International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 08 | Aug 2020 www.irjet.net p-ISSN: 2395-0072

Anomaly-based Intrusion Detection using Machine Learning Algorithms - A Review Paper

Ambreen Sabha

M.Tech. Student, Department of Computer Science and IT, University of Jammu, J&K, India

---

Abstract - *An intrusion is termed as an activity that attempts to compromise the confidentiality or availability of a resource. An intrusion detection system i.e. IDS is the most important field of network security, that monitors the state of software and hardware running in the network. In the past few years, Intrusion detection using machine learning technique has captured the attention of most of the researchers, and every researcher proposes a different algorithm for the distinct features used in the dataset. KDD-Cup99 intrusion detection dataset plays a vital role in the network intrusion detection system and NSL-KDD is an updated or revised version of KDD-Cup99. The dataset which is mostly used by the researchers working in the field of intrusion detection is KDD-Cup99. This paper presents an overview of various IDS and also the detailed analyses of various machine learning techniques and datasets used for improving IDS.*

Key Words: **IDS, Information security, Network based Intrusion detection system, NSL-KDD, Network traffic.**

1. INTRODUCTION

Internet has increasing influences on modern life, making cybersecurity an important field of research for the researchers. Cybersecurity techniques mainly include anti-virus software, firewalls, and intrusion detection systems i.e. IDS. Nowadays computer attack has become very common with the enormous growth of computer networks and usage of a large number of applications. One of the most important fields on network security is Intrusion Detection System.

Intrusion detection is the problem of identifying unauthorized use and abuse of computer systems by both system insiders and external intruders and it is the process of detecting malicious patterns in the large data sets. IDS is a device or software which monitors the network for malicious Activity [3]. **IDS** is one component of network security that protects data and information by monitoring the data packets in the network traffic to detect an intrusion. It monitors the network to check malicious activities. And reports events that do not meet the security criteria to the network administrator.

Intrusion detection systems are classified into two different categories i.e. HIDS and NIDS. Host-based IDS runs in any individual host or device. HIDS monitor only the inbound and outbound packets in the network traffic and when suspicious or harmful activities are identified it sends the alert to the administrator. NIDS analyses the passing traffic on the whole subnet and matches the traffic into the known traffic library. It uses techniques like packet sniffing, analyses the network data, and discovers the unauthorized access to computer networks [7]. A typical NIDS makes use of Signature detection also known as misuse detection and Anomaly detection.

The **Signature-based IDS** is used to identify attacks in a form of signature or pattern and it uses a database of known attack signatures that are developed by any experts or intrusion analysts. The Signature detection monitors packets in the network and compares them to the known signatures in the database. If there is a match with database entries, the IDS generate an alert message [5]. The major restriction of these signature-based IDS is that they can only detect the intrusions whose attack patterns are already stored in the database, but these systems cannot identify new and novel attacks. The attacks whose patterns are not present in the database cannot be detected. In contrast to signature-based IDS, Anomaly-based IDS look for the kinds of unknown attacks that signature-based IDS, finds hard to detect.

Anomaly-based IDS monitors the network traffic and sends an alert to the system administrator on the detection of anomalous behavior. Many machine learning techniques are being used to detect the anomalous behavior on a host or network, they function on the assumption that attacks are different from the "normal" activity. An Anomaly detection system monitors the behavior of a system or the network and flags significant deviations from the normal activity as an anomaly. Anomaly detection is used for identifying attacks in computer networks and malicious activities in a computer system.

2. DATASET

2.1 KDD-Cup99 Dataset

The KDD'99 has been the most widely used data set for the evaluation of anomaly detection method [13]. The dataset was used in the 3rd International Knowledge Discovery and Data Mining Tools Competition for building a intrusion detector i.e. a predictive model capable of distinguishing between "bad" connections, called intrusions or attacks, and "good" normal connections [8]. As a result of this competition, a mass amount of internet traffic records was collected and bundled into a data set called the KDD-Cup99 dataset.

KDD-Cup99 Dataset consists of 41 attributes and one label or class. There are 5 main classes in KDD-99 dataset can be categorized as (one normal class and four main intrusion classes i.e. probe, DOS, U2R, and R2L [3]. The KDD dataset mainly consists of four classes of attacks:

Denial of Service (DOS): DOS attack mainly floods the server, system, or network by superfluous packets. Due to that, the system's buffer becomes full and unable to respond and return to the valid request. This attack makes a network or system to shut down and it became inaccessible to the intended users. The main objective of this attack is to consume memory or other resources of the target.

Remote to Local (R2L): R2L is defined as unauthorized access in which the attacker intrudes into a remote machine and gains local access of the victim machine. Attacker uses existing vulnerabilities to access target account e.g., password guessing, IMAP, and FTP write [11].

User to Root (U2R): In this attack, the attacker got access to the target system using valid user authentication, and is able to exploit some vulnerability or issues to gain root access to the system [13]. The intruder begins with the access of a normal user account and then becomes a root-user by exploiting various vulnerabilities of the system. Most common exploits of U2R attacks are – e.g., buffer overflow attacks, load-module and Perl.

Probing: Probe is defined as an attack that scans a network to gather information or to find known vulnerabilities, Intruder having a map of machines and services that are available on a network can use the information to look for exploits [19]. The attacker accesses the security frame and obtain relevant data from the target system– e.g., port scanning.

2.2 NSL-KDD Dataset.

NSL-KDD intrusion detection data set is anupdated version of KDD cup'99 data set [12]. The reason for using NSL-KDD dataset is that the KDD'99 data set has a large number of redundant or duplicate records in the training and testing set. For binary classification, the NSLKDD classifies the network traffic into two main classes i.e. normal and anomaly [9]. NSL-KDD dataset is developed to resolved the data problems in KDD-Cup99 dataset [7]. To solve the issues in KDD-99 cup dataset, researchers proposed a new dataset called NSL-KDD which consists of only selected records from the complete KDD'99 and the above dataset does not suffer from any of the issues. The merits of using the NSL-KDD dataset are:

1. No redundant or duplicate records in the training set, due to this the classifier will not produce any biased result.

2. No duplicate records in the testing set which leads to better reduction rates [11].

3. The NSL-KDD has sufficient number of records in testing and training phase. It can help to perform experiments appropriately.

3. LITERATURE REVIEW

S. Omar et al. [14] published an overview of "Machine Learning Techniques for Anomaly Detection", the supervised and unsupervised methods were applied for the problem of anomaly detection. The experimental results showed that the supervised learning methods significantly performs better than the unsupervised ones, if the test data contains no unknown attacks. This paper shows the pros and cons of all the supervised and unsupervised machine learning algorithms. Among the supervised methods, the best performance is achieved by the SVM, multi-layer perceptron and the rule-based methods.

Shilpashree. S et al. [2] published a paper on "Decision Tree: A Machine Learning for Intrusion Detection" that deals with the performance of intrusion detection system by applying machine learning techniques based on decision trees. In this work, a system was built which is based on the decision tree and different strategies were also contrasted with this approach. The Bayesian three modes were analyzed for different size of datasets. The Multinomial naïve Bayes gets the least computation time then Bernoulli naïve Bayes, and Gaussian naive Bayes is the last one among all the test cases. Information gathering is obtained through some capturing devices, such as Libdump, TCPdump and Wireshark. The execution time taken by the classifier to build the model is analyzed and the accuracy is done.

Kajal Rai et al. [9] published a paper on "Decision Tree Based Algorithm for Intrusion Detection". In which the decision tree algorithm was developed based on C4.5 decision tree approach. Feature selection and split value are the important issues for the constructing of a decision tree. In this paper The Decision Tree Split (DTS) algorithm was designed to address these two issues. The DTS algorithm was implemented using tools WEKA and MATLAB. The proposed algorithm was compared with the existing tree algorithms such as Classification and Regression Tree (CART), C4.5, and AD Tree. The analysis was based on different parameters such as how many seconds the classifier took for the construction of the model, false positive rate (FPR), true positive rate (TPR), and accuracy.

M. Gupta et al. [4] published a paper on "Intrusion Detection Using Decision Tree Based Data Mining Technique", in which the J48 decision tree algorithm was used. Machine learning tool WEKA was used for the implementation and it also analyses the performance of datasets. In the WEKA tool, J48 is an open source Java implementation of the C4.5 algorithm. This J48 algorithm gave higher accuracy over Naïve Bayes and SVM. This algorithm gains equilibrium of flexibility and it shows 99.73% of accuracy.

M. Tabash et al. [1] published a paper on "Intrusion Detection Model Using Naive Bayes and Deep Learning Technique", in which the hybrid model was developed to explore any penetrations inside the network. This model divides into two basic stages. The first stage includes the Feature selection

technique i.e. Genetic Algorithm (GA) which depends on a process of Discretization and dimensionality reduction after that combining the Naïve Bayes classifier (NB) and Decision Table (DT) at the end of the first stage. The second stage depends on the output of the first stage and reclassified with multilayer perceptron using Deep Learning algorithm i.e. Stochastic Gradient Descent (SGD). so, this proposed hybrid model based on deep learning technology improves the detection rate, accuracy and reduces the false alarms. In order to improve the performance, the comparison is done between the proposed model and the previous conventional model. The experimental result showed that the proposed hybrid model had classification accuracy of 99.9325, the detection rate is 99.9738 and 0.00093 is the false alarm rate.

A. Nur Cahyo et al. [3] presented a paper on "Performance Comparison of Intrusion Detection System based on Anomaly Detection using Artificial Neural Network and Support Vector Machine". This study uses all the features of the dataset and presented a comparison between the ANN and SVM using the anomaly-based IDS and Pre-processing was performed on the datasets for the normalization and scaling attribute. Artificial Neural Network obtained high accuracy in all categories compared to that of SVM. The detection rate of DoS is 92.20%, probe detection rate is 90.60%, R2L rate is 89%, and that of U2R is 90.80%, the result showed that ANN shows better performance than SVM in attack detection.

A. Abd Ali Hadi et al. [12] published a paper on "Performance Analysis of Big Data Intrusion Detection System Over Random Forest Algorithm". In this paper the Random forest algorithm was applied to classify the network data. To increase the accuracy of Random forest algorithm the information gain method was used as features selection method, the 13 most significant features were generated from the original 41 features, and the proposed model used these significant features. It is observed that the accuracy was increased and it reduced the execution time of the model. The various performance measures were employed to test the proposed model. The accuracy of model is 99.33%, and it performs better than the existing classifiers. The Random Forest algorithm was implemented using tools WEKA and MATLAB.

O. P. Akomolafe et al. [16] published a paper on "An Improved KNN Classifier for Anomaly Intrusion Detection System Using Cluster Optimization". This paper presented an improved or modified KNN classifier using clustering optimization for an anomaly-based IDS, which would provide an effective classification scheme for existing anomaly intrusion detection system and this concluded that the nearest neighbour is a good classifier for anomaly intrusion detection system but with addition of the cluster optimizer. The performance of the improved KNN classifier was compared with the existing KNN classifier, and the experimental results showed that the existing classifier had an efficiency of 98.7% for correctly classified instance and 0.2395% for the incorrectly classified instances while the newly developed classifier had an efficiency of 99.6% for the correctly classified instances and 0.3222% for the incorrectly classified instances. Machine learning tool WEKA was used for the implementation.

R. Wankhede et al. [5] published a review paper on "A Review on Intrusion Detection System Using Classification Technique". This paper presented an overview of various IDS and detailed analyses of various techniques used for improving IDS, KDD-Cup99 intrusion detection dataset was used and it plays a vital role in intrusion detection system and is mostly used by the researchers working in the field of intrusion detection. This paper showed that Intrusion Detection System i.e. IDS still needs a lot of improvements in case of the detection rate of the attack and as well as reducing the error rate of the attacks. The IDS should be able to detect the known as well as the unknown attacks by improving their strategy for intrusion detection.

S. Taruna R et al. [8] published a paper on "Enhanced Naïve Bayes Algorithm for Intrusion Detection in Data Mining". This paper proposed a new method of Naïve Bayes Algorithm i.e. Enhanced Naïve Bayes which find the detection rate and false positive rate of given data very efficiently. The results showed that the Naïve Bayes classifier model more efficiently detect the network intrusions, compared to the neural network-based classification techniques and it also improved the detection rates as well as reduces the false positive rates for different types of network intrusions. The proposed algorithm was tested on KDD-Cup99 network intrusion detection dataset that shows that the detection rate for 4 attack classes is maximized in KDD-Cup99 dataset and also minimized the false positives at an acceptable level.

S. Revathi et al. [10] published a paper on "A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection". This paper presented a detailed analyses on various dataset used to improve the accuracy of system and to reduce false positive rate based on DAPRA 98 dataset and later used the updated version as KDD-cup99 which shows some statistical issues and it degrades the evaluation and affects the performance of anomaly detection, which leads to the replacement of KDD99 to NSL-KDD dataset. This paper mainly focused on NSL- KDD dataset which contains only selected records, and those records provides a good analysis on various machine learning techniques for intrusion detection.

Balogun et al. [13] published a paper on "Anomaly Intrusion Detection Using an Hybrid of Decision Tree And K-Nearest Neighbor". This paper proposed a hybrid classification algorithm based on decision tree and K-Nearest neighbour. The data set was firstly passed through the decision tree to generate the node information and this information is determined according to the rules generated by the decision tree. The original set of attributes along with the node information is passed through the KNN to obtain the final output. The evaluation is performed using a 10-fold cross validation technique on the individual base classifiers i.e. decision tree and KNN and the proposed hybrid classifier (DT-KNN) using the KDD-Cup99 dataset. The Machine learning

tool WEKA was used for the implementation. Results showed that the hybrid classifier (DT-KNN) gives the best result in terms of accuracy and the efficiency was compared with the individual base classifiers i.e. decision tree and KNN.

TABLE- 1: REVIEW OF LITERATURE

Author	Research Statement	Methodology used	Dataset	year	Performance Analysis
K. Rai *et. al.* [9]	Decision Tree Based Algorithm for Intrusion Detection.	Decision Tree Split (DTS) based on C4.5 decision tree Algorithm.	NSL-KDD	2016	Resulted work increased the effectiveness, accuracy and detection rate of the classifier.
M. Gupta *et. al.* [4]	Intrusion Detection Using Decision Tree Based Data Mining Technique.	J48 decision tree algorithm	ORNL	2016	J48 algorithm gave 99.73% of classification accuracy over Naïve Bayes and SVM.
M. Tabash *et. al.* [1]	Intrusion Detection Model Using Naive Bayes and Deep Learning Technique.	Naïve Bayes classifier, Decision table and stochastic gradient descent.	NSL- KDD	2018	The hybrid model gave 99.9325 of classification accuracy, detection rate is 99.9738 and 0.00093 of false alarm.
A.N Cahyo *et. al.* [3]	Performance Comparison of Anomaly based Intrusion Detection System using ANN and SVM.	Artificial Neural Network (ANN) and Support Vector Machine (SVM).	KDD-Cup99	2017	ANN shows better performance than SVM in attack detection. DoS 92.20%, Probe 90.60%, R2L 89%, and U2R 90.80%
A. Abd Ali Hadi [12]	Performance Analysis of Big Data Intrusion Detection System Over Random Forest Algorithm.	Random forest algorithm	NSL- KDD	2018	Accuracy of model is 99.33%, and performs better than the naïve Bayes, SVM and KNN.
O. P. Akomolafe *et al.* [16]	An Improved KNN Classifier for Anomaly Intrusion Detection System Using Cluster Optimization	modified KNN classifier	NSL- KDD	2017	Results showed that existing KNN had 98.7% efficiency of correctly classified instance and 0.2395% for the incorrectly classified instances while the Modified KNN had a 99.6% efficiency for the correctly classified instances and 0.3222% for the incorrectly classified instances.
Balogun *et al.* [13]	Anomaly Intrusion Detection Using an Hybrid of Decision Tree And K-Nearest Neighbour	Hybrid of decision tree and KNN (DT-KNN)	KDD-Cup99	2015	Hybrid classifier (DT-KNN) gives the best result in terms of accuracy and the efficiency was compared with the individual base classifiers (decision tree and KNN)

4. CONCLUSION

This paper describes that Intrusion detection is the most important feature of security and is used for detecting fraud and unauthorized access. In this review paper, we have introduced an overview of different machine learning techniques for the Intrusion Detection System i.e. IDS and different detection methodologies. Each technique has its merits and demerits i.e. limitations, so that precautions should be taken about the selection of the different machine learning algorithms. And we also analyzed the network intrusion datasets i.e. KDD'99 dataset and its updated version as NSL-KDD, and it shows that NSL-KDD solves some of the issues of KDDcup99 dataset. The analysis shows that NSL-KDD dataset is very much used by the researchers in the field

of network intrusion for comparing different intrusion models using machine learning techniques.

REFERENCES

[1] M. Tabash, M. A. Allah, B. Tawfik **"Intrusion Detection Model Using Naive Bayes and Deep Learning Technique"** International Arab Journal of Information Technology, 2018.

[2] Shilpashree. S, S. C. Lingareddy, N. G. Bhat, S. Kumar G **"Decision Tree: A Machine Learning for Intrusion Detection"** International Journal of Innovative Technology and Exploring Engineering, Vol. 8, Issue-6S4, pp. 1126-1130, April 2019.

[3] A. N. Cahyo, R. Hidayat, D. Adhipta **"Performance Comparison of Intrusion Detection System based on Anomaly Detection using Artificial Neural Network and Support Vector Machine"** Advances of Science and Technology for Society AIP Conf. Proc. 1755, pp.070011-1–070011-7 ,2017.

[4] M. Gupta, J. Shriwas, S. Farzana **"Intrusion Detection Using Decision Tree Based Data Mining Technique"** International Journal for Research in Applied Science & Engineering Technology**,** Vol. 4 Issue 7, pp. 24-28, July 2016.

[5] R. Wankhede, V. Chole, S. Kolte **"A Review on Intrusion Detection System Using Classification Technique"** International Journal of Advanced Computational Engineering and Networking, Vol.3, Issue 12, pp. 62-65, Dec. 2015.

[6] DNSstuff. **Intrusion detection system.** https://www.dnsstuff.com/intrusion-detection system

[7] Dzone. **Machine learning algorithm for Intrusion detection system.** https://dzone.com/articles/evaluation-of-machine-learning-algorithms-for-intrusion

[8] S. Taruna R., S. Hiranwal **"Enhanced Naïve Bayes Algorithm for Intrusion Detection in Data Mining"** International Journal of Computer Science and Information Technologies Vol. 4 (6), pp. 960-962, 2013.

[9] K. Rai, M. S. Devi, A. Guleria **"Decision Tree Based Algorithm for Intrusion Detection"** International Journal of Advanced Networking and Applications Vol. 07, Issue- 04 pp. 2828-2834, 2016.

[10] S. Revathi, A. Malathi **"A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection"** International Journal of Engineering Research & Technology, Vol.2, Issue 12, pp. 1848- 1853, Dec. 2013.

[11] R. R. Devi, M. Abualkibash **"Intrusion Detection System Classification using different Machine Learning Algorithms on KDD-99 and NSL-KDD datasets - A Review Paper"** International Journal of Computer Science & Information Technology, Vol.11, Issue 3, pp. 65-80, June 2019.

[12] A. A. Ali Hadi **"Performance Analysis of Big Data Intrusion Detection System Over Random Forest Algorithm"**, International Journal of Applied Engineering Research, Vol.13, Issue 2, pp. 1520-1527, 2018.

[13] Balogun, A. O., Jimoh, R. G. **"Anomaly Intrusion Detection Using an Hybrid of Decision Tree And K-Nearest Neighbor"**, A Multidisciplinary Journal Publication of the Faculty of Science, Vol. 2, pp. 67-74, 2015.

[14] S. Omar, A. Ngadi, H. H. Jebur **"Machine Learning Techniques for Anomaly Detection"** International Journal of Computer Applications, Vol. 79, No.2, pp. 33-37, 2013.

[15] D.P. Gaikwad, S. Jagtap, K. Thakare, V. Budhawant **"Anomaly Based Intrusion Detection System Using Artificial Neural Network and Fuzzy Clustering"** International Journal of Engineering Research & Technology, Vol. 1, Issue.9, pp. 1-6, 2012.

[16] O. P. Akomolafe, A. I. Adegboyega **"An Improved KNN Classifier for Anomaly Intrusion Detection System Using Cluster Optimization"** International Journal of Computer Science and Telecommunications Vol. 8, Issue 2, 2017.

[17] U. Kumari, U. Soni "**A Review of Intrusion Detection using Anomaly based Detection"** IEEE ICCES, pp.824-826, 2017.

[18] V. Jyothsna, V. V. R. Prasad "**A Review of Anomaly based Intrusion Detection Systems"** International Journal of Computer Applications, Vol. 28, No.7, pp. 26-35, Sep. 2011.

[19] M. Kumar, M. Hanumanthappa, T.V. Suresh Kumar **"Intrusion Detection System Using Decision Tree Algorithm"** IEEE, pp.629-634, 2012.

[20] V. D. Mane, A. Sayar, S. Pawar "**Anomaly Intrusion Detection System Using Neural Network"** International Journal of Computer Science and Mobile Computing Vol.2, Issue.8, pp. 76-81, 2013.