International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-7 Issue-4s2, December 2019

An Effective Way of Cloud Intrusion Detection System using Decision tree, Support Vector Machine and Naïve Bayes Algorithm

T. Nathiya, G. Suseendran

Abstract: ***Cloud computing is a vast area, use the resources with cost-effectively. The service provider is to share the resources anywhere at any time. But the network is the most vital to accessing data in the cloud. The cloud malicious takes advantages while using the cloud network. Intrusion Detection System (IDS) is monitoring the network and notifies attacks. In Intrusion Detection System, anomaly technique is most important. Whenever Virtual Machine is created, IDS track the known and unknown data's. If any unknown data found, Intrusion Detection System detects the data using anomaly classification algorithm and send the report to admin. This paper proposes we are using support vector machine (SVM), Naive Bayes, and decision tree (J48) algorithms for predicting unwanted data's. In these algorithms are help us to overcome the high false alarm rate. Our proposed work implemented part using the WEKA tool to give a statistical report, which gives a better outcome in little calculation time.***

Keywords: SVM, Naive Bayes, Decision Tree (J48), NSL-KDD dataset, H-IDS.

I.INTRODUCTION

Cloud computing is present at the remote location and it's providing services over the network. The user can be configured, accessing and manipulating the application such as data storage, infrastructure, server and application[1]. The user can access anything as services such as infrastructure, platform, and software anywhere in the world from the cloud through the internet. The Cloud computing communication with two ends. In see the figure 1 the front end must communicate with user and cloud. When the user need resources may like hardware/ software to execute for maintaining the database, developing the applications and its deliver the services via the network. Another end must communicate with cloud and third party. The virtual machine monitor (i.e. IDS) is run on multiple virtual machines on the physical layer. The third party has fully maintaining web and application server, database server and developing tools [2]. The government, business sector, a variety of academic, medical and lot of organizations are increasingly using Information and Technology (IT) in cloud computing. But they need to bring lots of security. Because lots of network attack intrudes in the cloud. The traditional attacks are IP spoofing, DDOS, User to Port, Port Scanning etc. An IDS haven a new efficient solution of the traditional network for securing packets. The role of IDS is observed the network and to predict the malicious activity and report to the cloud administrator. If an intrusion has detected, The IDS is creating issues alert signal to continuously watch about this event. Whether this alert is true positive or false alarms. The cloud network IDS have placed at cloud server and administrated managed by service provider.IDS handling large scale of the computing system, automated, scalability, and synchronization of IDS[3][4][5].

The network intrusion detection system must choose the feature and reduced the number of features can be easily extracted out of high speed of data. Because the local area network forwarding the packets with one gigabit per second depends upon the hard disk speeds. However hard disk speeds much slower. The minimal framework size is 64 bytes. So, one to 14.8 million frames can be transferred per second. During this transaction, the network is monitoring the data that's a major challenge in cloud computing. The most critical challenge is the real-time detection of data's. [6].

This paper aims to predictive data using four anomaly-based algorithms for making an effective system for detecting intrusion in cloud computing. These Abnormality based techniques are discussed later in this paper. To make the feature selection dataset attributes using various tools.

The balance of this paper is going to discuss on next section related work and briefly describe the Anomaly based techniques. And continuing section is Algorithm classification and then section dataset and preprocessing. And the next section is an experimental result and concludes our paper and informs about future work.

II.ANOMALY BASED TECHNIQUES AND RELATED WORK

Hybrid network intrusion detection system is designed to install in the virtual network at each host layer. H-NIDS has monitored the network traffic and reports into the higher layer and it has using both signature/ anomaly based techniques[7]. The misuse technique has identified only well-known attacks from signature database. Using snort rule using fast multi- pattern matching algorithm to detect the attacks. But anomaly-based technique helps in unknown attacks.

Manuscript published on 30 December 2019.
*Correspondence Author(s)
T. Nathiya, Ph.D. Research Scholar, Department of Computer Science, School of Computing Science, Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai, Tamil Nadu, India.
G. Suseendran, Department of Information Technology, School of Computing Science, Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai, Tamil Nadu, India.

Data mining, statistical modeling, and machine learning techniques are used in anomaly based. The last paper I had to create security framework, we are included in the classification of algorithms such as decision tree, SVM, and Naïve Bayes. These techniques are applied in anomaly based IDS and it provides better accuracy and confidentiality, low communication cost and low false alerts[8][9]. Md. Al Mehedi Hasan et al.[10] proposed a two classification model, one is Support Vector Machine (SVM) and another one Random Forest algorithm for classifying attacks. The 90% of the dataset used for training and 10% dataset used for testing which is not sufficient to verify the accuracy. It took low execution time but the accuracy of detection is not expected rate. So it requires more test case results. Nabila Farnaaz et al.[11] proposed Random forest classifier in IDS. And they are compared to other traditional attacks. Using NSL-KDD dataset to evaluate the performance by using random forest algorithm to detect attacks like DOS, probe, U2R, and R2L. But they are improving the accuracy of the classifier in feature selection measure. G.V. Nadiammai et al.[12] proposed data mining concept is integrated with IDS and it's to identify the related data, hidden data with less execution time. They are used various algorithms for classification using KDD dataset. The proposed algorithm had produced good accuracy and false alarm rate. But two issues such as lack of user information and these techniques have not achieved an automatic intrusion detection system. Xueyan Jing et al.[13] proposed the Fuzzy C-Means (FCM) algorithm employed to clustering centers of training dataset and K-Nearest Neighbors algorithm to combine with dem unknown attacks. But they need more training and testing results.

Rahimeh Rouhi et al.[14] proposed anomaly detect pster shafer theory. These proposed algorithms used in KDD'99 datasets. The performance of result was effective to detect ion techniques applied in feature selection from KDD Cup 99 dataset. These paper used a feed-forward neural network was trained to predict the normal/ attack packets in the dataset. But these papers need more training and testing datasets. Opeyemi Osanaiye et al.[15] proposed a decision tree classification algorithm to detect the DDoS attacks. This paper proposed feature selection methods. But these paper not used the confusion matrix values. The accuracy and detection rate are not mentioned. How to work the decision tree classifications are not mentioned. Ozge cephelli et al.[16] proposed detection system adopted the network with traffic packets along sensitivity parameters. The proposed Hybrid- IDS to detecting the DDoS attack. But the result has decreased the performance. They need training performance. There are used a limited number of DARPA 2000 datasets. The proposed model was not clear. They are used penetration testing tool to get the commercial bank detailed dataset.

III.CLASSIFICATION TECHNIQUES

The IDS are classified into two techniques, one is signature based and another one is anomaly-based techniques. The signature based, we are using snort rule to detect the known attack and anomaly based, and we are using different classification such as naïve Bayesian, decision tree, SVM to detect an unknown attack in anomaly detection[17].

A. Snort

Observation of real time traffic is very difficult to detect the intrusion while heavy load. It gives the solution of network intrusion detection. Snort rule is an extremely flexible rule and it's easy to modify the nothing like commercial NIDS. Snort can be running four approaches (sniffer, packet logger, IDS, and IPS)[16][18]. Snort rule, if the user to write own rule for incoming and outgoing network packets and its combing two parts "The Header" and "The Options" segment. When packets must meet the threshold condition that only need to follow the snort rule.

B. Naive Bayesian Classification

It is one of the supervised learning algorithm as well as a statistical method of classification. The learning algorithm produce the function to predictions of the output values. The system is providing the targets of new input values after training data. Given the Bayesian algorithm is representing a class variable and the set of attributes are$h_1, h_2, \dots, h_n$.

$$p(g/h_1, h_2, \dots, h_n) = \frac{p(h_1, h_2, \dots, h_n/g)p(g)}{\mathrm{p}(h_1, h_2, \dots, h_n)} \quad (1)$$

For all i = 1, 2…, n, it

becomes, $p\left(\frac{hi}{g}\right)$ (2)

Where $p(h_1, h_2, \dots, h-1, hi+1, \dots hn)$

$$p(g/h_1, h_2, \dots, h_n) = \frac{\mathrm{p(g)} \prod_{i=1}^{n} p(hi/g)}{p(h_1, h_2, \dots, h_n)} \quad (3)$$

The classification equation as:

$$P(g/h) = p(g/h_1, h_2, \dots, h_n) \alpha P(g) \prod_{i=1}^{n} P(h_i/g)$$

$$= argmax\{P(h_1/g)P(g), P(h_2/g)P(g), \dots\dots, P(h_n/g)P(g)\} \quad (4)$$

$P(h/g)$ is a probability of g given h. $P(g)$is an prior probability of hypothesis g. $P(h)$ is an prior probability of training data h. $P(g/h)$ is an probability of g given h. In this**eq**.4 classification algorithm to help of improving the speed and accuracy of IDS[19].

C. Decision Tree Classifier

This is a family of the supervised learning algorithm. The Decision tree rules are easy to understand the user and using knowledge system such as Weka tool. C4.5 is termed as (J48 in Weka software). The main motive of using their decision tree rule is to create the training model and which is predicted the class value. Here the information gain ratio as an amount to choose the splitting features. Decision tree is classified into tree structure, the tree contains decision node and leaf node. Decision node: it is root node, each internal node corresponding to an attribute, Leaf node: corresponding to a class values. The windows consist of various classifiers like bays, function, Meta and tree. The entropy of an attribute E is computed as eq.5:

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

$$entropy(E) = -\sum_{i=1}^{n} p(E,i)\log(p(E,i)) \qquad (5)$$

Let A is the total no of intrusion classes in the given dataset $p(E,i)$ gives the ratio of instance in E and these are assigned to i^{th} class.

The Information Gain of the dataset G is calculated as eq.6:

$$gain(E,G)entropy(E)\sum_{m\varepsilon Values(G_E)} \frac{|G_{E,m}|Entropy(Em)}{|G_{E,m}|} \qquad (6)$$

Then, eq.7, we prepare the gain ratio of an attributes E is given by,

$$Gainratio(E,G) = \frac{\text{gain(E, G)}}{\text{splitinfo(E, G)}} \qquad (7)$$

Whereas $splitinfo(E,G)$ is calculated as,

$$splitinfo(E,G) = -\sum_{m\varepsilon Values(G_E)} \frac{|G_{E,m}|}{|G_E|} log \frac{|G_{E,m}|}{|G_E|} \qquad (8)$$

In **eq.**8, we are select the best split node to select the feature with maximum gain ratio. Here, this one reduced the computational complication[20].

D. Support vector machine Classifier

The SVM learning algorithm is run for classification and regression. But it is mainly used in classification problems. The process of SVM works in two class with a hyperplane. The classification is complete using hyperplane which training data created by the maximum margin. [10][21]

IV.PROPOSED APPROACH

In this section, we fully discussed about the network intrusion detector using machine learning algorithm. See figure 1 detailed explain our proposed work.

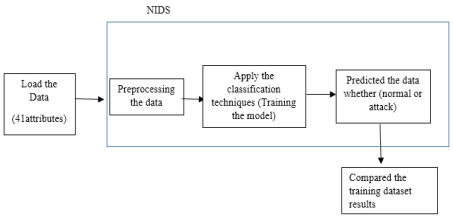

Figure 1: Sequence of Proposed approach.

The network intrusion detector is detecting the network intrusion. Some stages the NIDS is confused whether network packets are normal or abnormal and that critical situation, we are using machine learning techniques to classify the normal and abnormal packets.

Our proposed procedures is described below

Step 1: Load the intrusion dataset, which is containing 41 attributes features

Step 2: Preprocessing the data, which is reduced the irrelevant and redundant the data

Step 3: Apply the machine learning classification techniques (SVM, naïve bayes and decision tree)

Step 4: The classification techniques used to build the model (trained the model)

Step 5: Predicted the data, whether the data is normal or abnormal

Step 6: Finally compared the three classification techniques and its performance results.

V. DATASET DESCRIPTION

Herewith we are using NSL KDD dataset, it's advanced of KDD dataset. The NSL KDD dataset attributes can be used to detect the attacks like DOS, Probe, R2L, U2R etc. Also, it is mainly used for abnormal detection.

The advantages of NSL KDD data set are listed one by one, first, In the training set is not included the redundant records, so the classification will not produce a partial result.

And next one is processing the duplicate records along with testing dataset. NSL-KDD have produced better reduction rates[22]. Each record has 42 attributes contains data and the five various classes of the network.

One is original class and another four is attack classes and these attack classes are DOS, Probe, R2L, and U2R. Table 1, shows the major types of attack in both training and testing dataset[23].

Table 1: Display the Attack Classes and its Type

Attack Classes	Type of Network Intrusion classification
Dos = 1	Back = 1, land = 1, Neptune = 1, pod = 1, Smurf = 1, teardrop = 1
Probe = 2	Ipsweep = 2, nmap = 2, Portsweep = 2, satan = 2
R2L = 3	ftp_write = 3, guess_passwd = 3, imap = 3, Multihop = 3, Phf = 3, spy = 3, Warezclient = 3, warezmaster = 3
U2R = 4	Buffer_overflow = 4, Loadmodule = 4, Perl = 4, rootkit = 4
Normal = 0	Normal = 0

VI. EXPERIMENTAL RESULT AND ANALYSIS

A. Experimental Setup

For experimental setup, we have using NSL-KDD intrusion dataset and computerized data analyzing tool WEKA is used to perform the classification testing. Because the Weka tool is data mining process. It contains the clustering, pre-processing, classification, regression and feature selection models. It's working on windows. Only 20% NSL KDD dataset are charity to perform the classification. The presentation of the classifier is valuated with help of altered parameter like true detection rate, false detection rate, accuracy and execution time

B. Processing, Feature selection and Classification

The dataset can be classified initially preprocessing and that range is 0 to 1 (i.e. the researcher choose this level equal to 0.01, 0.05, or 0.10). The feature selection is nearly 41 available in the dataset. The SVM, Naïve Bayes and decision tree algorithms are used in this classification work.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

C. Result Analysis

The explorer carried out to the WAKA (Waikato Environment for Knowledge Analysis), in the first step to preprocessing only 1075 sample data and enter the classifying the Naïve Bayes, SVM, J48 (Decision Tree) algorithms. We have using WEKA tool for classification for algorithm shown in fig. 1, 2, and 3 is using three algorithms and its show the result in visualizing tree.

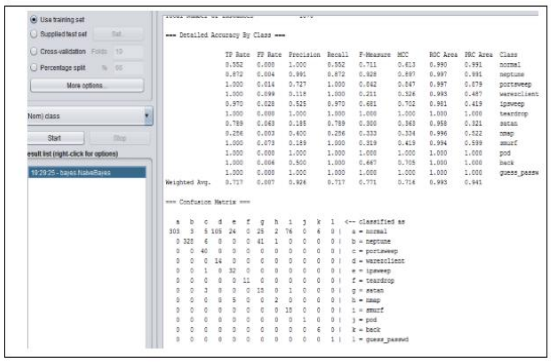

Figure 2: Classified NaiveBayes Algorithm of IDS Dataset.

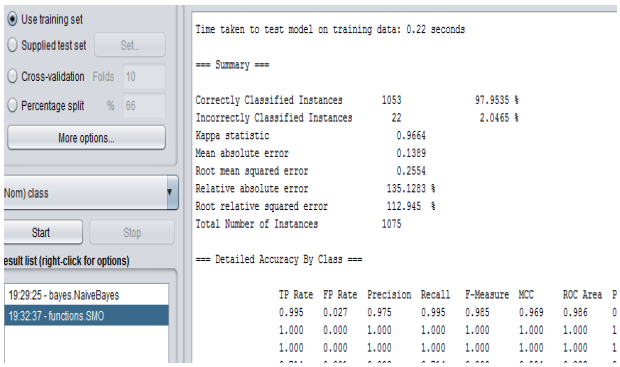

Figure 3: Confusion Matrix of SVM Algorithm.

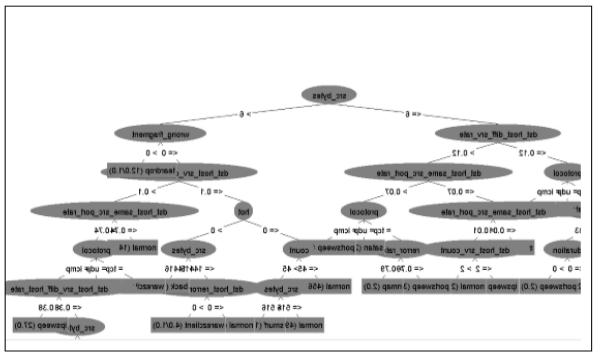

Figure 4: Decision Tree for IDS dataset.

VII. PERFORMANCE ANALYSIS

The three-supervised learning algorithm can be combined to predict different output results are given and only 20% dataset to analyzing training set in Weka tool. it gives better results for providing and Table 2 display the three-classification algorithm produced TPR, FPR, Accuracy, and ET (Execution Time) in percentage. NSL-KDD dataset has been using the accessing the data.

A. TP Rate

$$TPR = \frac{TP}{(TP + FN)}$$

B. FP Rate

$$FPR = \frac{FP}{(FP + TN)}$$

C. Accuracy

The display accuracy value is the proportion of correctly classified instance from the total amount of instance.

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)}$$

Table 2: Display the compared training dataset results

Classification Algorithms	**Dataset**	*TPR (%)*	*FPR (%)*	*Accuracy (%)*	*ET (Milliseconds)*
Naive Bayesian	*41 Attributes*	*71.70*	*0.7*	*92.60*	*500*
SVM	*41 Attributes*	*98.0*	*1.4*	*97.50*	*125*
J48	*41 Attributes*	***99.30***	***0.5***	***99.30***	*180*

During packets from source IP address to Designation IP address travel the network many intrudes are attacked in the virtual machine. While using the classification of naïve Bayes, SVM, J48, and Random Forest algorithm to gives a better result. The graphical representation of performance result is given below the figure 4 and figure 5 display the TP rate and FP rate respectively three machine learning algorithm with NSL_KDD data set.

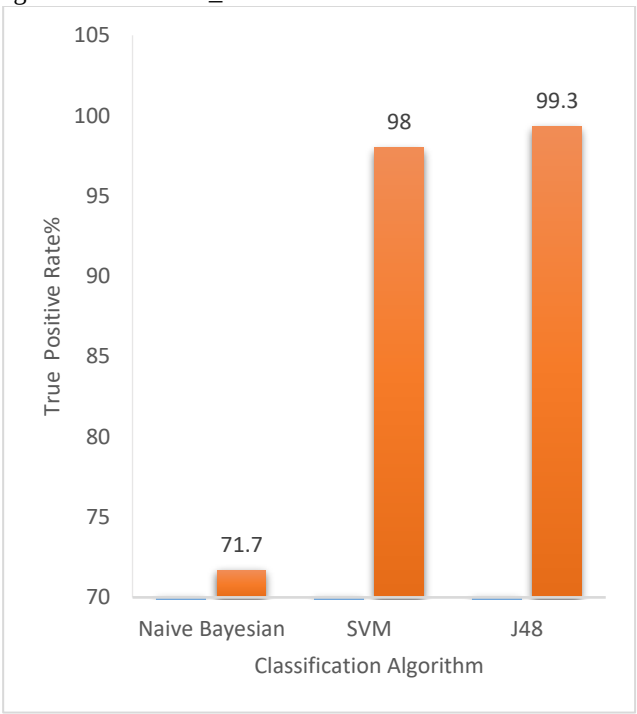

Figure 5: True positive rate of the NIDS

It shows the true positive rate results using 41 attributes for training data compared to three algorithm decision tree (>99%) with other. And it shows the false positive rate results using 41 attributes for training data compared to other the SVM algorithm only (<220 ms).

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

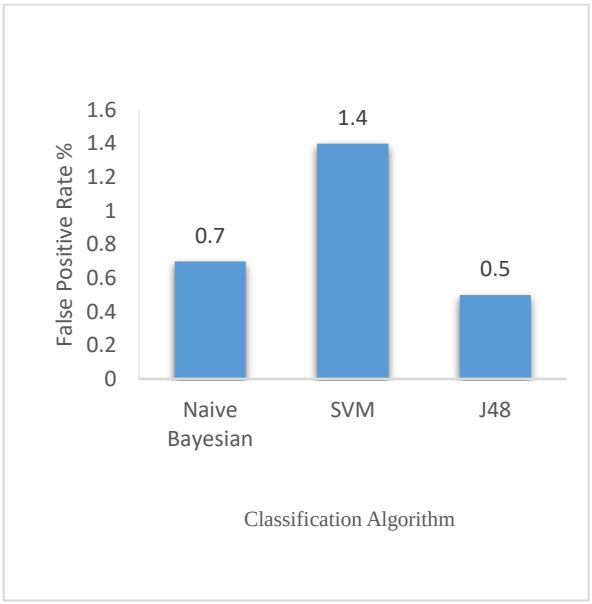

Figure 6: False positive rate of the NIDS

In figure 6 and figure 7 is display the accuracy and execution time.

The accuracy value of decision tree value is higher than SVM and naïve Bayesian. And the execution time is compared to other the SVM is low.

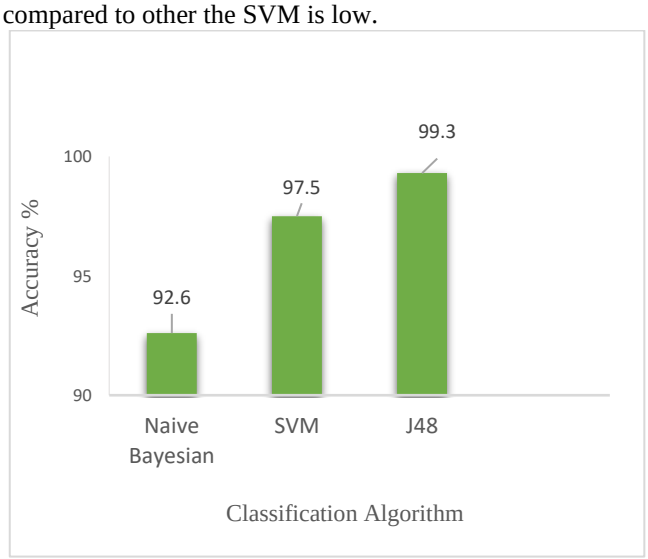

Figure 7: Accuracy of the NIDS

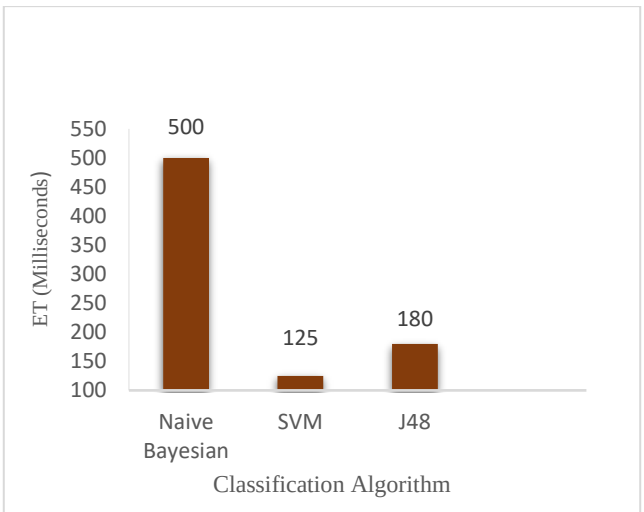

Figure 8: Execution time (in ms) of the NIDS

Figure 4,5,6,7 demonstrates only the top three machine learning algorithms performance result. Maximum all algorithm reached 80 % but j48 have 99% of TPR and FPR is very low time and accuracy is a maximum percentage and ET is compared to other in these two algorithms is a very short time to complete execution. The graph displays the high true positive rate, high accuracy and small execution time but compared to other random forest and j48 is produced the highest performance.

VIII.CONCLUSION & FUTURE WORK

This paper, we proposed a naive Bayes, SVM and decision tree algorithms using NSL-KDD dataset. In this dataset 41 attributes are available. The proposed technique need to 41 attributes is using. It is simulating the pre process dataset into simulating the training data. Because lots of network attack intrudes in the cloud. The traditional attacks are IP spoofing, DDOS, User to Port, Port Scanning etc. An IDS has a new efficient solution of the traditional network for securing packets. I have using sample data in anomaly techniques to improve the high accuracy and low false alarm. Simulation results decision tree is better than in terms of TP rate is almost 2% of SVM, 14% of naïve Bayesian. The decision tree is better than FP rate almost 1% Lower than SVM, 11% lower than naïve Bayesian. Accuracy value of decision tree is better than almost 2% of SVM, 20% of naïve Bayesian. And execution time of SVM is better than other. The main conclusion is that decision tree performance is better than other SVM and naïve Bayesian. As per our proposed work an effective approach of network IDS in cloud computing. In future work using the optimal feature selection algorithm for reducing the attributes and to build the training model.

REFERENCES

1. U. Kumar, “A Survey on Intrusion Detection Systems for Cloud Computing Environment,” International Journal of Computer Applications, vol. 109, no. 1, pp. 6–15, 2015.
2. K. Arjunan and C. N. Modi, “An enhanced intrusion detection framework for securing network layer of cloud computing,” ISEA Asia Security and Privacy Conference 2017, ISEASP 2017, 2017.
3. B. Mahalakshmi and G. Suseendran, “Effectuation of Secure Authorized Deduplication in Hybrid Cloud,” Indian Journal of Science and Technology, vol. 9, no. 25, Jul. 2016.
4. T. Nathiya, “Reducing DDOS Attack Techniques in Cloud Computing Network Technology,” International Journal of Innovative Research in Applied Sciences and Engineering (IJIRASE), vol. 1, no. 1, pp. 23–29, 2017.
5. R. K. Bathla, G. Suseendran, and Shallu, “Research analysis of big data and cloud computing with emerging impact of testing,” International Journal of Engineering and Technology(UAE), vol. 7, no. 3.27 Special Issue 27, pp. 239–243, 2018.
6. R. Staudemeyer and C. Omlin, “Extracting salient features for network intrusion detection using machine learning methods,” South African Computer Journal, vol. 52, no. July, pp. 82–96, 2014.
7. S. G. Kene and D. P. Theng, “A Review on Intrusion Detection Techniques for cloud computing and Security Challenges,” IEEE sponsored 2nd International Conference on Electronics and Communication Systems (ICECS), pp. 227–232, 2015.
8. N. Modi, “An Efficient Security Framework to Detect Intrusions at Virtual Network Layer of Cloud Computing,” 19th international ICIN conference- Innovations in clouds, Internet and Network, pp. 133–140, 2016.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

9. T. Nathiya and G. Suseendran, An Effective Hybrid Intrusion Detection System for Use in Security Monitoring in the Virtual Network Layer of Cloud Computing Technology,Data Management, Analytics and Innovation, Advances in Intelligent Systems and Computing,839, pp.483-496,2019. Doi: 10.1007/978-981-13-1274-8_3.
10. Hasan, M. Nasser, B. Pal, and S. Ahmad, "Support Vector Machine and Random Forest Modeling for Intrusion Detection System (IDS)," Journal of Intelligent Learning Systems and Applications, vol. 6, no. February, pp. 45–52, 2014.
11. Farnaaz and M. A. Jabbar, "Random Forest Modeling for Network Intrusion Detection System," Procedia Computer Science, vol. 89, pp. 213–217, 2016.
12. G. V. Nadiammai and M. Hemalatha, "Effective approach toward Intrusion Detection System using data mining techniques," Egyptian Informatics Journal, vol. 15, no. 1, pp. 37–50, 2014.
13. Jing, Y. Bi, and H. Deng, "An innovative two-stage fuzzy kNN-DST classifier for unknown intrusion detection," International Arab Journal of Information Technology, vol. 13, no. 4, pp. 359–366, 2016.
14. Rouhi, F. Keynia, and M. Amiri, "Improving the Intrusion Detection Systems' Performance by Correlation as a Sample Selection Method," Journal of Computer Sciences and Applications, vol. 1, no. 3, pp. 33–38, 2013.
15. O. Osanaiye, H. Cai, K. K. R. Choo, A. Dehghantanha, Z. Xu, and M. Dlodlo, "Ensemble-based multi-filter feature selection method for DDoS detection in cloud computing," Eurasip Journal on Wireless Communications and Networking, vol. 2016, no. 1, 2016.
16. Ö. Cepheli, S. Büyükçorak, and G. Karabulut Kurt, "Hybrid Intrusion Detection System for DDoS Attacks," Journal of Electrical and Computer Engineering, vol. 2016, 2016.
17. N. M. Turab, A. Abu, and T. Shadi, "C Loud Computing Challenges and Solutions," International Journal of Computer Networks & Communications (IJCNC), vol. 5, no. 5, pp. 209–216, 2013.
18. N. Hubballi and V. Suryanarayanan, "False alarm minimization techniques in signature-based intrusion detection systems: A survey," Computer Communications, vol. 49, pp. 1–17, 2014.
19. K. Chai, H. T. Hn, and H. L. Cheiu, "Naive-Bayes Classification Algorithm," Bayesian Online Classifiers for Text Classification and Filtering, pp. 97–104, 2002.
20. H. Chauhan and A. Chauhan, "Implementation of decision tree algorithm c4.5 1," International journal of scientific and research publications, vol. 3, no. 10, pp. 4–6, 2013.
21. O. Catak and M. E. Balaban, "CloudSVM: Training an SVM classifier in cloud computing systems," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 7719 LNCS, no. January, pp. 57–68, 2013.
22. M. S. Revathi, "A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection," International Jornal of Engineering Research and Technology, vol. 2, no. 12, pp. 1848–1853, 2013.
23. L. Dhanabal and S. P. Shantharajah, "A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms," International Journal of Advanced Research in Computer and Communication Engineering, vol. 4, no. 6, pp. 446–452, 2015.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)