PREDICTION OF CYBER-ATTACKS USING DATA SCIENCE TECHNIQUE

Sparjan S[1], Deepan Raj M[2], Suriya prakash T[3], Senthil K[4], Preetha M[5]

[1,2,3,]Student, Information Technology, Prince Shri Venkateshwara Padmavathy Engineering College, Chennai, India

[4]Associate Professor, Information Technology, Prince Shri Venkateshwara Padmavathy Engineering College, Chennai, India

[5]Professor, Computer Science & Engineering, Prince Shri Venkateshwara Padmavathy Engineering College, Chennai, India

ABSTRACT

A cyber-attack is an assault launched by cybercriminals using one or further computers against a single or multiple computers or networks for destroying the integrity of the data or stealing the information. The cyber-attack can peril disable computers, use a traduced computer as a position point for some attacks or steals public data. The Hacker use different styles to start a cyber-attack, for illustration phishing, Dos, R2L, probe, malware, U2R among other styles. Though a plethora of extant approaches, models and algorithms have handed the base for cyber-attack prognostications, there's the need to consider new models and algorithms, which are grounded on data representations other than task-specific ways. Still, its non-linear information processing architecture can be shaped towards learning the different data representations of network traffic to classify type of network attack. Networking sectors have to predict the type of Network attack from given dataset using machine learning ways. The analysis of dataset by supervised machine learning techniques(SMLT) to capture several information's like, variable identification, variant analysis, bi-variant and multi-variant analysis, missing value treatments etc. A relative study between machine literacy algorithms had been carried out in order to determine which algorithm is the most accurate in predicting the type cyber Attacks. We classify four types of attacks are DOS Attack, Remote to user (R2L), User to Root (U2R) Attack, probe attack. The results show that the effectiveness of the proposed machine literacy algorithm fashion can be compared with accuracy with entropy calculation, performance, Recall, F1 Score, perceptivity, Particularity and Entropy.

Keyword : - *Denial-of-service attack, Remote To User attack, User to root attack, Probing, Machine learning.*

INTRODUCTION

Data science is the domain of study that integrates domain expertise, programming skills, and knowledge of mathematics and statistics to separate meaningful insights from data. Data science practitioners apply machine learning algorithms to numbers, text, images, video, audio, and more to produce artificial intelligence (AI) systems to perform tasks that ordinarily require human intelligence. In turn, these systems generate insights which analysts and business users can translate into tangible business value. Data science can be defined as a blend of mathematics, business acumen, tools, algorithms and machine learning techniques, all of which help us in finding out the hidden insights or patterns from raw data which can be of major use in the formation of big business decisions.

LITERATURE SURVEY

A. A PREDICTION MODEL OF DOS ATTACK'S DISTRIBUTION DISCRETE PROBABILITY

The process of prediction analysis is a process of using some method or technology to explore or stimulate some unknown, undiscovered or complicated intermediate processes based on previous and present states and then speculated the results. In an early warning system, accurate prediction of DoS attacks is the prime aim in the network offence and defense task. Detection based on abnormity is effective to detect DoS attacks.

B. APRIORI VITERBI MODEL FOR PRIOR DETECTION OF SOCIO-TECHNICAL ATTACKS

Socio-technical attack is an organized approach which is defined by the interaction among people through maltreatment of technology with some of the malicious intent to attack the social structure based on trust and faith. Awful advertisement over internet and mobile phones may defame a person, organization, group and brand value in society which may be proved to be fatal.

C. NEW ATTACK SCENARIO PREDICTION METHODOLOGY

Intrusion detection systems (IDS) are used to detect the occurrence of malicious activities against IT system. Through monitoring and analyzing of IT system activities the malicious activities will be detected. In ideal case IDS generate alert(s) for each detected malicious activity and store it in IDS database. Some of stored alerts in IDS database are related. Alerts relations are differentiated from duplication relation to same attack scenario relation.

D. CYBER ATTACKS PREDICTION MODEL BASED ON BAYESIAN NETWORK

The prediction results reflect the security situation of the target network in the future, and security administrators can take corresponding measures to enhance network security according to the results. To quantitatively predict the possible attack of the network in the future, attack probability plays a significant role. It can be used to indicate the possibility of invasion by intruders. As an important kind of network security quantitative evaluation measure, attack probability and its computing methods has been studied for a long time.

EXISTING SYSTEM

They proposed first to produce a contrastive self-supervised learning to the anomaly discovery problem of attributed networks. ColaSoft, is sustainably consists of three factors: contrastive instance pair sampling, GNN-grounded contrastive learning model, and multi round slice-grounded anomaly score calculation. Their model captures the relationship between each node and its neighboring structure and uses an anomaly-related idea to train the contrastive learning model. We believe that the proposed module opens a new way to expand self-supervised knowledge and contrastive learning to increasingly graph anomaly discovery applications. The multiround anticipates scores by the contrastive learning model are further used to estimate the abnormality of each node with statistical estimation. The training phase and the conclusion phase. In the training phase, the contrastive learning model is trained with tried case pairs in an unsupervised fashion. After that the anomaly score for each node is attained in the conclusion phase.

DISADVANTAGES

1. The performance is not good and its get complicated for other networks.

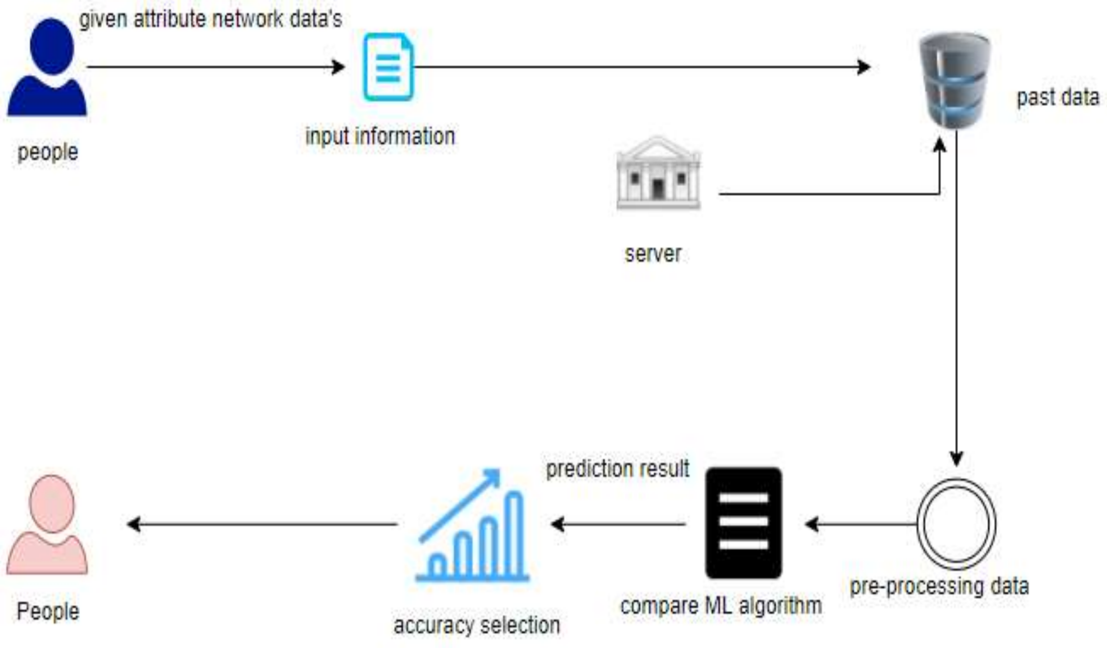
2. The performance criteria like recall F1 score and comparison of machine learning algorithm isn’t done.

PROPOSED SYSTEM

The proposed model is to build a machine learning model for anomaly detection. Anomaly detection is an important technique for recognizing fraud activities, suspicious activities, network intrusion, and other abnormal events that may have great significance but are difficult to detect. The machine learning model is built by applying proper data science techniques like variable identification that is the dependent and independent variables. Then the visualization of the data is done to insights of the data .The model is build based on the previous dataset where the algorithm learn data and get trained different algorithms are used for better comparisons. The performance metrics are calculated and compared.

SYSTEM ARCHITECHTURE

The data set in KDD Cup99 have normal and 22 attack type data with 41 features and all generated traffic patterns end with a label either as 'normal' or any type of 'attack' for upcoming analysis. There are varieties of attacks which are entering into the network over a period of time.



Module Description

1. Data validation process by each attack.
2. Performance measurements of DoS attack.
3. Performance measurements of R2L attacks.
4. Performance measurements of U2R attacks.
5. Performance measurements of Probe attacks.
6. Performance measurements of overall network attacks.
7. GUI based prediction results of Network attacks.

MODULE-01

Variable Identification Process / data validation process

Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, which can be considered as close to the true error rate of the dataset. If the data volume is large enough to be representative of the population, you may not need the validation techniques. However, in real-world scenarios, to work with samples of data that may not be a true representative of the population of given dataset. To finding the missing value, duplicate value and description of data type whether it is float variable or integer. The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyper parameters. The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration. The validation set is used to evaluate a given model, but this is for frequent evaluation. It as machine learning engineers uses this data to fine-tune the model hyper parameters. Data collection, data analysis, and the process of addressing data content, quality, and structure can add up to a time-consuming to-do list. During the process of data identification, it helps to understand your data and its properties; this knowledge will help you choose which algorithm to use to build your model. For example, time series data can be analysed by regression algorithms; classification algorithms can be used to analyse discrete data.

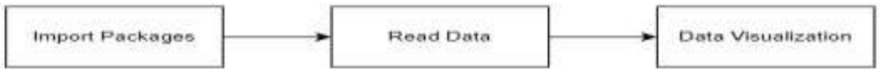
Data Validation/ Cleaning/Preparing Process

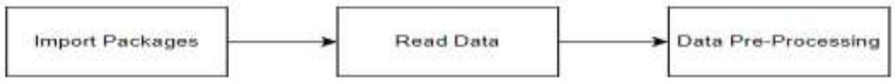
Importing the library packages with loading given dataset. To analysing the variable identification by data shape, data type and evaluating the missing values, duplicate values. A validation dataset is a sample of data held back from training your model that is used to give an estimate of model skill while tuning model's and procedures that you can use to make the best use of validation and test datasets when evaluating your models. Data cleaning / preparing by rename the given dataset and drop the column etc. to analyse the uni-variate, bi-variate and multi-variate process. The steps and techniques for data cleaning will vary from dataset to dataset. The primary goal of data cleaning is to detect and remove errors and anomalies to increase the value of data in analytics and decision making.

Data Pre-processing

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data Preprocessing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis. To achieving better results from the applied model in Machine Learning method of the data has to be in a proper manner. Some specified Machine Learning model needs information in a specified format; for example, Random Forest algorithm does not support null values. Therefore, to execute random forest algorithm null values have to be managed from the original raw data set. And another aspect is that data set should be formatted in such a way that more than one Machine Learning and Deep Learning algorithms are executed in given dataset.

MODULE DIAGRAM





MODULE-02

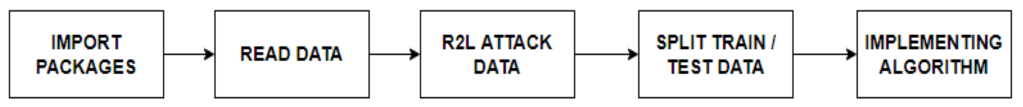
In computing , denial-of-service attack (DoS attack) is a cyber-attack in which the perpetrator seeks to make a machine or network resource unavailable to its intended users by temporarily or indefinitely disrupting services of a host connected to the Internet. Denial of service is typically accomplished by flooding the targeted machine or resource with superfluous requests in an attempt to overload systems and prevent some or all legitimate requests from being fulfilled. In a distributed denial-of-service attack (DDoS-attack)the incoming traffic flooding the victim originates from many different sources. This effectively makes it impossible to stop the attack simply by blocking a single source. A DoS or DDoS attack is analogous to a group of people crowding the entry door of a shop, making it hard for legitimate customers to enter, disrupting trade.A distributed denial-of-service (DDoS) is a large-scale DoS attack where the perpetrator uses more than one unique IP address, often thousands of them. A distributed denial of service attack typically involves more than around 3–5 nodes on different networks; fewer nodes may qualify as a

DoS attack but is not a DDoS attack. Since the incoming traffic flooding the victim originates from different sources, it may be impossible to stop the attack simply by using ingress filtering. It also makes it difficult to distinguish legitimate user traffic from attack traffic when spread across multiple points of origin. As an alternative or augmentation of a DDoS, attacks may involve forging of IP sender addresses (IP address spoofing) further complicating identifying and defeating the attack. An application layer DDoS attack is a form of DDoS attack where attackers target application-layer processes. The attack over-exercises specific functions or features of a website with the intention to disable those functions or features. This application-layer attack is different from an entire network attack, and is often used against financial institutions to distract IT and security personnel from security breaches.

MODULE-03

Now-a-days, it is very important to maintain a high level security to ensure safe and trusted communication of information between various organizations. But secured data communication over internet and any other network is always under threat of intrusions and misuses. To control these threats, recognition of attacks is critical matter. Probing, Denial of Service (DoS), Remote To User (R2L) attacks is some of the attacks which affect large number of computers in the world daily. Detection of these attacks and prevention of computers from it is a major research topic for researchers throughout the world.

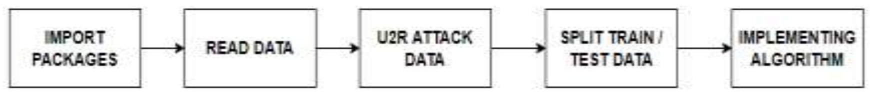
MODULE DIAGRAM



MODULE-04

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. Similarly user to root attack (u2r) is usually launched for illegally obtaining the root's privileges when legally accessing a local machine. Buffer overflow is the most common of U2R attacks. This class begins by gaining access to a normal user while sniffing around for passwords to gain access as a root user to a computer resource. Detection of these attacks and prevention of computers from it is a major research topic for researchers throughout the world

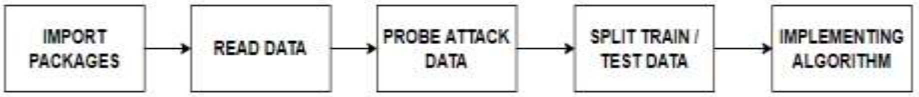
MODULE DIAGRAM



MODULE-05

Probing attacks are an invasive method for bypassing security measures by observing the physical silicon implementation of a chip. As an invasive attack, one directly accesses the internal wires and connections of a targeted device and extracts sensitive information. A probe is an attack which is deliberately crafted so that its target detects and reports it with a recognizable "fingerprint" in the report. The attacker then uses the collaborative infrastructure to learn the detector's location and defensive capabilities from this report. This is an attack where the attacker attempts to gather information about the target machine or the network, to map out the network. Information about target may reveal useful information such as open ports, its IP address, hostname, and operating system. Network Probe is the ultimate network monitor and protocol analyzer to monitor network traffic in real-time, and will help you find the sources of any network slow-downs in a matter of seconds.

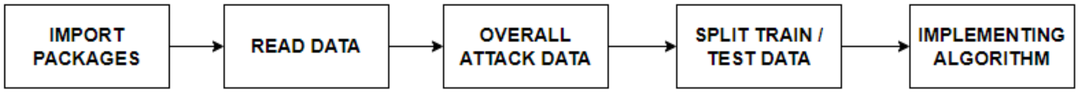
MODULE DIAGRAM



MODULE-06

Increasingly, attacks are executed in multiple steps, making them harder to detect. Such complex attacks require that defenders recognize the separate stages of an attack, possibly carried out over a longer period, as belonging to the same attack. Complex attacks can be divided into exploration and exploitation phases. Exploration involves identifying vulnerabilities and scanning and testing a system. It is how an attacker gathers information about the system. Exploitation involves gaining and maintaining access. At this stage, the attacker applies the know-how gathered during the exploration stage. An example of a complex attack that combines exploration and exploitation is a sequence of a phishing attack, followed by an exfiltration attack. First, attackers will attempt to collect information on the organization they intend to attack, e.g., names of key employees. Then, they will craft a targeted phishing attack. The phishing attack allows the attackers to gain access to the user's system and install malware. The purpose of the malware could be to extract files from the user's machine or to use the user's machine as an attack vector to attack other machines in the organization's network. A phishing attack is usually carried out by sending an email purporting to come from a trusted source and tricking its receiver to click on a URL that results in installing malware on the user's system. This malware then creates a backdoor into the user's system for staging a more complex attack. Phishing attacks can be recognized both by the types of keywords used in the email (as with a spam email), as well as by the characteristics of URLs included in the message. Features that have been used successfully to detect phishing attacks include URLs that include IP addresses, the age of a linked-to domain, and a mismatch between anchor and text of a link.

MODULE DIAGRAM



MODULE-07

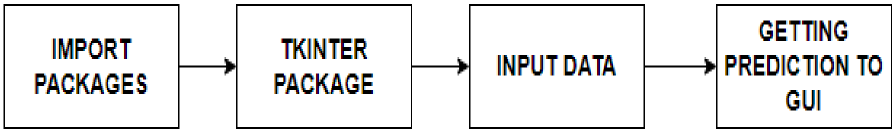
GUI means Graphical User Interface. It is the common user Interface that includes Graphical representation like buttons and icons, and communication can be performed by interacting with these icons rather than the usual text-based or command-based communication. A common example of a GUI is Microsoft operating systems.

The graphical user interface (GUI) is a form of user interface that allows users to interact with electronic devices through graphical icons and audio indicator such as primary notation, instead of text-based user interfaces, typed command labels or text navigation. GUIs were introduced in reaction to the perceived steep learning curve of command-line interfaces (CLIs) which require commands to be typed on a computer keyboard.

The actions in a GUI are usually performed through direct manipulation of the graphical elements. Beyond computers, GUIs are used in many handheld mobile devices such as MP3 players, portable media players, gaming devices, smartphones and smaller household, office and industrial controls. The term GUI tends not to be applied to other lower-display resolution types of interfaces, such as video games (where head-up display (HUD) is preferred), or not including flat screens, like volumetric displays because the term is restricted to the scope of two-dimensional display screens able to describe generic information, in the tradition of the computer science research at the Xerox Palo Alto Research Center.

Graphical user interface (GUI) wrappers find a way around the command-line interface versions of (typically) Linux and Unix-like software pplications and their text-based user interfaces or typed command labels. While command-line or text-based applications allow users to run a program non-interactively, GUI wrappers atop them avoid the steep learning curve of the command-line, which requires commands to be typed on the keyboard. By starting a GUI wrapper, users can intuitively interact with, start, stop, and change its working parameters, through graphical icons and visual indicators of a desktop environment, for example. Applications may also provide both interfaces, and when they do the GUI is usually a WIMP wrapper around the command-line version. This is especially common with applications designed for Unix-like operating systems. The latter used to be implemented first because it allowed the developers to focus exclusively on their product's functionality without bothering about interface details such as designing icons and placing buttons. Designing programs this way also allows users to run the program in a shell script.

MODULE DIAGRAM



CONCLUSION

The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. The best accuracy on public test set is higher accuracy score will be find out by comparing each algorithm with type of all network attacks for future prediction results by finding best connections. This brings some of the following insights about diagnose the network attack of each new connection. To presented a prediction model with the aid of artificial intelligence to improve over human accuracy and provide with the scope of early detection. It can be inferred from this model that, area analysis and use of machine learning technique is useful in developing prediction models that can helps to network sectors reduce the long process of diagnosis and eradicate any human error

ACKNOWLEDGMENT

We Would like to thank all our faculties, friends and family members who have directly and indirectly helped in easy completion of our project. Special thanks to Dr. K.Senthil , Professor, Department of Information Technology, Prince Shri Venkateshwara Padmavathy Engineering college, Chennai, Tamil Nadu, India for guiding us in completing this project successful.

REFERENCES

[1] A. B. Kulkarni and S. F. Bush. Detecting distributed denial-of-service attacks using kolmogorov complexity metrics. J.Network Syst. Manage., 14(1):69–80, 2006.

[2] B. S. Maulik U. Genetic algorithm-based clustering tech-nique. Pattern Recognition, 33(9):1455–1465, 2000.

[3] F. D. B. Evolution computation: toward a new philosophy of machine intelligence. Piscataway: IEEE Press, 1995

[4] G. Jian. Cluster analysis based on c-means and immune genetic algorithm. Computer Engineering, 29(12):65–66, 2003.

[5] G. R.-K. T. K.-R. Muller, S. Mika and B. Scholkopf. An in-troduction to kernel-based learning algorithms. IEEE Trans-action on Neural Networks, 12(2):181–201, May 2001

[6] J. Raychaudhuri.S., Stuart and R. Altman. Principal compo-nents analysis to summarize microarray experiments: Ap-plication to sporulation time series. In In Proc. Pacific Sym-posium on Biocomputing.

[7] Katharine E Worton, “Using Socio-Technical and Resilienceframeworks to Anticipate Threat”, Workshop on Socio-Technical Aspects in Security and Trust, Socio-Technical Center, University
of Leeds, UK 2012.

[8] L. C.-J. Lin Kuan-Ming. A study on reduced support vec-tor machines. IEEE Transactions on Neural Networks, 14(6):1449–1459, 2003.

[9] Robin Gandhi, Anup Sharma, William Mahoney, William Sousan,Qiuming Zhu and Phillip Laplante, “Dimension of Cyber-Attacks Social, Political, Economic and Cultural”, IEEE Technology and Society Magazine,2011

[10] S. Z. Z. X. Sun Qindong, Zhang Deyun. Detection of dis-tributed denial of service attacks based on flow connection density. Journal of Xi’an Jiaotong University, 38(10):1048–1052, 2004 .