iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** I **Month of publication:** January 2022

DOI: https://doi.org/10.22214/ijraset.2022.40128

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue I Jan 2022- Available at www.ijraset.com

A Survey on Intrusion Detection Systems Using Various Data Mining Techniques in Big Data Environments

Sandeep S Budhya[1], Siddharth Lohia[2], Deepak. G[3], Harish Kumar N[4]
[1, 2]*Student,* [3]*Associate Professor,* [4]*Assistant Professor, Department of CSE, Dayananda Sagar College of Engineering*

Abstract: ***Intrusion Detection is a topic that is of interest both in the corporate world as well as academia. In the advent of Big Data Analytics, multiple analytics techniques can be used on the enormous amounts of data that is being generated every single day in order to discover knowledge. This inherently poses a threat to the security and privacy of all the parties involved. Therefore, it is a necessity in today's world to reinforce the security systems with robust Intrusion Detection and Prevention Systems. A nominal Cybersecurity System can no longer suffice for detecting and minimizing the damage from cyber-attacks especially since many of the attacks do not fall under a pre-discovered category. In this paper we review the various works particularly concerning Big Heterogeneous Data as well as present opportunities for further research to be conducted in these areas.***

Keywords: Intrusion, Detection, Cybersecurity, Big Data, Machine Learning, KDDCup99

I. INTRODUCTION

Cybersecurity has become critical for society to function. This is due to it becoming increasingly dependent on computer systems for providing financial services, industrial work, healthcare, and other important aspects. One of the most important areas of research in Cybersecurity is Intrusion Detection. Awareness about an attack is quintessential to mitigate or prevent attacks, it is essential to being able to react and defend against attackers.

Intrusion Detection and Cybersecurity Analytics plays a vital role in forensics to identify successful breaches post their occurrence. By looking for patterns and trends of attacks that are generally not identifiable using Intrusion Detection, defenses can be further improved.

Important knowledge can be gained by knowing if financial information such as credit/debit card data has been leaked, in order to take additional precautions or legal actions if it is a serious issue.

Intrusion Detection can also be very helpful beyond detecting cyber-attacks such as catching out abnormal network behavior, detect accidents or undesired conditions.

This study draws comparisons to some of the advancements in Intrusion Detection technology that have risen to popularity, along with important propositions such as monitoring a wide array of security event sources that are heterogeneous in nature. Cyber-attacks have metamorphosed to become more malicious as well as sophisticated, Intrusion Detection products too have caught up in quite effective at handling them.

They are now capable of monitoring an ever-increasing number of diverse heterogeneous security event sources. IDSs were the first proprietary products developed to flag potential cyber-attacks.

They can be configured in one of two ways, for misuse detection or for anomaly detection. An IDS configured for misuse detection evaluates data that it is monitoring against a known database of signature based attacks to discern if they are identical. An IDS configured for anomaly detection, on the other hand, evaluates by monitoring against a baseline which is considered normal, and can notify based on abnormal behavior.

In this paper we discuss the various ways of implementing an Intrusion Detection System(IDS) using the most widely used benchmark, the KDDCup99 Dataset using Big data and Machine learning techniques. We survey a number of papers and come up with various different methods for feature selections on the dataset. This is an important step in making a more efficient and accurate Intrusion Detection System with less false positive rates, as the dataset is enormous and contains a lot of features some of which are not needed in the training of the model and decrease efficiency.

Our Goal through this project is to develop an efficient and accurate Intrusion Detection System to make the world a cybercrime free place and to protect the Integrity, confidentiality and availability of the information and data of an organization and individual.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue I Jan 2022- Available at www.ijraset.com

II. BACKGROUND AND MOTIVATION

The incremental increase of enormous amounts of data have recently become the driver of change in the technological world. Information security and Data Analytics for Big Data have become top priority in the industry today. Due to the availability of advanced open source tools, a general purpose security system can prove to be effective when it comes to detecting most cyberattacks (a one for all solution) however, when faced with attacks that have never been seen before, they tend to be inefficient or inaccurate. The goal of this paper is to conduct a literature on a number of solutions that have been proposed, some which are general purpose and some, purposefully built to detect anomalies.

III.LITERATURE SURVEY

1) *Tchakoucht TA, Ezziyyani M on "Building a fast intrusion detection system for high-speed-networks: probe and DoS attacks detection." Procedia Comput Sci. 2018.*

Proposed solution was lightweight and suitable for detection of attacks in high speed networks. Among the 41 features used in KDD'99 the most important features were used to improve accuracy and efficiency of the model. Six feature selection methods were used. A resampled version of the KDD'99 was used to assess the system. Results showed good detection accuracy of around 99.6% and false positive rates of around 0.3% for DoS attacks using C4.5. An accuracy of 99.8% and false positive rate of around 2.7% for Probe attacks using Naïve Bayes. It was demonstrated that processing time was also saved when evaluated using the best selected feature subset. The feature subset proposed was thus recommended for use in high speed networks, it has 19 features for probe detection and 9 features for DoS detection.

2) *Sahasrabuddhe A, et al. Survey on "Intrusion detection system using data mining techniques" in Int Res J Eng Technol. 2017.*

This paper talks about the different types of SQL injection attacks and various data mining algorithms used in intrusion detection. It then talks about a system which uses the Naive Bayes Classifier algorithm to detect SQL Injection attacks. It sheds light on the various types of SQL injection attacks that can be performed on a system such as Tautologies, Blind Injection, Piggy-Backed Queries Et al. It discusses the need for an intrusion detection system to detect SQL Injection attacks since it is one of the oldest forms of cyber-attacks. The different data mining techniques available are discussed in brief, finally arriving at Naïve Bayes Classifier algorithm which is used in the proposed model.

3) *Ferhat K, Sevcan A on "Controlling fraud by using machine learning libraries on Spark" in Int J Appl Math Electron Comput. 2018.*

This paper proposes the use of K-Means clustering algorithm to detect abnormal network behavior. One of the most widely used algorithm in data mining is the k-Means clustering algorithm. These are the type of algorithms that divide data on its own into smaller clusters or sub-clusters. It places statistically similar data in the same group. Apache Spark, which is an open-source unified analytics engine for Big data processing is being used. They detected ten abnormalities out of the four hundred thousand records from the KDD'99 10 percent dataset using the proposed method.

4) *Lee Y on "Toward scalable internet traffic measurement and analysis with Hadoop" in ACM SIGCOMM Comput Commun Rev 2013.*

A solution based on Apache Hadoop for packet analysis (Netflow trace analysis) has been suggested. A throughput of 14 Gbps was achieved utilizing a 200-node Hadoop testbed, for 5 TB of files data. The solution given incorporates a binary format for obtaining traces concurrently, MapReduce algorithms for Netflow, TCP, IP and HTTP analysis, and a Hive-based system for simplifying queries. The proposed solution includes, multiple tools built over this such as standard 5-tuple flow statistics, TCP re-transmission statistics and DDoS analysis. A series of experiments were performed for Accuracy, Scalability and a comprehensive comparison with CoralReef and RIPE's Pcap. The paper concludes with the proposal that this approach can provide the scale-out feature of Hadoop for handling the ever increasing traffic data.

5) *Peng K. Et al. on "Intrusion detection system based on decision tree over Big Data in fog environment" 2018.*

This paper discusses Intrusion Detection System in the perspective of Fog computing. The Fog nodes being close to the end users have limited computing ability and hence might come across some security challenges. The system of Fog nodes may get destroyed by traditional network security attacks and hence an Intrusion Detection System(IDS) can be a proactive security measure which can be used in the Fog environment.

Various Naïve Bayes methods as well as KNN is compared with the proposed model that is, the Decision tree algorithm. The Decision Tree algorithm was found to be the best if only the precision is considered for the IDS (Intrusion Detection System) issue. The BernoullliNB algorithm was found to be better when considering only the calculation time. After conduction above tests the Decision Tree method was found to be the most sustainable solution even though it is not the best method from calculation opinion.

6) *Manzoor MA, Morgan Y on "Real-time support vector machine based network intrusion detection system using Apache Storm" in IEEE 7th annual IEMCON, 2016.*

IDS is an essential component for managing a network. It works as a security mechanism for a network. This work presents a high speed Intrusion Detection System which can work in real time. The system proposed in the paper is a Support Vector Machine based Intrusion Detection System for networks which uses the KDD99 data-set. The system proposed is developed to handle enormous amounts of streaming data i.e. Big Data. The results of the experiment conducted on the proposed system show that it is viable for stream processing network traffic data for any intrusion detection with high accuracy.

7) *Dahiya P, Srivastava DK on "Network intrusion detection in big dataset using Spark" in Procedia Comput Sci. 2018.*

The model proposed in the paper is both fast and efficient while being effective for detection Intrusion. The UNSW NB-15 dataset consisting of both small and large dataset was used in the performance evaluation of the model proposed in the paper. Different feature extraction and classification methods/algorithms were used were used to prepare the model. The use of CNN led to the discovery that the accuracy of small dataset was affected but it also decreased the time taken was also reduced. Whereas, the LDA increases the accuracy but it comes at the cost of increase in time taken to train the model using both small and large dataset. The most optimal solution was found to be the Decision Tree algorithm and LDA was found to be the better method for feature reduction. The Intrusion detection approach was found to be faster and more efficient using the random forest algorithm and LDA. Accuracy wise Random Tree algorithm was found to be better than other Algorithms. It was able to classify the data accurately as normal traffic and into various attacks. Feature reduction methods when used were found to increase the accuracy of the model. After running the above tests using Apache Spark it can be concluded that the Apache Spark method is better, faster and more efficient.

8) *Wang H, Xiao Y, Long Y on "Research of intrusion detection algorithm based on parallel SVM on Spark." in 7th annual IEEE (ICEIEC), 2017.*

SVM is a proven tool, powerful when it comes to classification and regression. The SVM models that have been proposed so far are: SMO (Sequential Minimal Optimization), libSVM, lightSVM and so on. While they have been demonstrated to be feasible to some extent, none of them are suitable to be directly used for handling data sets which are on the larger scale. As the size of the data sample gradually increases, there is a dramatic increase in the time and memory taken to train a model using the SVM algorithm. It is also a known fact that a single SVM Algorithm cannot effectively deal with data sets that or on the larger end in terms of size. To solve this problem of optimization for SVM to handle large amounts of data, a solution was reached to parallelize the algorithm roughly. This method narrows the data by choosing the scheme of divide and conquer. This paper proposes an implementation, which is, an amalgamation of the parallelized implementation of SVM and PCA dimensionality reduction based on Bagging on Apache spark. Optimization was achieved by eliminating the bottleneck i.e. difficulty in dealing with large data sets, but also improving the efficiency of high dimensional network information data.

9) *Gupta GP, Kulariya M on "A framework for fast and efficient cyber security network intrusion detection using Apache Spark" in Procedia Comput Sci. 2016.*

This paper advocates the use of Apache Spark, a tool used for processing big data for the processing of immense amount of network traffic data. They proposed a working model in which a popular feature selection algorithm is selected for selecting the most important features. Then a classification based intrusion detection method is used for fast and efficient detection of intrusion among the enormous amount of network traffic. They used two well-known feature selection algorithm, namely, correlation based feature selection and Chi-squared feature selection and five well known classification based intrusion detection methods, namely, Logistic regression, Support vector Machines, Random forest, Gradient Boosted Decision trees & Naive Bayes. A real time DARPA's KDD'99 data set is used to validate the proposed framework and performance comparison of classification based intrusion detection schemes are evaluated in terms of training time, prediction time, accuracy, sensitivity and specificity.

10) Ahmed M, Anwar A, Mahmood AN, Hu J on "A survey of network anomaly detection techniques." 2016.

The survey of literature suggested on this paper has labeled the ambiguity detection techniques in 4 principal categories. For every category, we defined the assumptions for segregating ordinary records times from anomalous. These assumptions will offer a tenet to evaluate the performance of the strategies while implemented in a selected domain. Compared to different surveys, this paper furnished a dialogue on community visitors dataset problems which can be of vast subject to the studies network within the location of community visitors analysis. It became located that for the safety of huge networks and huge IT ecosystems (i.e. cloud services), collaborative strategies are extraordinarily green which include numerous video display units that act as sensors and accumulate records. Due to the unavailability of implementations of collaborative strategies which includes CIDSs (Collaborative Intrusion Detection Systems), destiny studies efforts are important for sizeable quantitative assessment with contemporary community infrastructure.

11) M. Mazhar Rathore, Awais Ahmad, Anand Paul on "Real time intrusion detection system for ultra-high-speed big data environments".

The proposed model consists of a four-layered IDS machine. These four layers are 1) Capturing layer 2) filtration and load balancing layer 3) processing (Hadoop layer) 4) decision-making layer. In addition to the evaluation of DARPA datasets, a characteristic choice scheme is proposed that selects 9 parameters for class using (FSR) and (BER). Five primordial Machine Learning techniques are used to assess the proposed system. They are: J48, REPTree, random forest tree, conjunctive rule, support vector machine, and Naïve Bayes classifiers. It can be ascertained from the results that out of the different classifiers, REPTree and J48 are the most satisfactory classifiers in terms of accuracy and performance. The proposed machine structure is evaluated with respect to accuracy in terms of True positive (TP) and false positive (FP) measures, and with respect to performance in terms of processing time. It has more than 99 % TP and much less than 0.001 % FP on REPTree and J48. The proposed machine has a better accuracy than present IDSs with the functionality to work in real time in an ultra-high-speed big data environment.

12) Sung-Hwan Ahn, Nam-Uk Kim, Tai-Myoung Chung on "Big Data Analysis System Concept for Detecting Unknown Attacks".

Unknown attacks should be defended against with top priority but the technology to do so does not exist at present. They advocated a brand new version primarily based on large records evaluation strategies. The statistics that can be extracted from a whole lot of assets will aid in stumbling upon future attacks. They propose their version of the concept to be the idea of the future Advanced Persistent Threat(APT) detection and prevention system implementations. In this paper they proposed a model driven by big data techniques for reacting to formerly unknown Cyber threats and researched the deduction of sensible technologies. Further research is to be carried out by them in the following areas: Classification of data with the aid of using the context of intrusion detection, Implementation of data relation analysis technique and its abnormal behavior detection strategy, Quantitative and qualitative evaluation of proposed model and overall performance evaluation.

13) S. C. Y. Ng ,a and M. Bakhtiarib on "Advanced Persistent Threat Detection Based On Network Traffic Noise Pattern and Analysis".

This study was conducted via all common features of the Remote Administrative Tool named "NjRAT v0.7". The traffic from IP address of only the Attacker and Victim were filtered to file, avoiding any confusion with normal operating system's service traffic. Their test demonstrated in-depth what was going on in the monitoring scenario. Analyzing the patterns of communication that was recorded on a graph offered a clear picture of the common pattern as proof of an ongoing attack to alert the victim. The important concept of this studies was that each cyber-assault concerning Advanced Persistent Threat with or without zero-day vulnerabilities, whether the attack was recognized or unknown, would without a doubt generate traffic within the attacker's device and victim's host. The objective of their research was to analyze the traffic between the attacker and victim's device by developing a Virtual Environment. Traffic of the attack would depart proof in real time therefore the Virtual Environment would reveal the activity of the attacker. From the traffic that is generated, the existence of an attack could be discerned. The traffic data collection should primarily be at the victim's side due to the fact that the victim's device is the one obeying the attacker's command through RAT. This research was subject to time constraints, the proposed framework will make a contribution multiple areas such as the preparation of an antivirus, network intrusion detection and prevention and manage a framework for preventing Advanced Persistent Threat attack by exploiting zero-day vulnerabilities. The proposed framework was developed primarily based on studies that had been carried out during the period of this project. Therefore, its implementation by detection-based organizations is still pending to offer the real-world result.

14) *Wenying Feng,a, Qinglei Zhangb , Gongzhu Huc , Jimmy Xiangji Huang "Mining Network Data for Intrusion Detection through Combining SVM with Ant Colony".*

In this paper, they introduced a new machine-learning based data classification algorithm that is applied to network intrusion detection. The objective was to classify network activities (connection records as per the network log) as regular or abnormal as well as minimizing misclassification simultaneously. Distinct classification models have been developed for network intrusion detection in the past, each one of them having their own strengths and weaknesses. These include the most commonly used Support Vector Machine method and the Clustering based on Self-Organized Ant Colony Network. Their new technique takes advantage of the strengths of both SVM and CSOACN of them while also avoiding the weaknesses of both. KDD99 dataset which is the benchmark when it comes to network intrusion detection was used to evaluate their algorithm. They demonstrated through experiments that CSVAC (Combining Support Vectors with Ant Colony) outperforms SVM alone or CSOACN alone in terms of both classification rate and run-time efficiency.

15) *Kayacik HG, Zincir-Heywood AN, Heywood MI on "A feature relevance analysis on kdd99 intrusion detection datasets." in Proceedings of the third annual conference on privacy, security and trust, Citeseer 2005.*

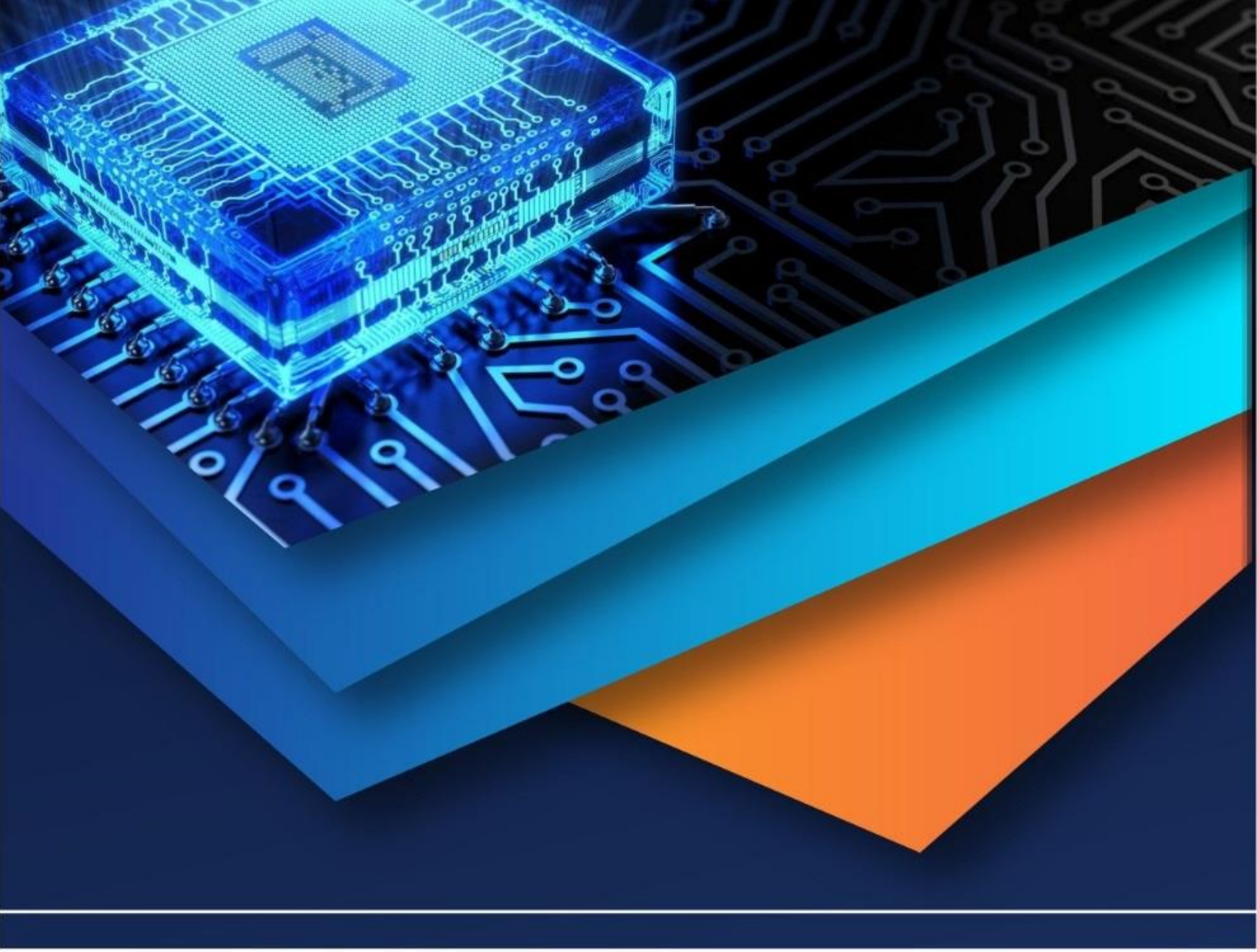
The researchers suggest that a characteristic relevance evaluation is carried out on KDD 99 training set, that is broadly used by the community. Feature relevance is expressed in terms of information gain, which gains weightage as the feature gets more discriminative. Information gain is calculated on binary classification for every feature ensuing in a separate information gain per class. This results in a feature relevance measure for all classes in training set. Recent research that made use of popular techniques, in terms of detection and false alarm rates, reported that U2R and R2L attacks are very tough to classify. This work analyzes the relevance of each feature to classification of the attack. Their results have discovered an association between certain features that make classification of three classes smurf, neptune and normal easier. These classes make up a high percentage of the training data, therefore making a machine-learning algorithm a viable approach to gain good result. It was also discovered that certain features do not contribute to the detection of intrusion showing that not all features are useful. "10% KDD" is the training data in the competition, therefore although test data indicates different traits than training data, evaluation of data marked for training shed light on the generalized performance of machine learning based IDS. Their future work involved extra steps to further improve the feature relevance and expand the evaluation to different datasets.

IV. CONCLUSIONS

From the survey of existing methodologies and the techniques used to design an effective Intrusion Detection System. It seems clear that there are quite a few challenges in handling big data like the KDDCup99 dataset and how it can be made more efficient. We came across numerous methods on creating an Intrusion detection system using apache spark, machine learning techniques like SVM, CNN, Naive Bayes, Decision trees Et al. Our job from here would be to try said methods and try to come up with a more competent and efficient solution which would help us move onto our Intrusion Prevention System in the next phase. In future work, can include extension of the existing model to a multi-classes model that could detect multiple types of attack instead of just binary classification. We would also be working on making the model competent enough to detect zero day vulnerabilities and newer cyber-attacks which might still be unknown to us. With the technological advancements we will also have new and different types of Cyber-attacks which would a much more complex to detect and prevent, hence there is always room for improvement in the model to make it better and more efficient using the new data that comes along with and training the model with it.

REFERENCES

[1] E. M. Tchakoucht TA, "Building a fast intrusion detection system for high-speed-networks: probe and DoS detection," 2018.
[2] S. A, "Survey on intrusion detection system using data mining techniques," 2017.
[3] S. A. Ferhat K, " Big Data: controlling fraud by using machine learning libraries on Spark," 2018.
[4] L. Y, "Toward scalable internet traffic measurement and analysis with hadoop," 2013.
[5] P. K, "Intrusion detection system based on decision tree over Big Data in fog environment," 2018.
[6] M. Y. Manzoor MA, "Real-time support vector machine based network intrusion detection system using Apache Storm," 2016.
[7] S. D. Dahiya P, "Network intrusion detection in big dataset using Spark," 2018.
[8] X. Y. L. Y. Wang H, "Research of intrusion detection algorithm based on parallel SVM on Spark," 2017.
[9] K. M. Gupta GP, "A framework for fast and efficient cyber security network intrusion detection using Apache Spark".
[10] A. A. M. A. H. J. Ahmed M, "A survey of network anomaly detection techniques," 2016.
[11] A. A. A. P. M. Mazhar Rathore, "Real time intrusion detection system for ultra-high-speed big data environments.".
[12] N.-U. K. T.-M. C. Sung-Hwan Ahn, "Big Data Analysis System Concept for Detecting Unknown Attacks".
[13] S. C. Y. N. a. M. Bakhtiarib, "Advanced Persistent Threat Detection Based On Network Traffic Noise Pattern and Analysis.".
[14] a. Q. Z. ,. G. H. ,. J. X. H. Wenying Feng, "Mining Network Data for Intrusion Detection through Combining SVM with Ant Colony".
[15] Z.-H. A. H. M. Kayacik HG, " Selecting features for intrusion detection: a feature relevance analysis on kdd99 intrusion detection datasets.," 2005.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY