The National Law Institute University, Bhopal

Offers Programme

Master

Of

Cyber Law and Information Security

Project On

"Effective Intrusion Detection System using Machine Learning"

SUBJECT

"Cyber Operation Security"

SUBMITTED TO:

Mr. Ankur Rajput

SUBMITTED BY:

Isha Gaur

2019 MCLIS 58

Table of Contents

ABSTRACT

The popularity of using Internet contains some risks of network attacks. Intrusion detection is one major research problem in network security, whose aim is to identify unusual access or attacks to secure internal networks. For a while intrusion detection systems have been approached by various machine learning techniques. All current IDSs are switching to machine learning techniques to combat ever-increasing security threats to networks. This not only automates the process of intrusion detection but does so with astonishing accuracy. In this paper, all four analytics techniques of machine learning and their implementation in Intrusion Detection systems will be discussed to develop an architecture for an IDS of future generation.

Keywords: Intrusion detection, machine learning, deep learning, analytic techniques.

1. INTRODUCTION

In this Internet era, organizational dependence on networked information technology and its underlying infrastructure has grown explosively. In conjunction with this growth, the frequency and severity of network-based attacks have also drastically increased. Despite concerted efforts on preventative security measures, vulnerabilities remain. These are due to programming errors, design flaws in foundational protocols, and the "insider" abuse problem of legitimate users misusing their privileges. Because of this, intrusion detection (ID), the subdiscipline of information security that monitors network events for signs of malicious or abnormal activity, has become an integral component in many organizations' approaches to security. Intrusion detection systems (IDS) assist security analysts by automatically identify potential attacks from network activity and produce alerts describing the details of these intrusions.

2. STATEMENT OF PROBLEM

Existing Intrusion Detection Systems lack the ability to detect unknown attacks as well as do not prescribe the solution or action to predicted intrusions. Because network environments change quickly, attack variants and novel attacks emerge constantly. Thus, it is necessary to develop IDSs that can detect unknown attacks and also prescribe the solution to them. To address the above problems, In this project we will be analysing how predictive and prescriptive analytics could be used to improve the efficiency of existing IDSs.

3. RESEARCH IMPLEMENTATIONS

3.1 LITERATURE REVIEW

For writing this paper we have gone through various books, case studies, operational security incidents and articles related work done on the area of Intrusion detection systems as a security solution. They are **Supervised and Unsupervised Machine Learning in Security Applications book by Charles Givre** in which it is described how machine learning algorithms are playing role in providing security solutions. What kind of security problems can be addressed by machine learning and how machine learning fits in security environment of an organization. Another book **Machine Learning and Security by Clarence Chio and David Freeman** in which they explored a range of different computer security applications for which machine learning has shown promising results. Also, they discussed the most common issues faced by practitioners when designing machine learning systems, whether in security or otherwise. **Machine Learning for Cybersecurity Cook book by Emmanuel Tsukerman** in which he described how to apply modern AI to create powerful cybersecurity solutions for malware, pentesting, social engineering, data privacy, and intrusion detection which would helping creating a machine learning based IDS. **A Case Study on Using Deep Learning for Network Intrusion Detection by Gabriel C. Fernandez and Shouhuai Xu** in which they proposed using a feedforward fully connected Deep Neural Network (DNN) to train a NIDS via supervised learning. They also proposed using an autoencoder to detect and classify attack traffic via unsupervised learning in the absence of labelled malicious traffic and they evaluated

these models using two network intrusion detection datasets with known ground truth of malicious vs. benign traffic.

3.2 OBJECTIVES OF STUDY

- To identify how machine learning would increase the effectiveness of Intrusion Detection Systems.
- To analyse how four types of analytic techniques would help in building IDS.
- To compare the existing IDS with the advanced machine learning based IDS.
- To analyse currently faced issues with the existing IDS.
- To find out how Machine learning based IDS would change the security solutions approach with the modern technology.

3.3 HYPOTHESIS

Predictive and prescribe analytics techniques are the future of existing security solution Intrusion Detection System.

3.4 RESEARCH QUESTIONS

- What are the algorithms used in machine learning based IDS?
- What is the effectiveness of IDS using Machine Learning?
- What would be the changes in the existing way of proving security solution in network security after adapting advance technologies?
- How machine learning based IDS would benefit the organization?

3.5 RESEARCH METHODLOGY

The research methodology adopted for this paper is Doctrinal.

4. INTRUSION DETECTION SYSTEM

Intrusion Detection Systems (IDS) are used to identify malicious network activity leading to confidentiality, integrity, or availability violation of the systems in a network. Many intrusion detection systems are specifically based on machine learning techniques due to their adaptability to new and unknown attacks. In recent days, the demand for cyber security and protection against various types of cyber-attacks has been ever increasing. The main reason is the popularity of Internet-of-Things (IoT), the tremendous growth of computer networks, and the huge number of relevant applications that are used by individuals or groups for the purpose of either personal or commercial use. Cyber-attacks such as the denial-of-service (DoS) attack[1],

[1] A DoS is an attack which is launched to make networks' and systems' resources unavailable for the legitimate users so that no one else can access it. Hackers can create a situation in which the organizations come to a grinding halt. The main targets of these attacks are web servers, default gateways, personal computers, etc. Available at

computer malware[2], or unauthorized access led to irreparable damage and financial losses in large-scale networks. A cyber security system typically consists of a network security system and a computer security system. Although various systems, such as firewall and encryption, are designed to handle Internet-based cyber-attacks, an intrusion detection system (IDS) is more capable of resisting the computer network from external attacks. Thus, the main purpose of an IDS is to detect various types of malicious network communications and computer systems usage for prevention.

Providing security requires an integration of tasks that include ID, preventative technologies for "hardening[3]" systems, implementing encryption and authentication schemes, and educating users in safety-smart work practices. The work of ID involves more than reviewing IDS alerts and occasionally responding to critical events. ID itself cannot be accomplished effectively in isolation, it also requires monitoring and analysing systems tangential to the IDS, as well as keeping abreast of the latest security information.

Three Phases of IDS includes:

MONITORING: Monitoring the status of the environment involves interaction with an IDS and other monitoring tools as well as following external information sources looking for vulnerabilities that match their particular environment. All of these monitoring tasks are part of routine ID work, time-consuming, but not as cognitively challenging as the subsequent analysis and response phases.

ANALYSIS: The transition from the monitoring phase to the analysis phase begins with a security trigger event. For network monitoring, this event is usually an IDS alert or recognition of an anomalous event occurring in the environment, such as a sudden spike in traffic or user complaints of slow systems. Analysis of alerts involves not only the alert itself, but many sources of data that provide the contextual information necessary to determine whether or not the alert is an actual intrusion and if so, to assess its severity.

RESPONSE: The most common forms of response in ID are intervention, feedback, and reporting. The response to an attack in progress could be as drastic as unplugging a network connection. More common are responses that occur after the fact, such as patching the vulnerability or reinstalling the compromised machine from backup. Feedback is usually directed at the IDS or other elements of the security infrastructure. It includes tweaking or removing IDS signatures that generate an excessive amount of false positives, even if the

---

"DoS and DDoS Attacks: Impact, Analysis and Countermeasures" by Nikhil Tripathi, https://www.researchgate.net/publication/259941506_DoS_and_DDoS_Attacks_Impact_Analysis_and_Countermeasures accessed on 13-06-2020.

[2] Malware, short for malicious software, is any software used to disrupt computer operation, gather sensitive information, or gain access to private computer systems. Malware is defined by its malicious intent, acting against the requirements of the computer user, and does not include software that causes unintentional harm due to some deficiency. Available at "Tools and Techniques for Malware Detection and Analysis" by Sajedul Talukder, https://arxiv.org/pdf/2002.06819.pdf accessed on 13-06-2020.

[3] Systems hardening is a collection of tools, techniques, and best practices to reduce vulnerability in technology applications, systems, infrastructure, firmware, and other areas. The goal of systems hardening is to reduce security risk by eliminating potential attack vectors and condensing the system's attack surface. Available at https://www.beyondtrust.com/resources/glossary/systems-hardening#:~:text=Systems%20hardening%20is%20a%20collection,condensing%20the%20system's%20attack%20surface. Accessed on 13-06-2020.

signature was not guaranteed to always generate a false positive. Responses also include generating incidence reports for legal action and reports for management.

5. TYPES OF IDS

Intrusion detection systems are classified as:

i. Network-based IDS
ii. Host-based IDS
iii. Hybrid based IDS

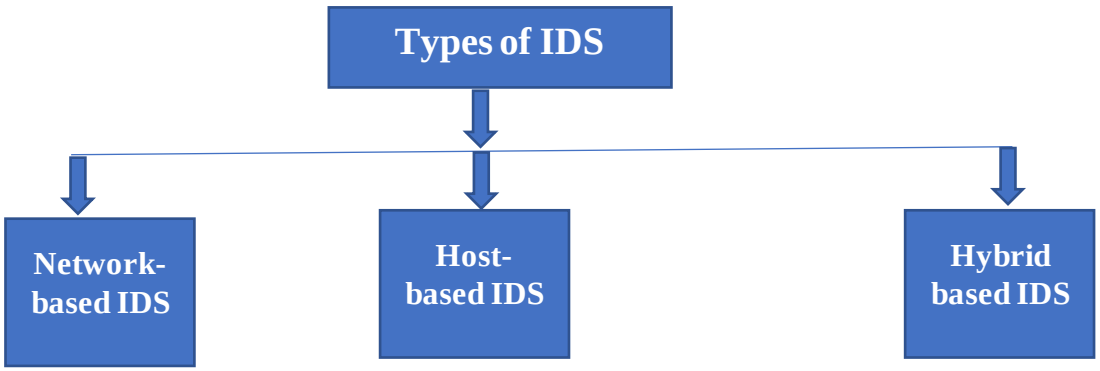

Figure 1: Categories of IDS

5.1 NETWORK-BASED IDS: Network-based IDS are standalone hardware appliances which include network intrusion detection capabilities. They are mostly deployed on strategic point in network infrastructure[4] such as at a boundary between networks, virtual private network servers, remote access servers, and wireless networks[5]. NIDS monitors network traffic going through particular network segments or devices[6]. It can capture and analyse data to detect known attacks or illegal activities or analyse network and application protocol activity to identify anomalous and suspicious activity by traffic

[4] Network infrastructure refers to all of the resources of a network that make network or internet connectivity, management, business operations and communication possible. Network infrastructure typically comprises hardware and software, systems and devices, and it enables computing and communication between users, services, applications and processes. Available at https://blog.gigamon.com/2019/03/06/what-is-network-infrastructure/ accessed on 14-06-2020.

[5] A wireless network allows devices to stay connected to the network but roam untethered to any wires. Access points amplify Wi-Fi signals, so a device can be far from a router but still be connected to the network. When you connect to a Wi-Fi hotspot at a cafe, a hotel, an airport lounge, or another public place, you're connecting to that business's wireless network. Available at https://www.cisco.com/c/en_in/solutions/small-business/resource-center/networking/wireless-network.html accessed on 14-06-2020.

[6] A network device is a node in the wireless mesh network. It can transmit and receive wireless HART data and perform the basic functions necessary to support network formation and maintenance. Network devices include field devices, router devices, gateway devices, and mesh hand-held devices. Available at https://www.sciencedirect.com/topics/engineering/network-device#:~:text=A%20network%20device%20is%20a,and%20mesh%20hand%2Dheld%20devices. Accessed on 14-06-2020.

scanning. NIDS can also be referred as "packet sniffers[7]", because it captures and collect the data in the form of internet packets passing through communication mediums.

5.2 HOST-BASED IDS: In Host-Based IDS, the characteristics of a single host are monitored and the events of that host are observed for any malicious activity. They can monitor network traffic, logs, processes, operations performed by applications, file access and modification, and any configuration change in system. The deployment of HIDS is usually done on critical hosts. Critical host includes servers or systems that are publicly accessible and have some sensitive information. They are placed on one server or workstation, where data is collected from different resources and machine analyse the data locally.

5.3 HYBRID BASED IDS: In Hybrid based IDS, both host-based and network-based IDSs are used simultaneously or we can say that the technique which combines the network intrusion detection system and host intrusion detection system is known as hybrid intrusion detection system. Some of the most current intrusion detection system only uses one of the two detection methods, misused detection or anomaly detection both of them have their own limitations, this is the technique which combines misuse detection system and anomaly detection system is known as hybrid intrusion detection system.

6. INTRUSION DETECTION TECHNIQUES

There are two types of IDS techniques:

i. Anomaly Based Detection Technique
ii. Signature Based Detection Technique

6.1 ANOMALY BASED DETECTION TECHNIQUE: An anomaly-based intrusion detection system, is a technique for detecting both network and computer intrusions and misuse by monitoring system activity and classifying it as either normal or anomalous. The classification is based on some rules, rather than patterns or signatures, and attempts to detect any type of malicious activity that falls out of normal system operation. While the signature- based systems can only detect attacks for which a signature has previously been created.

6.2 SIGNATURE BASED DETECTION TECHNIQUE: Signature-based IDS refers to the detection of attacks by looking for specific patterns, such as byte sequences in network traffic, or known malicious instruction sequences used by malware. The terminology is generated by anti- virus software, which refers to these detected patterns as signatures. Even though signature-based IDS can easily detect known attacks, it is impossible to detect new attacks, for which no pattern is available.

7. LIMITATION OF TRADITIONAL IDS

i. IDS monitor the whole network, so are vulnerable to the same attacks the network's hosts are. Protocol-based attacks can cause the IDS to fail.

7 *Packet Sniffers* are a type of Passive Service. Rather than opening up a TCP port and *actively* listening for requests, the Packet Sniffer *passively* reads raw data packets off the network interface. The Sniffer assembles these packets into complete messages that can then be passed into an associated policy. Available at https://docs.oracle.com/cd/E27515_01/common/tutorials/general_pcap.html accessed on 14-062020.

ii. Network IDS can only detect network anomalies which limit the variety of attacks it can discover.
iii. Host IDS rely on audit logs, any attack modifying audit logs threaten the integrity of HIDS.
iv. Network IDS can create a bottleneck as all the inbound and outbound traffic passes through it.
v. Constant software updates are required for signature-based IDS to keep up with the new threats.
vi. Noise can severely reduce the capabilities of the IDS by generating a high false-alarm rate.
vii. Several real attacks are far less than the number of false alarms raised. This causes real threats to go often unnoticed.

8. EVOLUTION OF MACHINE LEARNING FOR IDS

Machine Learning is the field of study that gives computers the capability to learn and improve from experience without being programmed explicitly automatically. Machine learning focuses on the development of programs that can use data to discover themselves. The process of learning begins with observations or data to look for patterns in data and make better predictions based on the examples provided. The primary aim is to allow the computers to learn without human assistance and adjust actions accordingly.

Machine Learning Algorithms can be broadly classified into:

i. Supervised machine learning algorithms[8]
ii. Unsupervised machine learning algorithms[9]

Unsupervised learning algorithms can "learn" the typical pattern of the network and can report anomalies without any labelled dataset. It can detect new types of intrusions but is very prone to false positive alarms. The supervised model can handle the known attacks deftly and can also recognise variations of those attacks. Supervised learning relies on useful information in labelled data. Classification is the most common task in supervised learning (and is also used most frequently in IDS). However, labelling data manually is expensive and time consuming. Consequently, the lack of sufficient labelled data forms the main bottleneck to supervised learning. In contrast, unsupervised learning extracts

[8] Supervised learning is the machine learning task of inferring a function from labelled training data. The training data consist of a set of training examples. In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal). A supervised learning algorithm analyses the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances. Available at http://www.datascienceassn.org/sites/default/files/Introduction%20to%20Machine%20Learning.pdf accessed on 15-06-2020.

[9] Unsupervised learning4 4 seems much harder: the goal is to have the computer learn how to do something that we don't tell it how to do! There are actually two approaches to unsupervised learning. The first approach is to teach the agent not by giving explicit categorizations, but by using some sort of reward system to indicate success. Note that this type of training will generally fit into the decision problem framework because the goal is not to produce a classification but to make decisions that maximize rewards. Available at https://www.aihorizon.com/essays/generalai/supervised_unsupervised_machine_learning.htm accessed on 15-06-2020.

valuable feature information from unlabelled data, making it much easier to obtain training data. However, the detection performance of unsupervised learning methods is usually inferior to those of supervised learning methods.

9. MACHINE LEARNING ALGORITHM USED FOR IDS

The common machine learning algorithms used in IDSs are shown in Figure 2.

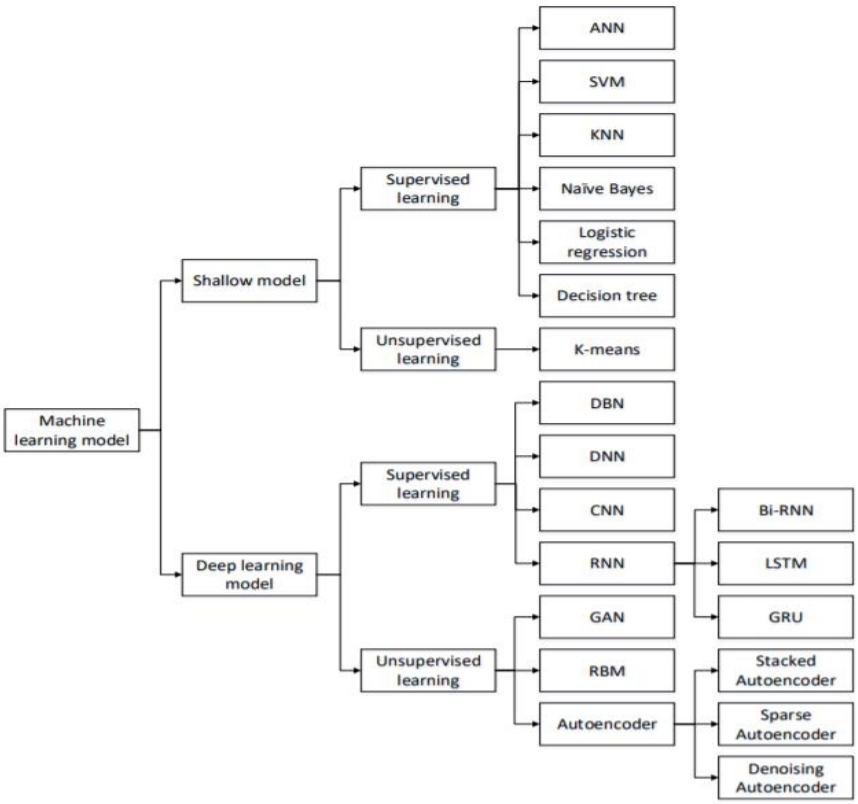

Figure 2: Algorithms used in machine learning

The Machine learning models are divided into two categories:

i. Shallow Model
ii. Deep Learning Model

Both of the models classified into the two types of algorithms used in Machine Learning i.e. supervised and unsupervised learning algorithms. In this paper we are not going to discuss deep learning methods.

9.1 SHALLOW MODELS: The traditional machine learning models (shallow models) for IDS primarily include the artificial neural network (ANN), support vector machine (SVM), K-nearest neighbour (KNN), naïve Bayes, logistic regression (LR), decision tree, clustering, and combined and hybrid methods. Some of these methods have been studied for several

decades, and their methodology is mature. They focus not only on the detection effect but also on practical problems, e.g., detection efficiency[10] and data management[11].

According to figure 2, the shallow model algorithms are as following:

9.1.1 Artificial Neural Network (ANN): The design idea of an ANN is to mimic the way human brains work. An ANN contains an input layer, several hidden layers, and an output layer. An ANN contains a huge number of units and can theoretically approximate arbitrary functions. Hence, it has strong fitting ability, especially for nonlinear functions. Due to the complex model structure, training ANNs is time-consuming.

9.1.2 Support Vector Machine (SVM): The strategy in SVMs is to find a max-margin separation hyperplane[12] in the n-dimension feature space. SVMs can achieve gratifying results even with small-scale training sets because the separation hyperplane is determined only by a small number of support vectors. However, SVMs are sensitive to noise near the hyperplane. SVMs are able to solve linear problems as well.

9.1.3 K-Nearest Neighbour (KNN): The core idea of KNN is based on the manifold hypothesis. If most of a sample's neighbours belong to the same class, the sample has a high probability of belonging to the class. Thus, the classification result is only related to the top-k nearest neighbours. The parameter k greatly influences the performance of KNN models. The smaller k is, the more complex the model is and the higher the risk of overfitting. Conversely, the larger k is, the simpler the model is and the weaker the fitting ability.

9.1.4 Naïve Bayes: The Naïve Bayes algorithm is based on the conditional probability and the hypothesis of attribute independence. For every sample, the Naïve Bayes classifier calculates the conditional probabilities for different classes. The sample is classified into the maximum probability class.

9.1.5 Logistic Regression (LR): The LR is a type of logarithm linear model. The LR algorithm computes the probabilities of different classes through parametric logistic distribution. An LR model is easy to construct, and model training is efficient. However, LR cannot deal well with nonlinear data, which limits its application.

9.1.6 Decision tree: The decision tree algorithm classifies data using a series of rules. The model is tree like, which makes it interpretable. The decision tree algorithm can automatically exclude irrelevant and redundant features. The learning process includes feature selection, tree

[10] Efficiency of machine learning algorithms in identifying the general state. Available at https://www.researchgate.net/publication/328234817_A_Suite_of_Intelligent_Tools_for_Early_Detection_and_Prevention_of_Blackouts_in_Power_Interconnections accessed on 16-06-2020.

[11] Large-scale data analytics using machine learning (ML) underpins many modern data-driven applications. ML systems provide means of specifying and executing these ML workloads in an efficient and scalable manner. Data management is at the heart of many ML systems due to data-driven application characteristics, data-centric workload characteristics, and system architectures inspired by classical data management techniques. Available at https://www.morganclaypool.com/doi/pdf/10.2200/S00895ED1V01Y201901DTM057 accessed on 16-06-2020.

[12] A hyperplane in an n-dimensional Euclidean space is a flat, n-1 dimensional subset of that space that divides the space into two disconnected parts. For example let's assume a line to be our one dimensional Euclidean space(i.e. let's say our datasets lie on a line). Now pick a point on the line, this point divides the line into two parts. The line has 1 dimension, while the point has 0 dimensions. So a point is a hyperplane of the line. Available at https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989 accessed on 16-06-2020.

generation, and tree pruning. When training a decision tree model, the algorithm selects the most suitable features individually and generates child nodes from the root node.

9.1.7 Clustering: Clustering is based on similarity theory, i.e., grouping highly similar data into the same clusters and grouping less-similar data into different clusters. Different from classification, clustering is a type of unsupervised learning. No prior knowledge or labelled data is needed for clustering algorithms; therefore, the data set requirements are relatively low. However, when using clustering algorithms to detect attacks, it is necessary to refer external information.

9.1.8 K-means: K-means is a typical clustering algorithm, where K is the number of clusters and the means is the mean of attributes. The K-means algorithm uses distance as a similarity measure criterion. The shorter the distance between two data objects is, the more likely they are to be placed in the same cluster. The K-means algorithm adapts well to linear data, but its results on nonconvex data are not ideal. In addition, the K-means algorithm is sensitive to the initialization condition and the parameter K. Consequently, many repeated experiments must be run to set the proper parameter value.

10. DATASETS IN IDS:

The task of machine learning is to extract valuable information from data; therefore, the performance of machine learning depends upon the quality of the input data. Understanding data is the basis of machine learning methodology. For IDSs, the adopted data should be easy to acquire and reflect the behaviours of the hosts or networks. The common source data types for IDSs are packets, flow, sessions, and logs. Building a dataset is complex and time-consuming. After a benchmark dataset is constructed, it can be reused repeatedly. The most common datasets used in machine learning based IDSs are DARPA1998[13], KDD99[14], NSL-KDD[15] and UNSW-NB15[16].

[13] The DARPA1998 dataset was built by the Lincoln laboratory of MIT and is a widely used benchmark dataset in IDS studies. To compile it, the researchers collected Internet traffic over nine weeks; the first seven weeks form the training set, and the last two weeks form the test set. The dataset contains both raw packets and labels Available at http://www.ll.mit.edu/r-d/datasets/1998-darpa-intrusion-detection-evaluation-dataset accessed on 17-06-2020.

[14] The KDD99 dataset is the most widespread IDS benchmark dataset at present. Its compilers extracted 41-dimensional features from data in DARPA1998. The labels in KDD99 are the same as the DARPA1998. There are four types of features in KDD99, i.e., basic features, content features, host-based statistical features, and time-based statistical features. Available at http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html accessed on 17-06-2020.

[15] To overcome the shortcomings of the KDD99 dataset, the NSL-KDD was proposed. The records in the NSL-KDD were carefully selected based on the KDD99. Records of different classes are balanced in the NSL-KDD, which avoids the classification bias problem. The NSL-KDD also removed duplicate and redundant records; therefore, it contains only a moderate number of records. Available at https://www.unb.ca/cic/datasets/nsl.html accessed on 17-06-2020.

[16] The UNSW-NB15 dataset was compiled by the University of South Wales, where researchers configured three virtual servers to capture network traffic and extracted 49-dimensional features using tool named Bro. The dataset includes more types of attacks than does the KDD99 dataset, and its features are more plentiful. The data categories include normal data and nine types of attacks. The features include flow features, basic features, content features, time features, additional features, and labelled features.

11. ATTACK DETECTION ON MACHINE LEARNING BASED IDS

Machine learning is a type of data driven method in which understanding the data is the first step. The different types of data reflect different attack behaviours, which include host behaviours and network behaviours. Host behaviours are reflected by system logs, and network behaviours are reflected by network traffic. There are multiple attack types, each of which has a unique pattern. Thus, selecting appropriate data sources is required to detect different attacks according to the attack characteristics.

There are for ways to detect attacks on machine learning based IDS:

1. Packet-Based Attack Detection
2. Flow-Based Attack Detection
3. Session-Based Attack Detection
4. Log-Based Attack Detection

11.1 Packet-Based Attack Detection: Packets, which are the basic units of network communication, represent the details of each communication. Packets consist of binary data, meaning that they are incomprehensible unless they are first parsed. A packet consists of a header and application data. The headers are structured fields that specify IP addresses, ports and other fields specific to various protocols. The application data portion contains the payload from the application layer protocols. There are three advantages to using packets as IDS data sources:

i. Packets contain communication contents; thus, they can effectively be used to detect U2L and R2L attacks.
ii. Packets contain IPs and timestamps; thus, they can locate the attack sources precisely
iii. Packets can be processed instantly without caching; thus, detection can occur in real time.

However, individual packets do not reflect the full communication state nor the contextual information of each packet, so it is difficult to detect some attacks, such as DDOS. The detection methods based on packets mainly include packet parsing methods[17] and payload analysis methods[18].

11.2 Flow-Based Attack Detection: Flow data contains packets grouped in a period, which is the most widespread data source for IDSs. The KDD99 and the NSL-KDD datasets are both flow data. Detecting attacks with flow has two benefits:

[17] The packet parsing-based detection methods primarily focus on the protocol header fields. The usual practice is to extract the header fields using parsing tools and then to treat the values of the most important fields as feature vectors. Packet parsing-based detection methods apply to shallow models.

[18] Payload analysis-based detection places emphasis on the application data. The payload analysis-based methods are suitable for multiple protocols because they do not need to parse the packet headers. As a type of unstructured data, payloads can be processed directly by deep learning models. It should be noted that this method does not include encrypted payloads. Shallow models depend on manual features and private information in packets, leading to high labour costs and privacy leakage problems. Deep learning methods learn features from raw data without manual intervention.

i. Flow represents the whole network environment, which can detect most attacks, especially DOS and Probe.
ii. Without packet parsing or session restructuring, flow pre-processing is simple.

However, flow ignores the content of packets; thus, its detection effect for U2R and R2L is unsatisfactory. When extracting flow features, packets must be cached packets; thus, it involves some hysteresis. Flow-based attack detection mainly includes feature engineering and deep learning methods. In addition, the strong heterogeneity of flow may cause poor detection effects. Traffic grouping is the usual solution to this problem.

11.3 Session-Based Attack Detection: A session is the interaction process between two terminal applications and can represent high-level semantics. A session is usually divided on the basis of a 5-tuple (client IP, client port, server IP, server port, and protocol). There are two advantages of detection using sessions:

i. Sessions are suitable for detecting an attack between specific IP addresses, such as tunnel and Trojan attacks.
ii. Sessions contain detailed communications between the attacker and the victim, which can help localize attack sources.

However, session duration can vary dramatically. As a result, a session analysis sometimes needs to cache many packets, which may increase lag. The session-based detection methods primarily include statistics-based features and sequence-based features.

11.4 Log-Based Attack Detection: Logs are the activity records of operating systems or application programs; they include system calls, alert logs, and access records. Logs have definite semantics. There are three benefits to using logs as a data source in IDSs:

i. Logs include detailed content information suitable for detecting SQL injection, U2R[19], and R2L[20] attacks.
ii. Logs often carry information about users and timestamps that can be used to trace attackers and reveal attack times.
iii. Logs record the complete intrusion process; thus, the result is interpretable.

However, one problem is that log analysis depends on cyber security knowledge. Additionally, the log formats of different application programs do not have identical formats, resulting in low scalability. The log-based attack detection primarily includes hybrid methods involving rules and machine learning, log feature extraction-based methods, and text analysis-based methods.

12. ROLE OF DATA ANALYTICS TECHNIQUES IN MACHINE LEARNING BASED IDS

There are four types of data analytics techniques used in machine learning based IDS as mentioned below. All four techniques play their role in identifying possibility of intrusion before it could cause harm to the organisation. Currently predictive analytics is used in most

[19] user to root attack (u2r) is usually launched for illegally obtaining the root's privileges when legally accessing a local machine. Available at http://www.ijcttjournal.org/archives/ijctt-v43p118 accessed on 18-062020.
[20] Remote to local attack (r2l) has been widely known to be launched by an attacker to gain unauthorized access to a victim machine in the entire network. Available at http://www.ijcttjournal.org/archives/ijctt-v43p118 accessed on 18-06-2020.

of the machine learning based IDS but in future, prescriptive analytics is going to be used. The four techniques of data analytics used in machine learning based IDS are as below:

12.1 Descriptive Analytics: This technique helps in IDS to describe or summarize the existing data of historical attacks. It helps the tools to better understand what has happened and what is going on. In short it explains what happened. This way it gets easy to identify and address the areas of strengths and weaknesses of IDS such that it can help in strategizing how to tackle the future coming attacks. The two main techniques involved are data aggregation and data mining stating that this method is purely used for understanding the underlying behaviour and not to make any estimations.

12.2 1Diagnostic Analytics: This technique focuses on past performance to determine what happened and why. Basically, it explains why the attack has happened. It is characterized by techniques such as drill-down, data discovery, data mining and correlations. Diagnostic analytics takes a deeper look at data to understand the root causes of the events. It is helpful in determining what factors and events contributed to the outcome. It mostly uses probabilities, likelihoods, and the distribution of outcomes for the analysis. Training algorithms for classification and regression also fall in diagnostic analytics.

12.3 Predictive Analytics: This technique emphasizes on predicting the possible outcome using statistical models and machine learning techniques. It helps in forecasting what might happen. It is important to note that it cannot predict if an event will occur in the future; it merely forecasts what are the probabilities of the occurrence of the event. A predictive model builds on the preliminary descriptive analytics stage to derive the possibility of the outcomes. Predictive analytics relies on machine learning algorithms like random forests, SVM, etc. and statistics for learning and testing the data. Usually, companies need trained data scientists and machine learning experts for building these models.

12.4 Prescriptive Analytics: It is a type of predictive analytics that is used to recommend one or more course of action on analysing the data. It recommends an action based on the forecast. it goes beyond the three mentioned above to suggest the future solutions. It can suggest all favourable outcomes according to a specified course of action and also suggest various course of actions to get to a particular outcome. Hence, it uses a strong feedback system that constantly learns and updates the relationship between the action and the outcome.

Predictive analytics forecasts the upcoming events based on the prior occurring and knowledge from the dataset analysis. It is widely used in machine learning for predicting future but for coming time, where intruders with also equip themselves with the techniques to bypass the predictive analysis we will need to use prescriptive analytics. Prescriptive analytics is still at the budding stage and not many firms have completely used its power. It could play the role of predictive analytics and in addition to that it could also tell the actions need to be taken for the forecasted event and automatically can be used to perform the task to avoid the intrusion by a machine learning based IDS.

13. CHALLENGES

Although machine learning methods have made great strides in the field of intrusion detection, the following challenges still exist. Some of the challenges that have been observed during this analysis are mentioned below:

i. **Lack of available datasets:** New types of attacks are emerging, and some existing datasets are too old to reflect these new attacks. The most widespread dataset is currently KDD99, which has many problems, and new datasets are required. Systematic datasets construction and incremental learning may be solutions to this problem.
ii. **Inferior detection accuracy in actual environments:** when the dataset does not cover all typical real-world samples, good performance in actual environments is not guaranteed even if the models achieve high accuracy on test sets.
iii. **Low efficiency:** Most studies emphasize the detection results; therefore, they usually employ complicated models and extensive data pre-processing methods, leading to low efficiency. However, to reduce harm as much as possible, IDSs need to detect attacks in real time.

14. CONCLUSION AND SUGGESSIONS

The traditional IDS have several limitations as they are protecting the organizations from the previously occurred events. Machine learning based IDS protects from the fore coming intrusions using machine learning algorithms. Most of the IDS uses predictive analysis for predicting the future based event. The use of machine learning in Intrusion Detection Systems has built the advancement in the organization. In the coming era of quantum computing where the speed of calculation would be way too fast than the current speed, the prescriptive analysis would be the best technique. As it could provide the actions for the predictive events. So, here we can say Predictive and prescribe analytics techniques are the future of existing security solution Intrusion Detection System.

Suggestions:

The limitations of current machine learning based IDS could be removed by the following suggestions:

i. Combining domain knowledge with machine learning can improve the detection effect, especially when the goal is to recognize specific types of attacks in specific application scenarios.
ii. Improvements in machine learning algorithms are the main means to enhance the detection effect.
iii. Develop practical models as practical IDSs need to have high detection accuracy, high runtime efficiency and interpretability

15. REFERENCES

i. A. Buczak and E. Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection", Published in IEEE Communications Surveys & Tutorials, vol. 18,2016, Available at http://doi.org/10.1109/COMST.2015.2494502 accessed on 12-06-2020.

ii. C. Nila, "MACHINE LEARNING DATASETS FOR CYBER SECURITY APPLICATIONS",Published on march 2020, Available at https://www.researchgate.net/publication/339617181_MACHINE_LEARNING_DATASETS_FOR_CYBER_SECURITY_APPLICATIONS accessed on 17-06-2020.
iii. F. Jemili, M. Zaghdoud, and A. Ben, "A framework for an adaptive intrusion detection system using Bayesian network," Intelligence and Security Informatics, IEEE, 2007. Available at https://www.academia.edu/3343445/A_Framework_for_an_Adaptive_Intrusion_Detection_System_using_Bayesian_Network accessed on 14-06-2020.
iv. R. Das and T. Morris, "Machine Learning and Cyber Security", conference paper, published at 2017 International Conference on Computer, Electrical & Communication Engineering (ICCECE), Available at https://www.researchgate.net/publication/328815330_Machine_Learning_and_Cyber_Security accessed on 15-06-2020.
v. O. Yavanoglu and M. Aydos, "A Review on Cyber Security Datasets for Machine Learning Algorithms" Published at 2017 IEEE International Conference on Big Data (Big Data), Available at https://www.semanticscholar.org/paper/A-review-on-cyber-security-datasets-for-machine-Yavanoglu-Aydos/c93b33619105127483298a8497e0ff29d1438b55 accessed on 15-06-2020.
vi. V. Ford and A, Siraj, "Applications of Machine Learning in Cyber Security" Published at 27th International Conference on Computer Applications in Industry and Engineering Available at https://www.researchgate.net/publication/283083699_Applications_of_Machine_Learning_in_Cyber_Security accessed on 18-06-2020.
vii. IBM Security Report. Available at https://www.ibm.com/security/data-breach accessed on
viii. S. Matzner, L. Pierce and C. Sinclair, "An Application of Machine Learning to Network Intrusion Detection", Published at 15th Annual Computer Security Applications Conference, Available at https://www.acsac.org/1999/papers/fri-b-1030-sinclair.pdf accessed on 12-06-2020.
ix. Y. Abushark, F. Alsolami and I. Sarker, "IntruDTree: A Machine Learning Based Cyber Security Intrusion Detection Model" Available at https://www.semanticscholar.org/paper/IntruDTree%3A-A-Machine-Learning-Based-Cyber-Security-Sarker-Abushark/01d94d49ef6a3f1936d493c3b8ae2728568babca accessed on 16-06-2020.
x. L. Ferretti, M. Colajanni and G. Apruzzese, "On the effectiveness of machine and deep learning for cyber security", Published at 2018 10th International Conference on Cyber Conflict (CyCon), Available at https://www.researchgate.net/publication/326276522_On_the_effectiveness_of_machine_and_deep_learning_for_cyber_security accessed on 13-06-2020.
xi. M. Lehto, "Artificial intelligence in the cyber security environment Artificial intelligence in the cyber security environment", Published at The 14th International Conference on Cyber Warfare and Security ICCWS2019, At Stellenbosch, South Africa, Available at

https://www.researchgate.net/publication/338223306_Artificial_intelligence_in_the_cyber_security_environment_Artificial_intelligence_in_the_cyber_security_environment accessed on 17-06-2020.

xii. J. Patel and K. Panchal, "Effective Intrusion Detection System using Data Mining Technique", Available at http://www.jetir.org/papers/JETIR1506034.pdf accessed on 15-06-2020.

xiii. S. Biswas, "Intrusion Detection Using Machine Learning: A Comparison Study", Published on International Journal of Pure and Applied Mathematics in 2018, Available at https://www.researchgate.net/publication/326572673_Intrusion_Detection_Using_Machine_Learning_A_Comparison_Study accessed on 12-06-2020.

xiv. S. Soni and B. Bhushan, "Use of Machine Learning algorithms for designing efficient cyber security solution", 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT) , Available at https://ieeexplore.ieee.org/document/8993253 accessed on 14-06-2020.

xv. H. Liu and B. Lang, "Machine Learning and Deep Learning Methods for Intrusion Detection Systems: A Survey", Available at https://www.semanticscholar.org/paper/Machine-Learning-and-Deep-Learning-Methods-for-A-Liu-Lang/236dfdeb4511754cf71ba220ac569b11973502cd Accessed on 16-06-2020.