JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

AI-Based Intrusion Detection System in Cloud Computing

Kalpana Verma
Department of Computer Science and Engineering Institute of Technology& Management Lucknow,

ABSTRACT – This global network, made up of hundreds of millions of computers running different hardware and software combinations, allows people to interact and do business. Because computers are connected to one another, this makes it easier for hackers to misuse resources and launch Internet attacks. The creation of a security-focused strategy that might be adaptable and agile in view of the rising frequency of cyberattacks faces substantial obstacles. To identify online dangers, one needs an intrusion detection system (IDS). An intrusion detection system, or IDS, is necessary to keep a network secure. We must also create a strong IDS for the cloud platform because it is constantly growing and permeating more aspects of our daily life. However, using standard intrusion detection systems in the cloud may provide challenges. The additional detection overhead introduced by the pre-defined IDS design may lead a cloud segment to become overloaded. within the framework of a flexible architecture networked system. We demonstrate how to fully utilize available resources without overtaxing any particular cloud server by using a neural network-based intrusion detection system. Even more efficiently, the suggested IDS employs machine learning with neural networks to detect new threats.

KEYWORDS- Artificial Intelligence, Intrusion Detection System, Machine learning Algorithm.

I. INTRODUCTION

The Internet is a vital component of modern life, used for everything from commerce to education to leisure. Businesses are using the Internet more and more frequently to obtain information. Because there is so much information available on the Internet, there are many ways in which a computer system might be compromised. Attacks and intrusions via the internet are growing more frequent. "Any combination of actions that seek to breach the security objectives" is one way to define an invasion or assault. The following are some of the most crucial security goals: assurance, accountability, integrity, confidentiality, and availability. Four types of attacks can be distinguished: Denial of Service, Probing, User to Root, and Remote User. Numerous anti-intrusion solutions have been created to thwart a significant percentage of cyberattacks. Halma and Bauer (1995) have detailed six anti-intrusion systems, including IDS for them. Countermeasures and detection are included in the remaining five. Perfectly identifying an intrusion is the most important of these components.

II. RELATED WORK

Saud Mohammed Othman and Fadl Mutaher Ba-Kiwi [1] presented their work on an intrusion detection model using machine algorithms in a big data setting. This study shows that the growing volume of Big Data has changed the importance of analytic and data security technologies. An intrusion detection system (IDS) keeps an eye on and evaluates data to find any potential breaches into the system or network. Because of the amount, variety, and speed at which data is generated within networks, traditional methods for identifying network attacks have grown more intricate. IDS analyzes big data accurately and efficiently by utilizing big data methodologies. They were able to build an intrusion detection model that can handle massive volumes of data by employing the Spark-Chi-SVM architecture. For the purpose of expediting the handling and analysis of data, the Spark Big Data platform was employed in the suggested manner. The categorization process is made more complicated and time-consuming by the enormous dimensionality of big data. According to Sharmila Wagh and Vinod K. Pachghare, a survey of intrusion detection systems using machine learning techniques was given [3]. The writers claim that in today's world, computers and network-based technologies are becoming more and more typical. In

the age of computers, the significance of network security cannot be emphasized. An Intrusion Detection System (IDS) is designed to identify system attacks and categorize system activity into normal and abnormal forms. Machine learning-based intrusion detection systems (IDS) are becoming more and more popular. This work describes an intrusion detection system based on cloud computing and distributed machine learning [4]. Cloud providers' edge network components will be integrated with the recommended cloud system. Consequently, inbound network communication may be intercepted by the physical layer edge network routers. Before being forwarded to a module that applies the Naïve Bayes classifier to find abnormalities, the network data acquired by each Cloud router is preprocessed using a time-based sliding window approach. Each anomaly detection module has access to server nodes powered by MapReduce and Hadoop when there is an accumulation of network congestion. Each time frame is assigned to a server for syncing anomalous network traffic data. Every router in the system provides data about aberrant network traffic to this server. After that, each attack is subjected to Random Forest classifiers to determine the type of attack that was carried out. The paper "Machine Learning-Based Network Intrusion Detection Detection: Dimensionality Reduction Approaches," by Hassan Musafer and Ali Alessa, was just released. [5]. They assert that concerns regarding the security of computer networks are shared by all stakeholders, including corporations, governments, and customers. The tactics that attackers employ to carry out such attacks also change as it gets harder to defend networked systems against attacks. Increasing the efficacy of the current intrusion detection systems is a part of the answer. Because machine learning is so effective at creating intrusion detection systems, this method is gaining popularity (IDS). Enhancements to IDS attributes such as discrimination and representation result in a notable rise in the system's overall

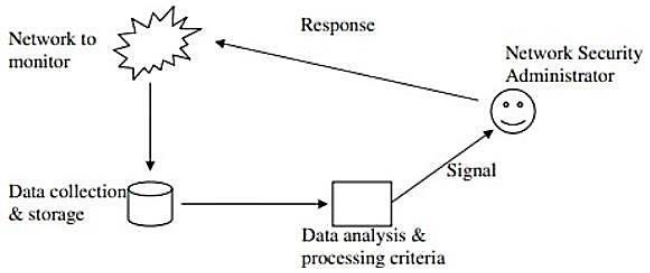


Figure 1: Architecture of Intrusion Detection System

In response to the increase in computer assaults, numerous IDS designs have been put forth. Figure 1 shows a popular IDS architecture proposed by Axelson (1999). According to Axelson (1999), the most typical IDS components are as follows: To keep an eye out for any unwanted intrusions, the network to be watched over must be identified. As an alternative, the entire network might be used for this. Data from multiple sources is gathered and stored by a disc-based event data collection and storage system. The brain of the IDS is its data processing and analysis unit. This device has all the features required to identify abnormal traffic patterns. A signal is produced upon the identification of an attack. When an intrusion detection system (IDS) detects an issue that has to be rectified, the system might decide to handle it on its own or notify the network administrator. This program may alert a network security administrator about potentially dangerous activities or trigger an automatic intrusion response system. There are multiple ways to categorize the modules of an IDS. The way that data is gathered and stored allows for the differentiation of two types of IDS: Host-based IDSs are those that gather information from a host. Operating system logs, system calls, and logs (such) For the purposes of this investigation, the dimensionality of the characteristics was decreased using Principal Components Analysis (PCA) (PCA). Combining these two methods might produce low-dimensional features that can be used to create classifiers like Bayesian networks, Random Forests, Linear Discriminate Analyses (LDA), and Quadratic Discriminate Analyses (QDA).

A. Intrusion Detection System-

An efficient security system that can recognize, stop, and possibly even react to cyber attacks is one of the essential elements of security infrastructure. For Security services monitors target sources of activity using a range of techniques, such as audit and network traffic data in computer or network systems. An intrusion detection system must promptly and accurately identify every threat (IDS). Network managers may find objective security issues with the aid of IDS. There's a chance that outsiders will try to get unauthorized access to the network. security objectives are violated when resources are made inaccessible to insiders who misuse their system resources or when security infrastructure is compromised. Additional sources of information include application logs, NT events and CPU utilization logs, and other logs. Host-based IDS can easily identify buffer overflow attacks because they are operating system-independent. Switched networks and encrypted data are not compatible with these methods. Network-based intrusion detection systems (IDS) are those that collect data from the network in the packet format. These IDS can be installed on almost any platform and configured on almost any kind of system.

B. Descriptions of CICIDS2017 dataset

Researchers in the field of intrusion detection have already reported accuracy rates as high as 98 percent or higher and false alarm rates as low as 1 percent. This higher rate of precision forced scientists and businesses to invest time and resources into creating effective products. Actually, the industry has only acknowledged a few number of models as capable of designing an IDS. In addition to including the most current network assaults, the dataset also satisfies all requirements for attacks that actually happen in the real world through the analysis of temporary IDS models and training and testing datasets. When we looked at the properties of this dataset, we found very few errors. One glaring weakness is the size of the dataset, which was assembled from five Traffic data from the Canadian Institute of Cyber security is dispersed among eight files. One dataset may be used to create an IDS. The dataset is not appropriate for training any IDS since it contains a large number of duplicate entries. We found that the dataset has a significant class disparity even if it includes modern attack scenarios. Class imbalance datasets have the potential to bias the classifier in favor of the majority class. In an effort to address these problems, a subset of the CICIDS2017 dataset was made available to the research community for use in the development and testing of detection algorithms. The description of the CICIDS2017 dataset, which I worked with, is displayed in the figure below. We are able to locate the source of this problem. The Cyber security Institute of Canada

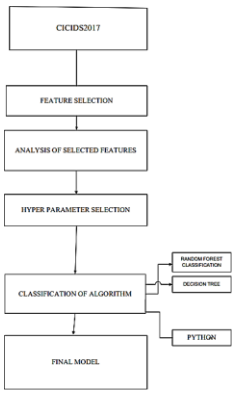


Figure 1: Description of the file containing the CICIDS2017 dataset

C. *Shortcomings of the CICIDS2017*

As we have already mentioned, the CICIDS2017 dataset has a number of shortcomings. The goal of our effort is to address such shortcomings so that subsequent researchers can gain a deeper understanding of the dataset. Scattered Presence: Table 1 shows that the CICIDS2017 dataset's data is now included in eight files. Taking care of individual files takes time. Consequently, we generated a single file containing 3119345 instances of every file under consideration.

Table 2: Attack found in a day wise

Name of Files	Day Activity	Attacks Found
Monday-WorkingHours.pcap_ISCX.csv	Monday	Benign (Normal human activities)
Tuesday-WorkingHours.pcap_ISCX.csv	Tuesday	Benign, FTP-Patator, SSH-Patator
Wednesday-workingHours.pcap_ISCX.csv	Wednesday	Benign, DoS GoldenEye, DoS Hulk, DoS Slowhttptest, DoS slowloris, Heartbleed
Thursday-WorkingHours-Morning-WebAttacks.pcap_ISCX.csv	Thursday	Benign, Web Attack – Brute Force, Web Attack – Sql Injection, Web Attack – XSS
Thursday-WorkingHours-Afternoon-Infilteration.pcap_ISCX.csv	Thursday	Benign, Infiltration
Friday-WorkingHours-Morning.pcap_ISCX.csv	Friday	Benign, Bot
Friday-WorkingHours-Afternoon-PortScan.pcap_ISCX.csv	Friday	Benign, PortScan
Friday-WorkingHours-Afternoon-DDos.pcap_ISCX.csv	Friday	Benign, DDoS

D. *Huge Volume of Data*

Once all of the data files have been integrated, all of the possible labels for recent assaults may be located in one place. On the other hand, the merged dataset gets much larger. The challenge arises from the sheer volume of information available. One disadvantage is that data loading and analysis take longer.

E. *Missing Values*

There are 203 occurrences of missing metadata and 288602 cases with a missing class label in the aggregated CICIDS2017 dataset. We discovered that this was an issue. All but a handful of the original data points were removed to provide a dataset with 2830540 unique occurrences.

F. *Classification of Algorithm*

The study's findings indicate that a classifier algorithm's accuracy, scalability, speed, and learning capability are the most important factors to consider when making a decision. This theory has been supported by research and findings using five different classification algorithms: Random Forests, Bayesian Network, Random Trees, Naive Bayes, and J48 classifiers. This study demonstrates that random forest trees can learn and perform very well in terms of assault detection with the use of the Information Gain feature selection. The Bayesian Network performs better than other algorithms in classifying attacks. A scalable and effective technique is Random Tree. When it comes to data classification, Naive Bayes is a superior option than other algorithms because of its minimal model complexity.

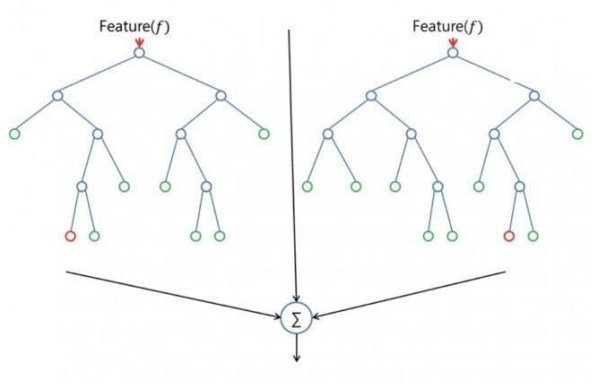


Figure 2: Random Forest classifiers tree.

1. ***Random Forest (RF)***

One method for ensemble classifiers is Random Forest. An ensemble of decision tree classifiers is a "forest" of classifiers. At each node, attributes are chosen at random to create each decision tree. Breich first presented the random forest algorithm in 2001.

2. ***Bayes Network (BN)***

A modeling technique known as Bayesian Network (BN) is used to express probabilistic links between variables of interest. Presumptions regarding the behavior of the model

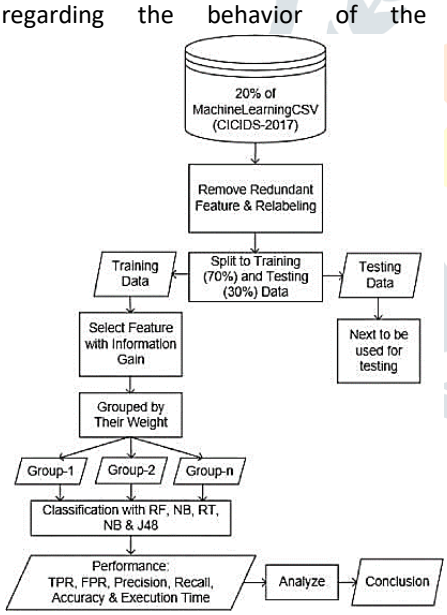


Figure 3: Experimental design

To classify each feature group or feature subset, the Random Forest (RF), Bayes Net (BN), Random Tree (RT), Naive Bayes (NB), and J48 classifiers are employed, in that order. The accuracy rate, precision rate, recall rate, accuracy rate, true positive rate, false positive rate, and proportion oftarget system are used to determine how accurate this technique is. If the assumption is significantly altered, then detection accuracy is reduced.

3. ***Random Tree (RT)***

A decision tree constructed with a random collection of attributes is referred to as a "random tree". A decision tree has numerous nodes and branches that can be connected to one another in a number of ways. An attribute under test is represented by a node, and the results are shown by branches. Decision leaves, which have the form of class albethey e, show the ultimate selection chosen following the computation of all characteristics.

4. ***Naive Bayes (NB)***

The Bayesian categorization system states that it is possible to statistically predict the likelihood of belonging to a particular class..

5. ***J48***

A popular machine learning algorithm that is a component of the decision tree algorithm is J48 or C4.5. This method forms a decision tree using the concept of entropy. When completing the analysis, factors like incorrectly classified data and analysis execution time are taken into consideration. A comparison of the True Positive Rate and the False Positive Rate is also provided. Additionally, a comparison is provided between the True Positive Rate andthe rate of false positives. At this stage of the process, the method is 10-fold cross-validation. It is essential to examine and contrast each classifier algorithm's TPR, FPR, Accuracy, Precision and Recall, Percentage of Incorrect Categorization, and Execution Time.Throughout the entire process of learning and testing,

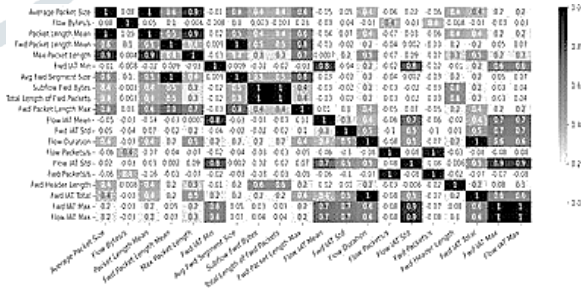


Figure 4: Accuracy graph

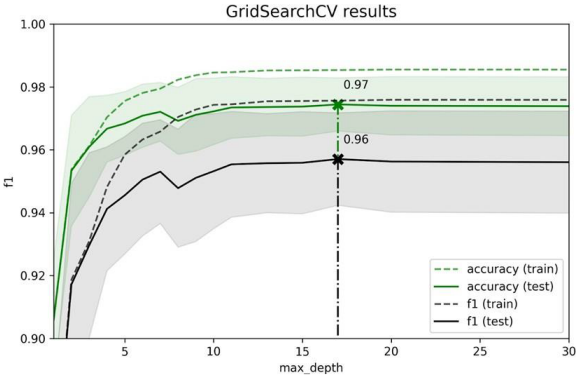


Figure 5: Correlated Heat Map

III. EXPERIMENTAL RESULT

Metrics like True Positive Rate (TPR), False Positive Rate (FPR), Precision, Recall, Accuracy, Percentage of incorrectly Classified, and Execution Time are used to assess the effectiveness of Information Gain in addition to the five different classifier approaches. When taken as a whole, these measurements are referred to as the metric set. Throughout the training, multiple times points in time are simulated to mimic the actual execution, with the aim of enhancing and honing it even more. In this trial, the J48, RT, NB, BN, and RT classifiers There is ten-fold cross-validation. At this point, it is imperative that you draw some conclusions or deductions. are grouped in a multitude of ways to categorize every distinct feature subset. For the random tree, RT, NB, and J48 stand for it; also, RT implies the random tree. Throughout this experiment, a 10-fold cross-validation method was employed to determine the effectiveness of categorizing algorithms. The 10-fold cross-validation is used because it reduces the total computation time while preserving the accuracy of the classification algorithms. Ten random folds of the input dataset of the same size will be created from it as a direct and immediate consequence of this. Nine of the ten-fold data sets will be used in the cross-validation procedure.

IV. OVERALL PROCESS

Start selecting features based on the feature weight, and

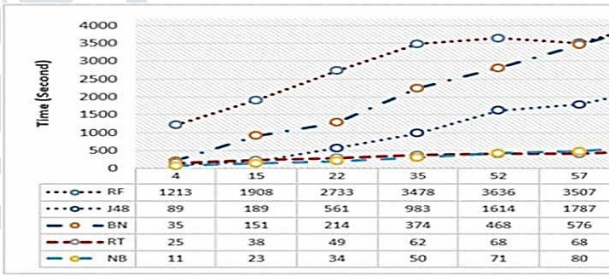
- For each function Fr in the data for Feature Ranked
- Reduce the number of features from 77 to n based on the feature weight
- Features subsets, TPR, FPR, accuracy, recall, and precision are included in the output
- Feature Ranked data are accepted as input
- A step-by-step procedure ()

Detection	RF	BN	RT	NB	J4
Normal	0.960	0.943	0.960	0.174	0.9
DoS/ DDoS	0.992	0.996	0.992	0.999	0.9
Port Scan	0.995	0.992	0.995	0.983	0.9
Bot	0.438	0.642	0.430	0.687	0.3
Web Attack	0.072	0.031	0.072	0.000	0.0
Infiltration	0.000	0.000	0.400	0.400	0.0
Brute Force	0.792	0.991	0.792	1.000	0.7
Recall	0.965	0.962	0.970	0.903	Na
Precision	NaN	0.953	0.965	0.335	0.9
FPR	0.016	0.010	0.016	0.026	0.0

Figure 6: Performance Metric Using Four Features

Then save them on Feature Groups. Group 1 is made up of all characteristics with weights greater than or equal to 0.6. Group 2 is made up of all characteristics with weights greater than or equal to 0.4%. Group 4 is made up of all features with weights greater than or equal to 0.2. Group 5 is made up of all features with weights greater than or equal to 0.1. Group 7 represents the entire set of traits.1) With respect to every grouping of features.2) Provide RF, BN, RT, NB, and J48 with specific characteristics using CICIDS-2017-20 percent

0.010), the lowest FPR. Using these four features, classifiers can identify only DoS/DDoS, Port Scan, and Brute Force assaults. Only NB and the use of these four (4) features are impacted in terms of regular traffic. As far as regular traffic is concerned, this just impacts NB. This study further analyzes the effect of execution time for the specific feature approach under discussion. The picture below shows a summary of the execution times for each feature subset using RF, J48, BN RT, and NB. A significant impact is seen on the pertinent characteristics procedure's RF, J48, and BN. The run times of RT and NB are incredibly short. Generally speaking, the more features to assess, the longer it takes to finish.



V. CONCLUSIONS AND FUTURE WORK

A series of experiments was conducted to show how feature selection can improve the accuracy of anomaly detection. Out of the feature sets 15, 22, and 35 that were tested, Information Gain was found to be the best information classifier because it was accurate in determining the amount of data contained in each feature set. On the other hand, feature sets 52, 57, and 77 are the best for J48. Using feature sets 52, 57, and 77.5, all communication could be detected, even though the precision of BN is lower than that of RF and J48 despite its lower level of accuracy. It was also shown that the traits selected reduce the FPR, especially for BN. Finally, the results of the experiments indicate that the number of features selected has an effect on the program's execution time. Information Gain suggests arranging attributes in order of weight values. However, determining the minimal weight value—which affects the

number of features selected—needs to be done by an expert. Our plan is to test multiple feature selection procedures in order to identify the most effective one. Future studies will examine every characteristic subset that affects an assault.

REFERENCES

1 Suad Mohammed Othman and colleagues, "Intrusion detection model using machine learning algorithm on Big Data environment." Big Data Journal 5.1 (2018): 34.

2 Aleksander Essex, Salo, Fadi, and Ali Bou Nassif. Computer Networks 148 (2019): 164–175, "Dimensionality reduction with IG-PCA and ensemble classifier for network intrusion detection."

3 . Wagh, Vinod K. Pachghare, Satish R. Kolhe, and SharmilaKishor. "Survey on intrusion detection system using machine learning techniques." 78.16 (2013)

4 International Journal of Computer Applications. In June 2019, Chiba, Z., Abghour, N., Moussaid, K., El Omri, A., & Rida, M. A Superb Network Intrusion Detector for Cloud Environments based on an optimized self-adaptive heuristic search algorithm combined with deep learning. (pp. 235-249)

5 in International Conference on Networked Systems. Springer, Cham Mustapha Belouch, Karim Adel, Mohamed, and Idhammad. "Distributed intrusion detection system for cloud environments based on data mining techniques." Computer Science Procedia 127 (2018):35–41.

6 "Features Dimensionality Reduction Approaches for Machine Learning Based Network Intrusion Detection," Abdulhammed, Razan, et al. 2019's Electronics 8.3: 322.
6. In the IOSR Journal of Computer Engineering (IOSR-JCE), Basaveswara Rao B and Swathi K (2016)

7 Variance-Index Based Feature Selection Algorithm for Network Intrusion Detection, Volume 18, Issue 4, Swathi, K. & Basaveswara Rao B (2017). Quick kNN Classifiers for Network Intrusion Detection Systems. Science and Technology in India, 10(14).

8 Bobba Basaveswara Rao, Kailas am, and Swathi. International Journal of Rough Sets and Data Analysis (IJRSDA) 6.2 (2019): 61-72. "Impact of PDS Based kNN Classifiers on Kyoto Dataset." The article

9 "IntrusionDetection System Based on Fast Learning Network in Cloud Computing" was written by Ali, Mohammed Hasan, and Mohamad Fadli Zimmeri. Letters on Advanced Science 24.10

10 Ahmed, H. A. S., Ali, M. H., Kadhum, L. M., Zolkipli, M. F., & Alsariera, Y. (2017). A review of the difficulties and security risks associated with cloud computing. Journal of Telecommunication, Electronic and Computer Engineering (JTEC),

11 Ali, Mohammed Hasan, et al. "A hybrid Particle swarm optimizationExtreme Learning Machine approach for Intrusion Detection System." 2018 IEEE Student Conference on Research and Development (SCOReD). IEEE, 2018.

12 Umer, Muhammad Fahad, Muhammad Sher, and Yaxin Bi. "Flow-based intrusion detection: Techniques and challenges." Computers & Security 70 (2017): 238-254.

13 "Nsl-kdd data set for network-based intrusion detection systems." Available on: http://nsl.cs.unb.ca/KDD/NSLKDD.html, March 2009.The article "A Fast KNN Based Intrusion Detection System for Cloud Environment" by Basaveswara Rao B and colleagues was published in the Journal of Adv Research in Dynamical & Control Systems in 2018. It can be found on pages 1509–1515.

14 Dieter Landes, Andreas Hotho, Deniz Scheuring, Sarah Wunderlich, and Markus Ring. Computers & Security (2019) published "A Survey of Network-based Intrusion Detection Data Sets." 18. Ahmim, A., Ferrag, M., and Maglaras, L.