International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2018): 7.426

Intrusion Detection with Machine Learning & Artificial Intelligence (ML&AI) Techniques to Reduce Cyberattacks (Network Traffic) (New Way to Improve Cybersecurity)

S K. Niamathulla

Product Manager – Content Researcher, EC-Council, Hyderabad, India

Abstract: ***Cybersecurity plays an important role in the field of Information Technology (IT). Securing information becomes one of the biggest challenges in the present day. Whenever we think about cybersecurity the first thing that comes to our mind is cybercrime which is increasing immensely day by day. As we know that billions of people affected by breaches for many years, government agencies and businesses are spending more time and money defending against it. In the existing scenario, many cybersecurity systems use DIDS (Distributed Intrusion Detection Sensor/systems) that allows a limited trained analyst (i.e., CSA/CTIA) to monitor several networks at the same time. However, this approach requires data to be transmitted from DIDS on the defended network to Central Analysis Server (CAS). Transmitting all the data captured by DID sensors and send summaries of activities to report back to a security analyst (CSA/CTIA). With only summaries report, cyber-attacks can go undetected because the analyst (CSA/CTIA) did not have enough information to understand the network activity. In this proposed research we mainly focus on to identifying a new way to improve cybersecurity and to reduce cyber-attacks for which we proposed to design a Scalable Distributed Intrusion Detection System (DIDS) is in Artificial Intelligence & Machine Learning (AI&ML) techniques (i.e. Classifiers & Lossless compression) that gives the security analyst (CSA/CTIA) a quicker, easier, more efficient method to identify attacks across multiple network segments by compressing the network traffic, and also to trace back the activities of the attacker. The DIDS is in AI & ML techniques that provide better facilitation of advance network monitoring, incident analysis, and instant attacks data across multiple network segments and as a result, provides a real-time accurate analysis report for early detection of malicious activities and instant attacks. The DIDS system gives the analyst (CSA/CTIA) a complete real-time accurate analysis of activities reports, it allows the analyst much more flexibility in discovering attack patterns. And to capture all the transmitting data by sensors required too much bandwidth, keeping in view of this we propose to increase the bandwidth of network to improve the data rate flow of network traffic. For which it is easy to reduce the cyber-attacks on the network and save a lot of time and money.***
i

Keywords: Internet, Firewall, DIDS, Bandwidth, Network Classifiers, Lossless Compression, Network Traffic, Certified SOC Analyst (CSA), Certified Threat Intelligence Analyst (CTIA).

1. Introduction

The Internet is one of the fastest-growing domain of technical infrastructure development. In today's world, every organization is utilizing information technology for sharing data or information online. Today our entire society, the Planet Earth, is connecting to the Internet. The level of Internet connection is outpacing our ability to properly secure it. And more than ninety percent of a total commercial transaction is done online, such as e-commerce, banking, and business-related highly confidential and valuable information communicated within the network, so this field required high quality of security for transparent and best transactions. Many cyber security systems use Distributed Intrusion Detection Systems (DIDS) which plays vital role in a network security environment that facilitates advanced network monitoring, incident analysis, and instant attacks data for early detection of malicious activity and instant attacks. By utilizing information provided by an IDS it is possible to apply appropriate countermeasures and mitigate an attack, otherwise, it seriously damages the network. However, the current high volume of network traffic over the most IDS techniques require new methods or techniques to handle huge quantities of network traffic during analysis while still maintaining high throughput.

Network traffic analysis and prediction resemble a proactive approach rather than reactive, where the network is monitored to ensure that security breaches do not occur within network. The network traffic classification analysis is a significant stage for developing successful preventive congestion/blockage control schemes and to find out regular and malicious packets.

The predictability of network traffic is of the most important benefits in many cases, such as network security, network planning, and dynamic bandwidth allocation, so on.

The predictability of network traffic classified into two categories:
1) Long-period network predictions
2) Short period network predictions

Long-period network traffic prediction gives detailed forecasting of traffic to evaluate future capacity requirements and therefore permits for more minute planning and safer decisions. **Short period** (milliseconds to minutes) network traffic prediction is linked to dynamic resource allotment. It can be used to improve the Quality of Service (QoS) mechanisms as well as for congestion control and for optimal resource management.

In this research, we proposed to design a scalable Distributed Intrusion Detection System (DIDS) is in AI & ML techniques

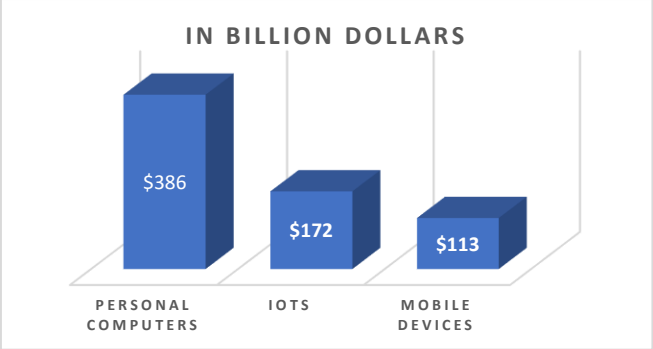
i.e. Network Classification and Lossless Compression techniques, which compressing the size of the file (network traffic) is reduced without losing data/image quality and gives a complete network traffic classification analysis to security analyst (CSA & CTIA) by which a quicker and more efficient method to identify cyberattacks across multiple network segments by compressing the network traffic, and also trace back the activities of the attacker. AI & ML techniques give very promising accuracy results in network traffic classification by compressing the network traffic. And, the sensor to capture all the data transmitting from IDS to CAS, requires too much bandwidth, for which we proposed to increase the bandwidth of the network, as shown in the fig-4, due to which the data rate flow of the network traffic increases, and it is difficult an attacker to attack on network. As a result, it is easy to reduce the cyber-attacks on the network and save a lot of time and money.

The remaining paper is organized as follows, a short description of why cybersecurity, follows by an extensive review of the latest cybersecurity issues in section two. Section three reviews various phases of the research (Improve cybersecurity) and section four reviews about proposed methodology i.e. DIDS is in AI & ML and section five reviews on AI & ML techniques and in the last, section six we give our conclusion.

Why Cybersecurity?

The cybersecurity becomes a major concern over the last few decades as hackers have penetrated the IT infrastructure of government agencies and firms with increasing frequency and complexity. Cybersecurity plays an important role in the current advancement of Information Technology, as well as Internet Services [1]. Improving cybersecurity and protecting critical information infrastructures are essential to each nation's security and economic wellbeing. Making the Internet safer has become integral to the development of new services as well as governmental policy. Preventing cybercrime is an integral component of national cybersecurity and critical information infrastructure protection strategy.

If we see the sizing of the cybersecurity market, the estimate of $655 billion will be spent on cybersecurity initiatives to protect Personal Computers (PCs), Mobile Devices, IoT devices between 2015 – 2020 [2].



https://www.businessinsider.com/

Figure 1: Sizing of Market between 2015-2020

2. Latest Cybersecurity Issues

Cybercrime is the greatest threat to every organization in the world, and one of the biggest problems with mankind. The impact on high society is reflected in the numbers.

According to AT&T Report, that billions of personal user information stolen from large companies. In truth, a data breach can happen to any company big or small and can involve data theft of only a few thousand records [3].

Companies	*Cyber Attacks*		*Percentage*
IT Governance U.K.	Data Breach		1.76 billion (Jan – 2019)
IBM	Criminal Attack	Data Breach	48% (25,575)
		HR Account	27%
		System Glitch	25%
Verizon	Criminal Attack	Data Breaches	48%
		Malware	30%
		Social Attacks	17%
		Misuse	12%
		Physical Action	11%
		Public sector entities	14%
		Accommodation & Food Services	15%
IT Professionals (End user devices)	Data Breach	Desktops & laptops	70%
		Smartphones	61%
		Tablets	53%
		Wireless access points	50%
		Servers rooms	50%
		Routers & Switches	47%

Most Vulnerable Devices to Cybercrimes -The following as having the highest level of risk to hacks, breaches, and other cyber threats

Devices	*Cybercrimes*
Desktop & Laptops	70%
Smartphones	61%
Tablets	53%
Wireless Access Points	50%
Servers & Servers Rooms	50%
Routers & Switches	47%

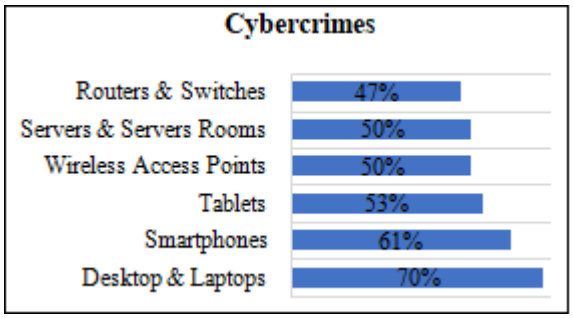


Figure 2: Most Vulnerable Devices to Cybercrimes

Privacy of Information and data-stealing will be the top security issues that organizations need to focus on. In today's world all information and data in digital form. Social media platforms provide a space where users feel safe and secure as they interact with friends and family.

Latest cybersecurity issues: According to a New Data Security Council of India (DSCI), Cyber-attacks are the fastest growing crime in the United States (U.S.), and they are increasing in size, sophistication, and cost. India has been the second most cyber-attacks affected country between 2016 and 2018. The rising cyber-attacks have resulted in more and more companies opting for cyber insurance policies to mitigate the cyber-breach risk. While Information Technology, Banking, and Financial services were early adopters of cyber insurance, a new demand has arisen among Manufacturing, Pharmaceutical, Retail, Hospitality, Research & Development and Internet Protocol (IP)-based organizations [4].

According to a CISCO Annual Cyber Security Report, 53% of all cyber-attacks led to financial damages of more than $500K for organizations in 2018. To minimize such cybersecurity issues, we initiated a new way to improve cybersecurity with latest techniques of AI & ML to provides a real-time accurate analysis report and to reduce the cyber-attacks on the network by compressing the network traffic without losing the originality and quality of the data [4].

3. Phases for Proposed Methodology

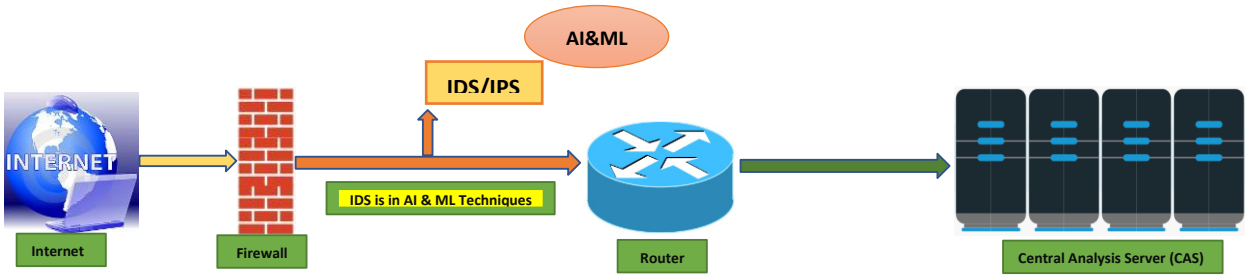


Figure 3: Proposed System - IDS with AI & ML Techniques

Internet
The Internet is widely used all over the world for communication and data transfer. The internet has also changed means of security surveillance, money transactions through online payments, operating devices (IoT), travel and appointment bookings and many more. Whenever people face difficulties with something like travelling to a new city or curiosity on information, they can simply turn on their internet devices such as phones, tablets, and computers to check for a solution. So far, the internet has made human daily life as possible as it can in many ways and has always been with us as the closest friend. With all these great features also come online security and threats where some unknown user can be monitoring the device and activities to retrieve any personal information known for the intrusion. Intrusion is unauthorized activities happening to the internet like identity theft, unauthorized logins, confidential file retrieval and unauthorized transactions without prior permission of the owner. This has been a huge cybersecurity threat today and is happening through various means like viruses, malware, phishing, spamming and many more. This paper teaches us about is to minimize such threats and protect from retrieval and unauthorized transactions.

Firewalls
A firewall program will monitor traffic both in and out of a computer and alert the user to apparently unauthorized usage.

DIDS – Distributed Intrusion Detection System
A Distributed Intrusion Detection System (DIDS) – A Multiple Intrusion Detection Systems (IDS) over a large network, all of which communicate with each other, or with a Central Analysis Server (CAS) that facilitates advanced network monitoring, incident analysis, and instant attacks data.

By having these co-operative agents distributed across a network, SOC Analyst, Incident Analysts, Network Engineers, and Security Personnel can get a wider view of what is occurring on their network as a whole. A DIDS also allows a company to efficiently manage its incident analysis resources by centralizing its attack records and by giving the analyst a quick and easy way to point new trends and patterns, to identify threats to the network across multiple network segments.

In face of the growing volume of network traffic and high transmission rates, software-based network IDSs present performance issues, not being able to analysis all the captured packets rapidly enough. Some hardware-based network IDSs offer the necessary analysis throughput, but the cost of such systems is too high about software-based alternatives. The present IDS technology is increasingly unable to protect the global information infrastructure due to several problems. The existence of single intruder attacks that cannot be detected based on the observations of only a single site. Normal variations in system performance and changes in attack behavior that cause false detection and identification. To overcome such problems, we mainly focus on to improve cybersecurity, and to reduce the cyberattack for which we proposed to design a Scalable Distributed Intrusion Detection System (DIDS) is in Artificial Intelligence & Machine Learning (AI & ML) techniques such as Network Classifier and Lossless Compression algorithms that gives the complete real-time analysis report on advance monitoring, incidence attacks, instant attack by compressing the network traffic to security analyst (CSA/CTIA), so that a quicker, more efficient method to identify attacks across multiple network segments and also trace back the activities of the attackers and as a result, provides a real-time accurate analysis report for early

detection of malicious activities and instant attacks. Detection of attack intention (IDS) and trending is needed for prevention (IPS). The sheer volume of attack notifications received by ISPs and host owners can become overwhelming.

Central Analysis Server (CAS)

The Central Analysis Server (CAS) is really the heart and soul of the operation. This server would ideally consist of a database and a web server. This allows the interactive querying of attack data for analysis as well as a useful web interface to allow the corporate guys upstairs to see the current attack status of your network. It also allows analysts to perform pre-programmed queried, such as attack aggregation, statistics gathering, to identify attack patterns and to perform rudimentary incident analysis, all from a web interface.

The co-operative agent network is one of the most important components of the DIDS. An agent is a piece of software that reports attack information to the central analysis server. The use of multiple co-operative agents across a network allows the incident analysis team (CTIA/CSA/ECSA) a broader view of the network than can be achieved with single IDS systems.

4. Proposed Methodology – DIDS is in Artificial Intelligence & Machine Learning Techniques (AI &ML)

To design a scalable distributed Intrusion Detection System (IDS) is in **AI & ML environment with Network classifier and Lossless compression techniques that compressing the size of the file (network traffic) is reduced without losing data/image quality and gives a complete network traffic classification analysis to security analyst (CSA & CTIA) by which a quicker, easier, more efficient method to identify attacks across multiple network segments, and to trace back the activities of the attacker**. The DIDS system gives the analyst a quicker, easier, more efficient method to identify coordinated attacks across multiple network segments and to trace back the activities of the attacker. By having all the attack records stored in a single place, it allows the analyst much more flexibility in discovering attack patterns, and other attack issues which may have otherwise gone unnoticed. The broad view is given by the DIDS system also allows the analyst to ensure a minimum of false positives and false negatives by being able to see beyond a single network segment, into the network. Scalable storage in distributed file system infrastructure. Scalable distributed data processing in AI & ML through classifiers. This strategy should be effective in reducing the amount of traffic sent from the sensor to Central Analysis System (CAS) and Ultimately these techniques could be used to increase the reliability and security of networks.

According to the below diagrammatic representation of network traffic, the broad view is given that transmitting all the data captured by sensors required too much bandwidth. And keeping in view of that we propose to increase the bandwidth of the network to improve the data rate flow of network traffic and to capture all the transmitted data flow, due to which it is easy to reduce the cyber-attacks on the network and save a lot of time and money. (**Shown fig-4**)

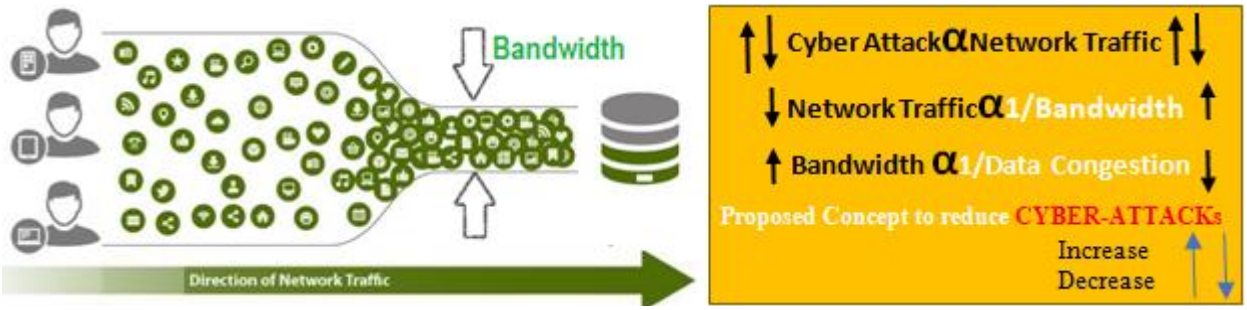


Figure 4: Direction of Network Traffic

Desired Characteristic of the DIDS with AI & ML Techniques (Network Classification & Lossless Compression)

An IDS is in AL & ML techniques, should have certain characteristic in order to be able to provide efficient security against serious cyberattacks. Those characteristics include the following:

- Real time instruction detection - while the attack is in progress or immediately afterwards
- Compressing the data without sacrificing data or quality by using compression technique
- Increases the data rate flow of the network traffic
- Provide complete network traffic classification analysis report (incident analysis, instant attack, monitoring) to security analyst
- It identifies coordinated attacks across multiple network segments
- It allows the analyst much more flexibility in discovering attack patterns, and other attack issues which may have otherwise gone unnoticed.
- False positive alarms must be minimized
- Human supervision should be reduced to minimum, and continuous operation should be ensured
- Recoverability from system crashed, either accidental or those resulting from attacks.
- Self- monitoring ability in order to detect attackers attempts to change the system
- Compliance to the security policies of the system that is being monitored

5. Artificial Intelligence & Machine Learning Techniques

Machine Learning techniques applied to many problems in

computer networking, such as data traffic prediction, data routing and data classification, data congestion control, resource and fault management, data compression and network security.
In this sub-section, we present few techniques of AI & ML algorithms that are being proposed for network traffic analysis in order to reduce cyberattacks.

1) Network Classification Technique
2) Lossless Compression Technique

(A)Network Classification Techniques
Classification is one of the foremost forms of technique of data analysis which takes each instance of a data packet and assigns it to class. A classification techniques-based network traffic analysis attempts to classify all traffics as either normal or malicious. Real-time analysis has become important to detect any suspicious activities. Network classification is the first step of network traffic analysis, and it is the core element of network IDS/IPS. Although the many techniques of network classification have improved and their accuracy has been enhanced,

In this paper, discusses on how we apply DIDS with AI & ML algorithms in several data classification & compression techniques, utilising the statistical properties of the network traffic flow from DIDS to CAS. The proposed different classification techniques (supervised, semi-supervise, and unsupervised) that use Machine Learning algorithms to cope with real-world network traffic.

The challenge of classification is to reduce the number of false positives (detection of normal network traffic as abnormal) and false negative (detection of malicious network traffic as normal).

In the last two decades, numerous network traffic classification techniques have been proposed to classify unknown classes. The first one is Port Based Technique. It is a good technique for network traffic classification. This technique includes a port, which is first registered in Internet Assign Number Authority (IANA). But this technique failed due to increase of Peer to Peer applications (P2P) in which use dynamic port numbers. Dynamic port number means unregistered number with Internet Assign Number Authority (IANA). Then second one is Payload Based Technique. This technique gives accurate results in network traffic classification. This technique is also called Deep Packet Inspection (DPI) technique. But there is a problem in this technique. The problem is that it cannot be used for encrypted data network applications as numerous network applications use encrypted techniques to protect data from detection. So, this technique also failed due to use of encrypted flow of application. In this research we proposed a new trend classification and compression techniques of AI & ML technology to classify internet traffic as well as to know what type of applications flow in the network. Machine Learning (ML) techniques give very promising accuracy results in network traffic classification.

Network Traffic Analysis
Network Traffic Analysis plays a vital role in the present days for monitoring network traffic. In the past years, the administrator was monitoring only a small number of network devices (i.e. less than a thousand networks). The bandwidth of the network was may be just less or 100 Mbps (Megabits per second). Currently, administrators must deal with a higher speed wired network (more than 1Gbps (Gigabit per second) and various networks such as wireless networks and ATM (Asynchronous Transfer Mode). They require advanced network traffic analysis techniques of AI and ML in order to manage network, solve the network problems quickly to avoid network failure, overcome the instant cyberattacks, and handle the network security.

Network traffic analysis presents several challenges in recent days. The Network is analyzed at different levels viz. at the data packet level, data flow level and network level for security management. Various techniques are being used for network traffic analysis, and but with new trend classifiers and lossless techniques of AI and ML technology which provide real-time and complete accurate analysis reports to a security analyst for early detection of malicious activities and instant attacks in the network systems. In Figure – 6 shows five main phases of network traffic analysis. The detailed description of these five phases is presented in subsequent subsections.

Machine Learning (ML) Techniques
Machine Learning technique is based on the data set (Labeled/Unlabelled data set). In this technique, a machine learning classifier is trained as input and then using the trained sample prediction, unknown classes are classified. As shown in figure – 5, the detailed description of each technique.

There are three main areas in Machine Learning technique:
1) Supervised Learning Technique
2) Unsupervised Learning Technique
3) Semi supervised Learning Technique

Figure 5: Machine Learning technique

Supervised Classification Technique
The Supervised learning technique is a machine learning technique. This technique needs a complete labelled data set to classify unknown classes. The supervised learning technique trains the model with some labelled data set and then it will produce prediction output in new data samples.

Unsupervised Classification Technique
The Unsupervised technique is also called a clustering technique. In this classification, there is no need for complete labelled data sets. Unsupervised is a type of machine learning. Thus, the result of machine learning training does not classify instances in predefined classes.

Semi-supervised Learning Technique
Semi-supervised learning is one of the machine learning techniques. This technique also makes use of unlabelled data for training – typically a small amount of labelled data with a large amount of unlabelled data. Semi-supervised learning falls between unsupervised and supervised learning. The Network Traffic Classification is a step by step process method will show you how to use network traffic classification technique to classify network traffic by using the Machine Leaning technique. (Fig-6)

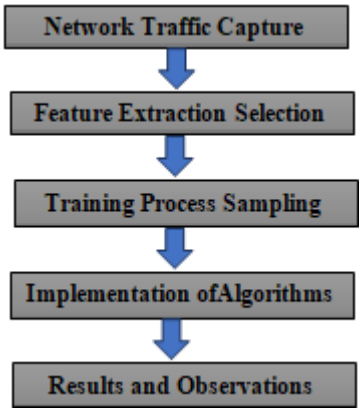


Figure 6: Network Traffic Classification

Network Traffic Capture
This is the foremost important step, which includes data collection. In this step, the real-time network traffic is captured. It is also known as a data collection step. There are many tools for network traffic capturing, but the Tcpdump command can be used to capture real-time network traffic. To capture network traffic, we use the Wire Shark tool for packet capturing and analyzing. We captured the traffic the duration of one minute of WWW, DNS, FTP, P2P, and Telnet application.

Wire Shark Tool: Wireshark is also called as Network Packet Analyzer. Wireshark is open source and is available for free. It is one of the best packets analyzers available today. Its captures packet data in as much details as possible [5].

Feature Extraction Selection
After capture network traffic data, the feature selection and extractions step follow. In this step, the features are extracted from the captured data such as packet duration, packet length, interarrival packet time protocol etc. then extracted features are used to train the Machine Learning (ML) classifier. For feature extraction, the Perl script can be used to extract the feature from the captured data set.

Perl Script –Perl Script is a general-purpose programming language, developed for data manipulation, network programming, web development and many more.

Training Process Sampling
In this training process sampling, the data sets sampled for supervised learning techniques. In supervised learning, data are first labelled to classify unknown network applications.

Implementation of AI &ML Algorithms
This is the implementation step which includes applying ML classifier on the instances. For example, applying Supervise, Unsupervised and Semi-Supervised learning algorithm. For the implementation of ML algorithms, there are many tools available on the internet, but most common nowadays are used MATLAB and Weka classification simulation tools. In this research, we use **Weka tool** and apply four ML classifier algorithms C4.5, Support Vector Machine, BayesNet and NaiveBayes to build a classification model.

Weka Tool: The Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.

Result and Observation
After the implementation of Machine Learning (ML) algorithms, the classification tools provide detailed accurate result about the applied algorithms such as accuracy detailed information, training time and recall etc. In this research, we proposed four classifiers C4.5, SVM, BayesNet and NaïveBayes. But the C4.5 algorithm gives very high accuracy results as compared to other algorithms.

(B) Lossless compression Techniques:
Lossless compression is a technique of data compression in which the size of the file is reduced without sacrificing data/image quality. Unlike lossy compression, no data is lost when this technique is used. Because the data is well-preserved, this technique will decompress the data and restore exactly to its original state. By applying the lossless compression technique, our intention is to identify how to compress network traffic as much as possible without losing the ability to detect and investigate malicious activity by increasing the data rate flow of network traffic. As data rate flow of the network traffic increases, the congestion control of traffic decreases and finally, it is easy to find out regular and malicious packets.

Figure 7: Process of Lossless Compression Techniques

The Lossless compression technique basically rewrites the data of the original file in a more efficient way. However, no quality is lost, the resulting files are typically much larger than image and audio files compressed with lossy compression. This technique should be effective in reducing the amount of traffic sent from the sensor to Central Analysis System (CAS) and Ultimately, this strategy could be used to increase the reliability and security of networks. The lossless compression techniques to reduce the volume of traffic that needs to be transmitted to the Central Analysis Systems (CAS) to less than 10% of the original network traffic while losing no more than 1% of cyber security alerts. This technique compresses the network traffic, which improves the data rate flow of network traffic, due to which it is difficult for an attacker to attack targeted infrastructure.

6. Conclusion

In the existing intrusion detection systems do not properly handle the complete amount of traffic and data transmitted in large scale networks. In this paper, we proposed an efficient

and scalable distributed intrusion detection system is in AI & ML with Network Classification and Lossless Compressors algorithms, which is capable of handling large volumes of data and faultlessly scale to handle network growth as well as efficiently detecting cyber-attacks which occurs in computer networks.

References

[1] Cyber Security Strategy of United Kingdom, 2009
[2] https://www.businessinsider.com/the-cybersecurity-report-threats-and-opportunities-2016-4-22?IR=T
[3] https://financesonline.com/cybercrime-statistics/#stats
[4] https://inc42.com/buzz/cyber-attacks-india/
[5] https://www.wireshark.org/docs/wsug_html_chunked/ChapterIntroduction.html