! Independent assumptions. With the built patterns, the framework detects attacks in the datasets using the naïve Bayes Classifier algorithm. Nevertheless, the results of the naïve Bayes classifier are often correct. But as a naïve Bayesian network is a restricted network that has only two layers and assumes complete independence between the information nodes. This poses a limitation to naïve Bayes, in order to alleviate this problem so as to reduce the false positives, active platform or event based classification.

**B. Bayesian networks**

Bayesian networks represent a new approach to detection and prevention of attacks in computer networks [16]. Bayesian Networks allow the representation of causal dependencies between random variables in graphical form and permit the calculation of the joint probability distribution of the random variables by specifying only a small set of probabilities, relating only to neighboring nodes.

System using Bayesian network offers a unique advantage over other systems when one calculates the influence of newly produced events on the other observed events; accordingly, all data and rules used in other systems can be built into IDS based on Bayesian networks. Bayesian networks provide a full compatibility of corresponding software products without respect to platform used for execution; this fact can speed-up the development and application of standalone and distributed IDS.

**C. Decision trees**

Decision trees (DTs) are popular in misuse detection systems, as they yield good performance and offers some benefits over other machine learning techniques. For example, they learn quickly compared with Artificial Neural Networks (ANNs), and DTs are not black boxes. DTs have been successfully applied to intrusion detection both as a standalone misuse detector and as a part of hybrid systems [17]. A good example of the success of DTs is an application of a C5.0 DT by P fahringer, which won the KDD Cup competition with bagging and boosting.

**D. Support Vector Machine (SVM)**

The SVM is already known as the best machine learning algorithm for binary classification. The SVM, originally a type of pattern classifier based on a statistical learning technique for classification and regression with a variety of kernel functions, has been successfully applied to a number of pattern recognition applications. Recently, it has also been applied to information security for intrusion detection. Support Vector Machine has become one of the popular techniques for anomaly intrusion detection due to their good generalization nature and the ability to overcome the curse of dimensionality. Another aspect of SVM is that it is useful for finding a global minimum of the actual risk using structural risk minimization, since it can generalize well with kernel tricks even in high-dimensional spaces under little training sample conditions [18]. One of the main advantage of using SVM for IDS is its speed, as the capability of detecting intrusions in real-time is very important. SVMs can learn a larger set of patterns and be able to scale better, because the classification complexity does not depend on the dimensionality of the feature space. But at the same time, as the SVM being a supervised machine learning method requires labelled information for efficient learning. Pre-existing knowledge is required for classification which may not be available all the time. Training of SVM is time-consuming for IDS domain and requires large dataset storage. Thus SVM is computationally expensive for resource-limited ad hoc network [18]. Moreover as SVM requires the processing of raw features for classification which increases the architecture complexity and decreases the accuracy of detecting intrusion.

**V. CONCLUSION**

As the use of computer and internet and over the network devices is going on increasing, leading to different network attacks. This paper performs the study of network Intrusion detection system with their types and methods. Different authors used different techniques to deal with this network attacks and protects the computing devices and important data from lost with different kinds of network attacks, review of them is also perform here. As machine learning techniques are necessary to deal with current networks flows and it becomes most important with current network scenarios. This paper also perform the study of different data mining and machine learning techniques that are useful for performing Intrusion detection system.

**REFERENCES**

[1] Mohammad Reza Parsaei, Samaneh Miri Rostami, Reza Javidan, “A Hybrid Data Mining Approach for Intrusion Detection on Imbalanced NSL-KDD Dataset”, *International Journal of Advanced Computer Science and Applications (IJACSA),* Vol. 7, No. 6, 2016.

[2] L. Morgan (2014), List of Cyber Attacks and Data Breaches in 2014. IT Governance, 23 Dec. Available from: http://www.itgovernance.co.uk/blog/list-of-the-hacks-and-breaches-in-2014/ [Accessed on 13 November 2015]

[3] M. Watson, (2014). JP Morgan suffers data breach affecting 76 million customers. IT Governance, 23 Dec. Available from: http://www.itgovernanceusa.com/blog/jp-morgan-suffers-data-breach-affecting-76-million-customers/ [Accessed on 14 November 2015].

[4] “Report and response regarding Leakage of Customers‟ personal Information.” (10 September 2014). Last accessed on 17 February 2015,

http://blog.benesse.ne.jp/bh/en/ir\_news/m/2014/09/10/uploads/news\_20140910\_en.pdf.

[5] S. Sathya Bama, M. S. Irfan Ahmed, A. Saravanan, “Network Intrusion Detection using Clustering: A Data Mining Approach”, *International Journal of Computer Applications (0975 – 8887) Volume 30– No.4, September 2011.*

[6] Zibusiso Dewa and Leandros A. Maglaras, “Data Mining and Intrusion Detection Systems”, *International Journal of Advanced Computer Science and Applications (IJACSA),* Vol. 7, No. 1, 2016.

[7] Md. Enamul Haque and Talal M. Alkharobi, “Adaptive Hybrid Model for Network Intrusion Detection and Comparison among Machine Learning Algorithms”, *International Journal of Machine Learning and Computing (IJMLC),* Vol. 5, No. 1, February 2015.

[8] R. Richhariya1, A.K. Manjhwar, R. R. Singh Makwana, “A Hybrid Approach for User to Root and Remote to Local Attack”, *International Journal of Computer Sciences and Engineering (IJCSE),* E-ISSN: 2347-2693, Volume-5, Issue-6, 2017.

[9]Amjad Hussain Bhat, Sabyasachi Patra, Dr. Debasish Jena, “Machine Learning Approach for Intrusion Detection on Cloud Virtual Machines”, *International Journal of Application or Innovation in Engineering & Management (IJAIEM),* Volume 2, Issue 6, June 2013 ISSN 2319 – 4847.

[10] Gaikwad, Sonali Jagtap, D.P. Kunal Thakare and Vaishali Budhawant. Anomaly Based Intrusion Detection System Using Artificial Neural Network and fuzzy clustering, International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, November- 2012; 1(9).

[11] Jaiganesh, V., Sumathi, P. and Mangayarkarasi, S., An Analysis of Intrusion Detection System using back propagation neural network, IEEE Computer Society Publication, 2013.

[12] JABEZ J, Dr. B. MUTHUKUMAR, “Intrusion Detection System (IDS): Anomaly Detection using Outlier Detection Approach”, International Conference on Intelligent Computing, Communication & Convergence, Procedia Computer Science 48 ( 2015 ) 338 – 346, Available online at www.sciencedirect.com

[13] Mustapha Belouch, Salah El Hadaj, “A Two-Stage Classifier Approach using RepTree Algorithm for Network Intrusion Detection”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No. 6, 2017.

[14] Mohammad Reza Parsaei, Mohammad Javad Sobouti, Seyed Raouf khayami, Reza Javidan, “ Network Traffic Classification using Machine Learning Techniques over Software Defined Networks”, International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 8, No.7, 2017.

[15]Mrutyunjaya Panda and Manas Ranjan Patra, “NETWORK INTRUSION DETECTION USING NAÏVE BAYES”, IJCSNS International Journal of Computer Science and Network Security, VOL.7 No.12, December 2007.

[16] MILAN TUBA, DUSAN BULATOVIC, “Design of an Intrusion Detection System Based on Bayesian Networks”, WSEAS TRANSACTIONS on COMPUTERS, ISSN: 1109-2750, Issue 5, Volume 8, May 2009.

[17] O. Depren, M. Topallar, E. Anarim and M.K. Ciliz. 2005. An Intelligent Intrusion Detection System for Anomaly

and Misuse Detection in Computer Networks. Expert systems with Applications, 29, 713-722.

[18] Jayshree Jha, Leena Ragha, “Intrusion Detection System using Support Vector Machine”, *International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868 Foundation of Computer Science FCS, New York, USA International Conference & workshop on Advanced Computing 2013 (ICWAC 2013) –* [*www.ijais.org*](http://www.ijais.org)*.*