IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 7, July 2021

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.569

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

|DOI:10.15680/IJIRSET.2021.1007281|

A Review on Intrusion Detection System Using Machine Learning Algorithm

Savita K, Gardanmare [1], Dr. B.R. Bombade [2]

P.G. Student, Department of Computer Engineering, SGGSIE&T, Nanded, Maharashtra, India[1]

Professor, Department of Computer Engineering, SGGSIE&T, Nanded, Maharashtra, India[2]

ABSTRACT: **:** For detecting the attacks and classifying them we use the machine learning techniques that is intrusion detection System at network level and host level. system security and networks are still the main problem for researchers and administrators. The no. of hardware devices and the software's are used to remove hacking, attacks, and cybercrimes. With the creativity of increasing attackers, complete security system is impossible to handle. For security purpose of system no other tools and services are available as (IDS), by examining network traffic the IDS have the ability to identify malicious activities. Malicious attacks are rapidly changing and are fall out very large volumes requires a solution. So the dynamic change in attacking methods of malware, the datasets available publicly and updated systematically and benchmarked. In this paper an Artificial Neural Network (ANN) a type of deep learning model is explored to develop effective IDS to classify and detect cyber attacks. A thorough assessment of ANNs and other ML classifiers are displayed on different freely accessible benchmark malware datasets. The ideal organization boundaries and organization geographies for ANNs are picked through the accompanying hyper boundary choice strategies with KDD Cup 99 dataset. Every one of the examinations of ANNs are run till 1,000 ages with the learning rate differing in the reach [0.01–0.5]. The ANN model which performed well on KDD Cup 99 is applied on other datasets. Through a thorough exploratory testing, it is affirmed that ANNs perform well in examination with the traditional ML classifiers. And finally we propose an ANNs framework which utilize losing of information or data and ready for alert of possible cyber attacks

KEYWORDS: Intrusion detection, Cyber Security, machine learning, artificial neural network(ANN)

I. INTRODUCTION

Intrusion can be defined as any kind of unauthorized activities that cause damage to an information system.
In Intrusion Detection System the term intruder is just a person who is trying to gain access to a system or a network with criminal intension.

There are two major types of intruder-hacker, thieves

a. Outside intruder(masquerader)
b. Inside intruder(misfeasor)

Outside Intruder: Means he is not an authorized access to the network, he is trying to gain an access through the credentials.
Inside Intruder: Means he is an authorized access to the network. But using legitimate privileges or access to the resources which he have provided using that he stole the data and misuse that information or data. Inside Intruder are more harmful than the outside intruder.
For identifying and detecting the attack Intrusion Detection System (IDS) is used. An IDS continues monitors network and system is there any malicious attack that can imbalance your network. It is running in the background of the system when he get any suspicious activity then he directly alerts to the system administrator. Then system administrator takes some actions in a very quick manner to not damage any information or data. Based on intrusive behaviors, Intrusion Detection is classified into two major parts one is network based intrusion detection system(NIDS) and second is host based intrusion detection system(HIDS).

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

|DOI:10.15680/IJIRSET.2021.1007281|

What are we protecting:-

a. Confidentiality- Whenever the data or information is copied by unauthorized , the result may be loss of confidentiality. Confidentiality is very important factor in the kind of information .for example in case of credit card, businesses, hospitals , doctor offices. In unsecure network it can be corrupt. The result is known as loss of integrity, when data is modified in unknown ways.

Every bit of data, a business has is valuable, especially in today's society. Financial information, credit card details, trade secrets, and legal documents all demand strict confidentiality. To put it another way, only those who are permitted should have access to sensitive information.
Failure to maintain confidentiality means that someone who shouldn't have access to sensitive information has gained access. A breach of confidentiality, whether deliberate or unintentional, can have devastating consequences.

b. Integrity- Implies ensuring data accuracy and doing the right thing in circumstances, even no one is watching you. This means that information cannot be changed without permission. For example, if an employee leaves an organization, all data for that employee in all departments, such as accounts, should be updated to reflect the employee's status as JOB LEFT so that data is comprehensive and accurate, and only authorized people should be permitted to alter employ data.

Methods of intrusions:-
Methods used by intruders can contain any one, or even combinations of following intrusion

Types :- Attacks distinguishes based on the attacker's activities and targets. Each attack type can be categorized into four categories below.

- Denial-of-Service (DoS) attacks aim to prevent or restrict users from accessing services provided by a network or computer.
- The goal of probing attacks is to obtain information about the network or computer system.
- The goal of a User-to-Root (U2R) assault is for a non-privileged user to get root or admin access to a computer where the intruder had user level access.
- In Remote-to-Local (R2L) attacks, packets are sent to the victim machine. The cybercriminal learns about the user's activity and gains access to the computer system's rights that an end user would have.

Implementation for detection of attacks-
Attacker nodes send data packets to the target node or nodes. Using a random function, attacks are created at random. A Probe, R2L, U2R, or Dos attack is the type of attack that is generated. If the attack was not generated; the packet has been classified as a regular packet. The frequency of single characters and frequency of groups of characters are displayed in the packets that reach the victim.

STAGES OF COMPROMISE: AN ATTACKER'S VIEW:
Unauthorized or not having credentials for accessing the network or system is called attacker. For successfully attack a system to detect an intrusion correctly requires to understand the method.
Attacks can be classified into 5 categories
(a) Reconnaissance, (b)Exploitation , (c) reinforcement, (d)consolidation, (e)pillage
An attack can be stopped or detected during the first three phases, if they have reached fourth or fifth phase then the system or network will be fully damaged or in risk. Then the system or network is difficult to classified between a normal or an attack behavior.
In the reconnaissance phase, an intruder or attacker tries to collect data related to reachable hosts and services, and also the versions of operating systems and application
In the exploitation phase, an attacker utilizes a service with the aim to access the target system , a service may be as abusing, subverting. An abusing includes the stolen password or dictionary attack, subversion includes SQL injection
After entering into a system or network, an intruder follows concealment activity and then installs extra tools and services to take the advantages of the concession gained during the reinforcement phase. An intruder trying to gain full access based on the misused user account. Finally, an attacker makes use of the applications that are accessible from available user account. An intruder obtains a fully control on the device or system in te consolidation phase. And finally the pillage phase in this the intruder make a possible harmful activities include the theft of information and cpu time and spoof. Since computers and networks are assembled and programmed by humans, there are

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

|DOI:10.15680/IJIRSET.2021.1007281|

possibilities for bugs in both the hardware and software. These human errors and bugs can lead to vulnerabilities [1] purchasers by item search and proposal. Fourth, apparel order will be completed for labeling things. It will naturally mark labels via on-line media footage or treat depictions of things on on-line search footage. A multi-brands on-line search with several complete retailers ought to total all things and apply a number of standards, what is additional, impediments for unification and quality norm. This interplay is known as 'manage'. All in all, once each complete seek has to offer things at the multi-manufacturers, the guiding leader can take a look at the approaching things and select in spite of whether or not or now not they're going to be permissible or denied. This guiding interplay has issues in 2 viewpoints. in the 1st vicinity, there may be fantastic rate for paying each one of human experts to superintend continuously the interaction. 2nd, it is going to be tedious for screening each one of the matters from various retail retailers with the aid of human.

II. RELATED WORK

This section summarizes the research that has been done on leveraging the KDD dataset to construct machine learning techniques. It also gives a quick rundown of the main machine learning techniques and demonstrates how the KDD dataset can be used to evaluate and test different types of machine learning algorithms. For intrusion detections, they used the KDD dataset as a benchmark market dataset. They stated that the SVM algorithm requires a considerable training time, and as a result, SVM's applicability is limited.

All of the prior research papers made significant contributions while also demonstrating how the KDD dataset provides the necessary environment for testing and evaluating the various machine learning techniques. To evaluate the selected classifiers, the most essential performance characteristics, such as false negative and false positive, are of importance. As a consequence of the tests, the focus will be on determining the effectiveness of the machine learning classifier that achieves the required accuracy rate with the least amount of false negatives.

KDD DATASET PREPROCESSING AND ANALYSIS -

At the same time, the dataset is generally used in numerous fields for testing and evaluating intrusion detection algorithms, providing a good grasp of several intrusion behaviors.

For implementing several measures values the KDD dataset was brought into SQL server 2008, such as distribution of instance records, attack types, and occurrence ratios, in this work.

Techniques: Decision trees- Rule-based systems, SVM, NB, RF, Artificial neural network , GB, LR are just a few of the categorization algorithms available to create a classification model, each strategy employs a learning mechanism. A suitable classification strategy, on the other hand, not even handle the training data, but also accurately identify records in the class. The learning algorithm's main purpose is to create classification models with a high degree of generalization capacity.

Decision tree:-The decision tree is made up of three main components. A decision node is the first component, and it is used to designate a test attribute. The second type is a branch, in which each branch indicates a possible decision based on the test attribute's value. The third element is a leaf, which contains the instance's class and CART are only a few of the decision tree algorithms available . Assumptions of naïve bayes among the qualities. Using conditional probability equations, Naive Bayes solves queries like "what is the likelihood that a particular type of assault is occurring, given the observed system activities?" The traits that have differing probabilities of arising in attacks and in regular conduct are used by Naive Bayes. The conditional independence assumption attribute of the Nave Bayes classification model makes it one of the most popular in IDS because of its ease of use and calculation efficiency . However, if this independence assumption is false, the system does not work well, as proved by the KDD'99 intrusion dataset of intrusion detection, which has complicated attribute dependencies.

(Machine Learning) is a major artificial intelligence subfield that allows machines to learn and perform specified tasks. Machine learning comprises a group of techniques and algorithms that may predict certain future occurrences or categorize certain information by learning the patterns in existing data. LR, SVM (Support Vector Machines), algorithm Naive Bayes [2], decision-making, gradient boosting, random forest & deep learning, are all of the most significant algorithms in this subject.

Naïve-bayes – The naive Bayes' principle is used in conjunction with powerful independence presumption among the attributes in this technique. Using conditional probability equations, Naive Bayes solves queries like "what is the likelihood that a particular type of assault is occurring, given the observed system activities?"

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

|DOI:10.15680/IJIRSET.2021.1007281|

The traits that have differing probabilities of arising in attacks and in regular conduct are used by Naive Bayes. The conditional independence assumption attribute of the NB classification model makes it one of the most popular in IDS because of its ease of use and calculation efficiency.

Random Forest - Random Forest method, which is an ensemble, supervised machine learning technique. In most circumstances, its classification performance outperforms that of other single classifier models, and it can handle both binary and multi classification problems. The primary idea behind RF is to generate several decision trees using randomly sampling and replacement, with the final outcome being determined by voting. The following is the procedure for creating RF.

a) Extract samples from the dataset using random sampling with replacement to create a training subset.

b) The training subset uses features chosen at random without replacement from the feature set as the basis for splitting each node in the decision tree. A full decision tree is constructed from the root node to the bottom node.

c) By repeating steps (1) and (2) K times the decision trees are created. Combining these decision trees yields the RF classifier. These decision trees vote on the classification outcome.

LR- is a method of ML that seeks to discover a linear pattern of variables by placing one line on the data curve [3]. It can also be used for grading reasons.

SVM algorithm- this is one of the most used algorithm in data classification [4]. The best data classification for the provided data is found. SVM can deliver very excellent performance generalization.

A decision tree- is another data modeling technique that employs tree-like structures to categorize decisions to output the given data class. A whole learning approach called random forest is developed when certain decision trees are used [5].

Gradient boosting-In some learning issues, a gradient boosting technique is commonly utilized. It creates an ensemble model by using certain weak models. The capacity to decrease bias and variation in the model is one of the benefits of this technique.

ANN- ANN also called as neural networks in this millions of neurons are presented, it works like human brain processes information. In data mining neural networks find good applications. For example pattern matching and economics. Nodes are divided into two parts one is submission part and other one is function part. The number of signals are passed to nodes and it signals are associated with their weighted sum will calculate and next passed to activation function.

An artificial neural network has following three layers-
Input layer – Actual input signals are coming which are been used to generating and processing appropriate output. For each observation receives the input values to how people explain to themselves. The number of explanatory variables are equally to the number of input layers. They never change the data means they are passive. They receives single number and gives the duplicate number to their many outputs. Sigmoid and linear function both are used in input layer as an activation function.
Hidden layer- how much hidden layer are more the accuracy and precision will be more of the generated output on the taken action. Sigmoid and linear function both are used in hidden layer as an activation function.
Output layer- From input layer and hidden layer which receives the connection to output layers. In output layer we use the heavy side steps function. It deals with only two states 1 or 0 because of output nodes are help in decision support.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

|DOI:10.15680/IJIRSET.2021.1007281|

ANN types

1. Feed forward network- This is a forward network nature means there no loops form, whenever any signal come and go it never come back word. Example when input layer giving values to hidden layer but hidden layer never come back to input layer. No feedback or no loops are available

2.Feedback network- In feedback network it allows loops and feedback. Whenever your signals transforms input layer to output layer in a one direction that will also come backward. Direction is allowed in both ways.

Disadvantages-

Its complexity is very high manse very complex to implement also called as recursive/recurrent network.

III. METHODOLOGY

This section offers a description of ways for evaluating intrusion detection and prevention system (IDPS) methodologies and the systems that are based on these methodologies. Table 1 can be used to evaluate any intrusion detection and prevention system (IDPS) whether it uses one of the three main methodologies or a combination of the two or more of the other methodologies.

a) Resistance to evasion

The intrusion detection and prevention system (IDPS) should be able to detect evasion attempts and stop them. These attempts are more common with the signature and stateful protocol analysis based intrusion detection and prevention system (IDPS) due their dependence on signatures. Anomaly based intrusion detection and prevention system (IDPS) have better resistance to evasion, but the hybrid based system offers the best resistance to evasion attempts due to the combination of other methodologies.

b) High Accuracy Rate

An IDPS should have a high accuracy rate when detecting and analyzing possible threats. The signature based methodology has a high accuracy rate on known threats but its overall rate is lower that the anomaly based methodology. which can detect previously known threats. The hybrid based methodology offers the best accuracy rates.

c) Market Share

Market share is the measure of the methodology's dominance in the deployed systems. The signature based methodology far outweighs the other three methodologies, followed by Stateful protocol analysis. The anomaly and hybrid based methodology are the bottom but their adaption is growing much faster and will soon surpass the first two methodologies.

d) Scalability

Scalability is the ability of an IDPS to scale and grow with environment once deployed. The signature and Stateful protocol analysis based methodologies are easy to scale since they are based on signatures that can be easily scaled. A hybrid based methodology can be easily scale depending on the underlying methodologies. The anomaly based methodology is the least scalable methodology due the time it requires to learn and build its baseline profiles.

IV. CONCLUSION

For detecting attacks there are many approaches in intrusion detection system. In our paper the many techniques are used LR, SVM, GB, DT, NB, RF and ANN. The above analysis of different models states that the ANN model not that much best fits our data considering both accuracy and time complexity even though ANN model is best because losses

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 7, July 2021 ||

|DOI:10.15680/IJIRSET.2021.1007281|

of data is very low and as amount of data is increasing performance of (Deep learning) is increasing that's why the ANN model is better than SVM or LR . When we want to work with large amount of data. We proposed the accuracy score of SVM and ANN and proved that ANN is better performed than SVM . In this we are giving high training time than SVM and also more testing time it gives more accuracy score than SVM and also number of losses are 0.5521 0,0497 and accuracy score is 0.9682 0.9848.

REFERENCES

[1] M. Tang, M. Alazab, Y. Luo, and M. Donlon, ''Disclosure of cyber security vulnerabilities: time series modelling,'' Int. J. Electron. Secur. Digit. Forensics, vol. 10, no. 3, pp. 255–275, 2018.
[2] Bonaccorso G. Machine learning algorithms. Packet Publishing Ltd; 2017
[3] Kleinbaum DG, Dietz K, Gail M, Klein M, Klein M. Logistic regression. Springer; 2002.
[4] Cortes C, Vapnik V. Support-vector networks. Machine learning. 1995;20(3):273–97.
[5] Koning M, Smith C. Decision Trees and Random Forests: A Visual Introduction for Beginners. Amazon Digital Services LLC - Kdp Print Us; 2017.
[6] Saritas MM, Yasar A. Performance analysis of ANN and Naive Bayes classification algorithm for data classification. *Int J Intell Syst Appl Eng*. 2019; **7**(2): 88- 91. https://doi.org/10.18201//ijisae.2019252786
[7] Bangyal WH, Ahmad J, Rauf HT, Shakir R. Evolving artificial neural networks using opposition based particle swarm optimization neural network for data classification. Paper presented at: Proceedings of the International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT). Sakhier, Bahrain; 2018:1-6.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.569

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details