A Systematic Literature Review on Intrusion Detection System using Machine Learning

[1]M.Nikhitha, [2]M.A.Jabbar
[1]Research Scholar, [2]Professor
[1]Computer Science Department,
[1]Vardhaman College of Engineering, Hyderabad, India.

Abstract: Network security has become more prominent with the advancement of the internet over years and the number of attacks also increased. Ongoing mitigation against various attacks remains unable to provide complete protection. A powerful detection system is required to ensure security of the network. Intrusion detection system (IDS) is active systems, which supervises and examine the network traffic to identify any deviations from network traffic. Various Machine Learning techniques are used to build an effective Intrusion Detection System. This paper gives a comprehensive review on Intrusion Detection System and discusses about various approaches to detect the attacks and their working procedures.

***Index Terms* - : Intrusion detection system, Attack, Security, Anomaly, Effective, processing.**

I. INTRODUCTION

Over the years, the usage of internet and the number of devices connected to the internet are also increasing with the advancement of the internet**.** With the help of internet, the attackers are gaining the access in order to use or destroy the unauthorized data [3]. So we need a powerful tool or software to protect the system from harmful attacks. The main purpose of IDS is to detect any a suspicious activity in the network and alarm the administrator. The network administrators use IDS which is a software or a device that monitors the network traffic by investigating different areas of networks looking for a suspicious activity which may represent an attack or unauthorized. IDS architecture is shown in below figure 1.

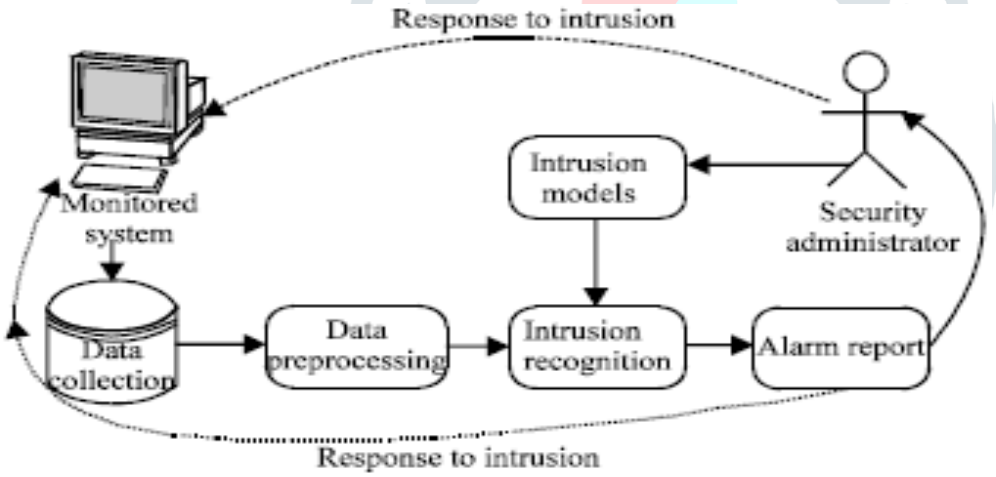

Figure 1: Intrusion detection system architecture [16]

Machine learning techniques play an important role in identifying network intrusions (or attacks) [10]. Intrusion Detection is a cyber security mechanism, to secure the network. The difficult things in developing Intrusion Detection System (IDS) are false positive rate and detection rate [21]. Machine learning algorithms are used in Intrusion detection system to reduce error rate and build effective IDS [20].

This paper discussed recent literature, reviews of existing literature and research efforts on IDS. The literature review has focused on the works done by different authors in the area of IDS with the help of scientific repositories like IEEE Explore, Scopus, Springer, Science Direct etc [5].The remaining paper is organized as follows:
Section II discussed about the background study.
Section III describes about the existing works on IDS.
Section IV presents a set of conclusion and future work considerations.

II. BACKGROUND STUDY

2.1 MACHINE LEARNING

Machine learning (ML) is a subset of artificial intelligence (AI) that provides the capability to educate itself without human intervention. The primary aim is to allow the computers to learn automatically and improve from experience [17]. Machine learning tools can detect key features from complex datasets and present their performance. In general, the machine learning algorithms are classified into Supervised, Unsupervised and Reinforcement. Different types of machine learning algorithms are shown in figure 2.

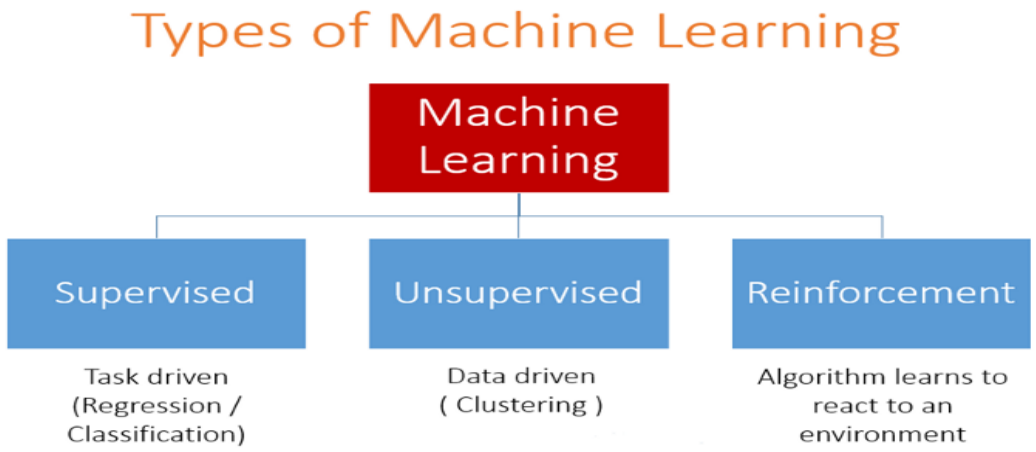

Figure 2: Different types of machine learning algorithms [22].

a. SUPERVISED: learns from past data, which can be applied to new data and predict future events by using labeled examples. It analyses the training data and produces an inferred function, which can be used to predict the output values based on target variables. The model compares its output with the desired output and also finds errors in order to modify accordingly [17]. It is further categorized into two types Classification and Regression.

b. UNSUPERVISED: is used to train the information which is not labeled or classified. Unsupervised learning can group and interpret data on the basis of input data. It is mainly used to explore the data and for discovering the hidden structure of data, but it cannot figure out right output. It is further grouped as (i).Clustering.

c. REINFORCEMENT: this method aims by using observations gathered from the interaction with the environment in order to take actions that would minimize the risk and maximize the reward. Reinforcement learning algorithm (agent) continuously learns in an iterative fashion from the environment. In this process, the agent from its experience until it explores the full range of possible states.

2.2. IDS (INTRUSION DETECTION SYSTEM)

Intrusion Detection System (IDS) monitor the networks or other systems for harmful or abnormal performances. Enhancing about preventive technologies like firewalls, Strong validation, and user advantage. IDS main goal is to protect the data's integrity confidentiality and availability of resources [9]. It is categorized as misuse or anomaly. Misuse is an inside attack and anomaly or intrusion is the outside attack. There are two types of IDS.

1. HOST BASED IDS: It is attached or deployed on the individual device in order to monitor the device. It not only monitors but also analyzes whether anything internal or external, has evaded the system security.

2. NETWORK-BASED IDS: It connects to more than one network modules and observes network traffic for un-usual actions from all segments over the network [5]. Depend on network-based detection IDS can be signature based or not [3].

a. SIGNATURE BASED IDS: follows a well-defined patterns and signatures, which are previously encoded and stored in IDS internal database. If any network attack matches with stored pattern or signature then an alert will be trigged. Whereas it is unable to detect new attacks because the matching signature of this attack is unknown [5].

b. ANAMOLY BASED IDS: The activities of a system are trained as normal network traffic and generates the model. The model is used to classify the new events or objects as normal or anomaly. A deviation from normal behavior is considered as an intrusion [5].

Types of IDS attacks are shown in figure 3. There are 4 different types of attacks that can occur in an IDS system [11][12], they are:

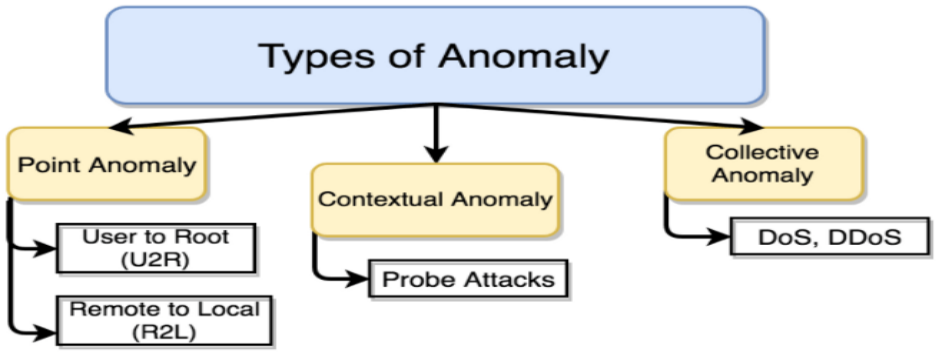

Figure 3: Types of IDS attacks [15]

1. DOS ATTACK: DOS is a kind of attack in which a resource is made unavailable to the intended user. So attack with respect to resource unavailability can be said to a DOS attack. This sort of attack targets the provision of the network. To detect DOS attack it is not be important to know whether "user logged in or not".
EX: Email bombs, Mail blood Smurf, UDP storms.

2. PROBE ATTACK: These attacks are focused at gaining information about the selected network for the purpose of identifying the possible vulnerabilities that may be utilized later so as to compromise the system.
EX: Satan, Saint, Mscan, Nmap.

3. REMOTE TO USER ATTACK (R2L**):** To obtain the benefits of local user to send packets to the networks, so the attackers remotely exploit vulnerabilities in a target system**.**
EX: imap, send-mail, multihop.

4. USER TO ROOT ATTACK (U2R): The intruder starts by obtaining the illegal access to the administrator account on the system by exploiting the vulnerability.
EX: Eject, Ffbconfig, Ps, and Perl.

IDS have two reaction modes after detecting the threats Passive mode and Active mode. Passive IDS will observe and analyze the actions of network. If find any threat then it only cautions. Active IDS not only observes the network traffic but also response to the threat.

2.3. Machine Learning in IDS:

Machine learning techniques are widely used to build IDS to have high accuracy and low error rate**.** In the past decade, however, for the hope of improving detection rates and adaptability several techniques have been involved in the intrusion detection. These techniques are often used to keep the attack knowledge bases up-to-date and comprehensive [13]. Here are a few of the most commonly used models are Decision Tree, K-means clustering**,** Neural Networks, Nearest Neighbor, Naive Bayes etc.

III. RELATED WORK

Over the years, various papers have been published on IDS using different Data Mining techniques or Machine Learning techniques. We mainly focus our attention on those works particularly targeting IDS by using different classifiers. Our literature review of IDS is mainly concerning the following features like accuracy, detection rate, and security threat.

In 2018 L. Haripriya and M. A.Jabbar proposed a novel IDS using ANN and feature subset selection [1].where the main objective is to achieve high accuracy in classification of attacks, in order to implement IDS, the authors adapted back propagation algorithm in ANN to classify the attacks along with feature subset selection on KYOTO dataset. The experiment result showed an increase in accuracy, precision, recall, and F-measure.

On their 2018 paper, Alemtsehay Adhanom and Henock Mulugeta proposed a hybrid behavioural based IDS [2]. The authors proposed hybrid detection by combining Knowledge-based IDS and Anomaly-based IDS along with two techniques Decision Tree and Association Rule Mining. Combination of detection methods could find more number of attacks. The main feature of BBCNIDS is hybrid applications that detect both stored and unsorted attacks. NSL-KDD dataset is used for training and testing. The results show that it provides an improved detection rate and lower false alarm rates.

Also in March 2018 Majid Latah, Levent Toker presents an IDS for Software-Defined Networks (SDN) [6]. The objective is to develop efficient anomaly-based intrusion detection for SDN. The authors investigated the performance of anomaly-based ID approach in terms of accuracy, false positive rate, area under ROC curve and execution time mainly and they focus on supervised learning approaches along with classifiers like a Decision tree, Naive Bayesian, Neural network, nearest neighbour , Random forest (RF)and SVM (Support vector machine). In this approach, the authors used NSL-KDD dataset and concluded that KNN is best in accuracy and area under ROC curve and the random forest is best in false positive rate.

In 2016 Nabila Farnazz and M.A.Jabbar proposed IDS [7] to identify the user actions using Random Forest classifier. The classifier has less classification errors than other algorithms. Feature selection process is used to reduce data dimensionality. The experimental result was done on dataset NSL-KDD and the model is efficient for high accuracy and classification of attacks.

Jayshree Jha and Leena Ragha in 2013 have developed an IDS using Support Vector Machine [8].SVM is used due to their good generalization nature and ability to overcome the curse of dimensionality. This research uses a hybrid approach for feature selection process. This will enhance the performance and accuracy of the SVM model. This model presents a study that incorporates Information Gain and K-mean algorithm to SVM for intrusion detection.

The summary of various techniques proposed by different authors are shown in table 1:

Table1: Different Algorithms used in IDS

Author Name	**Proposed Model**	**Algorithm Used**	**Dataset Used**	**Metrics Used**	**Values**
L.Hari Priya and M.A.Jabbar, 2018. [1]	A Novel IDS using ANN and Feature subset selection	Back Propagation Algorithm	KYOTO dataset	Accuracy Precision Recall	Accuracy=98.8 Precision=99.6 Recall=99.3 measure=98.5
Alemtsehay adhanom and Henock Mulugeta, 2018. [2]	Hybrid Behavioural Based IDS	Decision Tree and Association rule mining algorithm	NSL-KDD dataset	Accuracy Detection Rate	Accuracy=99.85 Detection Rate=99.88
Majd Latah and Levent Tokker, 2018. [6]	Efficient Anomaly Based ID for SDN	Different Supervised MLA	NSL-KDD dataset	Accuracy False Positive Rate	Accuracy(KNN)=98.23 False Positive Rate(RF)=0
Nabil Farnaaz and M.A.Jabbar, 2016. [7]	IDS using Random Forest	RF Algorithm	NSL-KDD dataset	Accuracy Detection Rate False Alarm Rate	Accuracy=99.6 Detection Rate=99.8 False positive Rate=0.00502
Jayshree Jha and Leena Ragha, 2013.[8]	IDS using SVM	SVM Algorithm	NSL-KDD dataset	Accuracy	Accuracy=99.37

IV. CONCLUSION

In this paper, we discussed the latest literature review on IDS and examined the kinds of techniques used for IDS. Several machine learning classifiers like SVM (support vector machine), Naive Bayesian, Neural Networks, and RF (random forest) etc has been discussed for effective intrusion detection. There are still many aspects of intrusion detection that require more research to make IDS's effective and usable The performance improvement in different data mining methods have been shown by researchers. A further enhancement to this study is improving the speed of processing and accuracy.

REFERENCES:

[1]. L.haripriya and M.A.Jabbar, A Novel intrusion detection system using ANN and feature subset selection, international journal of engineering and technology, 2018.

[2]. Alemtsehay Adhanom and Henock Mulugeta, A hybrid behavioral based cyber intrusion detection system, Int. J. Communication Networks and Distributed Systems June 2018.

[3]. Nabeela Ashraf and Waqar Ahmad, A comparative study of data mining algorithm for high detection rate in IDS, AETiC January 2018.

[4]. Fadi Salo, Mohammad Noor, Data mining techniques in Intrusion detection system: A systematic literature review, IEEE august 2018.

[5]. Lionel Santos and Carlos Raba Dao, Intrusion Detection system in Internet of Things: A literature review, Portuguese National Fund through FCT, 2016.

[6]. Majd Latah and Levent Tokker, Towards an Efficient Anomaly-Based Intrusion Detection for Software-Defined Networks, arXiv: 1803.06762v1 March 18, 2018.

[7]. Nabil Farnaaz and M.A.Jabbar, Random Forest Modeling for Network Intrusion Detection System, ELSEVIER, Science Direct 2016.

[8]. Jayshree Jha and Leena Ragha, Intrusion Detection System using Support Vector Machine, International Conference & workshop on Advanced Computing 2013.

[9]. Kavitha.N, Blessy Boaz, Survey on Intrusion Detection System Using Data Mining Techniques, International Journal of Innovative Research in Science, Engineering and Technology 2017.

[10]. M.A.Jabbar, Rajanikanth Aluvalu and S.Sai Satyanarayana Reddy, Intrusion Detection System Using Bayesian Network and Feature Subset Selection, 2017.

[11]. Ms.sonali et.al, Research Paper on Basic of Artificial Neural Network, International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 2 issue: 1 96-100.

[12]. K.KanakaVardhini et.al, Enhanced Intrusion Detection System using Data Reduction: An Ant Colony Optimization Approach, International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 9 (2017) pp.1844-1847.

[13]. Mahdi Zamani, Machine Learning Techniques for Intrusion Detection, December 2013.

[14]. Mohammad Almseidin, Maen Alzubi, Szilveszter Kovacs and Mouhammad Alkasassbeh, Evaluation of Machine Learning Algorithms for Intrusion Detection System, arxiv.org on 8 January 2018.

[15]. https://medium.com/@EbubekirBbr/ids-ips-with-ml-3761cc44ac5

[16]. V.Moraveji Hashemi,Z.Muda and W.Yassin, Improving Intrusion Detection Using Genetic Algorithm. Science Alert, Research Article in 2013.

[17]. www.expertsystem.com/machine-learning-definition.

[18]. https://ecmapping.com/2018/02/21/the-10-machine-learning-algorithms-to-master-for-beginners.

[19]. https://searchenterpriseai.techtarget.com/definition/machine-learning-ML.

[20]. Suad Mohammed Othman,Fadl Mathaher Ba-Alwi, Nabeel T.Alsohybe and Amal, Y. Al-Hashida, Intrusion Detection model using machine learning on Big Data environment ,journal of big data, Othman et al. J Big Data (2018) 5:34.

[21]. M.A.Jabbar, Rajinikanth Aluvalu, S.Sai Satyanarayana Reddy, Cluster Based Ensemble Classification for Intrusion Detection System, ICMNC 2017,pp 253-257,2017.

[22]. https://www.analyticsvidhya.com/blog/2015/06/machine-learning-basics/