Intelligent Intrusion Detection System using Supervised Learning

Sandipan Roy[1*], Apurbo Mandal[2], and Debraj Dey[3]

[1]Dept. of Computer Science and Engineering, Aliah University, Kolkata, India.
[2]System Engineer, Precision Infomatic (M) Private limited, Berhampore, India.
[3]Dept. of Computational Science, Brainware University, Kolkata, India.

*Corresponding author
doi: https://doi.org/10.21467/proceedings.115.3

ABSTRACT

Going digital involves networking with so many connected devices, so network security becomes a critical task for everyone. But an intrusion detection system can help us to detect malicious activity in a system or network. But generally, intrusion detection systems (IDS) are not reliable and sustainable also they require more resources. In recent years so many machine learning methods are proposed to give higher accuracy with minimal false alerts. But analyzing those huge traffic data is still challenging. So, in this article, we proposed a technique using the Support Vector Machine & Naive Bayes algorithm, by using this we can solve the classification problem of the intrusion detection system. For evaluating our proposed method, we use NSL-KDD and UNSW-NB15 dataset. And after getting the result we see that the SVM works better than the Naive Bayes algorithm on that dataset.

Keywords: Intrusion Detection System, Network Security, Support Vector Machine, Machine Learning

1 Introduction

In this digital era, every technology or system is a primary need to build, grow and follow success for every government, organization, or individual. All personal, organizational, and governmental information or data validate through a digital system. If any unauthorized access won that data or information that it's become a nightmare. So, we need to ensure that our data stays protected on that digital system. To secure our data we need a system that can detect malicious access to our pieces of information and system resources. The system that is trying to detect those activities and inform us about them, known as Intrusion Detection System. A general overview of an IDS is shown in Figure 1. The main aim of an Intrusion Detection System is to check and identify security issues automatically like, malicious traffic, contents, unauthorized access, file integrity, and many more and block them or send an alert to the network administrators. Also, there were certain problems like bugs in software, broken access control, vulnerable cryptographic systems that lead us to invest time and money for developing an Intrusion detection system (IDS) [1, 3, 11, 12, 15]. Generally, an IDS has three major steps:

A. Collecting the Information: Collect all data from various hosts, nodes, applications, and networks. B. Analyzing Data: Those captured or stored data are analyzed and check if they are malicious or not. C. Response: After analysis of data, stored them and send the real-time alert accordingly with the response to that incident.

Proceedings DOI: 10.21467/proceedings.115; Series: AIJR Proceedings; ISSN: 2582-3922; ISBN: 978-81-947843-5-7

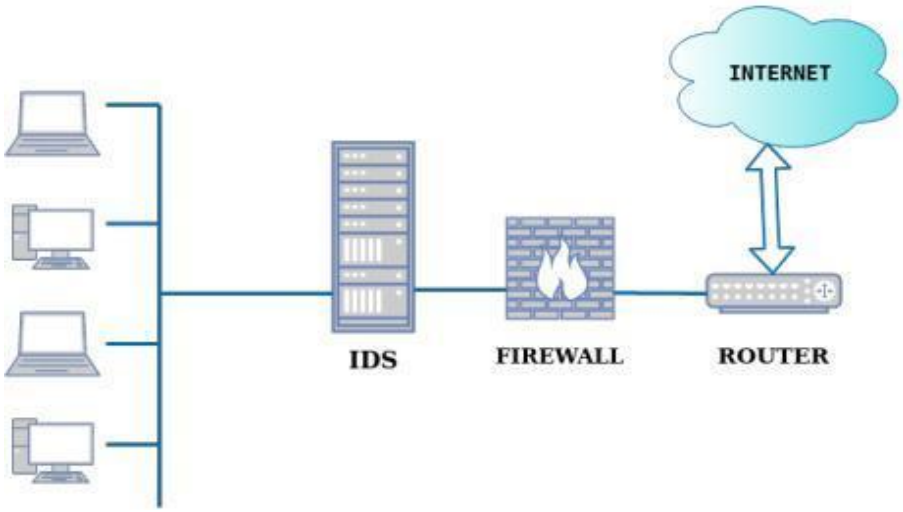

Fig. 1: An Intrusion Detection System (IDS).

Normally there are two methodologies to develop an intrusion detection system, 1. Using various data sources IDS can build. Those are usually divided into two categories again, Network-Based and Host-Based IDS. 2. Also using various analysis methods IDS can build up. That kind of IDS also has two major categories, Anomaly Detection [7, 20] and Misuse Detection [21]. Comes to Host-based IDS, the data or activity recorded from various hosts, system logs, operating systems, applications, audit records, and many more. The advantage of this host-based system is that it can judge more accurately because it comes from logs and does not require additional hardware support. Besides, host-based IDS can decrypt the encrypted traffic packets. But the main disadvantage of this system is that it is required to be installed on every host or node. For network-based IDS, it is focused on collected network segments like network packets. It is a low cost IDS that can detect attacks like DoS (Denial of Service) [9, 10], that cannot be detected by any Host-based IDS. But this type of IDS requires more resources like CPU, Memory to analyze huge data traffic, also Network-based IDS can't decrypt encrypted data packets. On the other hand, analysis-based methods like Misure detection also known as signature-based detection can detect attacks via checking or comparing signatures from a stored signature database. So, it is giving us the result very first for known attacks, but the signature database needs to update frequently. And the Anomaly detection [20] technique can be used as a hybrid Intrusion detection system to reduce false alerts. But nowadays Machine Learning is used in every field like computer vision, medical diagnosis, emotion detection, and many more. Recently researchers are also using these machine learning algorithms to intrusion detection, that can help us for speeding up the intrusion detection process with less number of false results with accurate detection rate. In our system, we used Support Vector Machine [16] also known as SVM that comes under the category of supervised learning and creates a hyper-plane or multi hyper-planes in a high dimensional space. For non-linear classifiers, it uses various kernel functions and maximizes the margin between hyper-planes. On the other hand, we use the Naive Bayes Algorithm [17], which is a statistical classifier which can help us to check whether a model fits any particular class or not.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Proceedings DOI: 10.21467/proceedings.115
ISBN: 978-81-947843-5-7

Series: AIJR Proceedings
ISSN: 2582-3922

2 Review of related literature

In recent trends, there are so many approaches given by researchers that can solve intrusion detection problems. A few summarized pieces of literature are described as follows.

2.1 Random Tree Classifier for IDS

M. Almseidin et. al [2] introduced a method using Random Tree Classifier that uses a fixed number of classified trees and each tree represented by a single decision point. So random tree classifiers can be featured as a finite set group the decision tree. With this method, they use the same KDD dataset and get a 90.57% of accuracy rate.

2.2 Supervised Logistic Regression

A Logistic Regression based system developed by C. Belavagi et. al. [4] For binary and multi class classification they used this LR (Logistic Regression) based function. They fitted the data-set by the probability of occurrence of any event using the LR function. They use the Equation,$h_\theta(x) = g(1/1 + e^{-\theta^T}\ x)$ for this. After applying the Logistic Regression to the NSL-KDD dataset they get a 0.84 accuracy rate along with 0.83 Precision, 0.85 Recall, and 0.82 F1-Score.

2.3 Adaptive Voting Algorithm for IDS

Another research done by Gao et. al. [5] using the Adaptive Voting Algorithm. In this, the voting classifiers used logistic regression, kNN(k-nearest neighbors), Support Vector Machine, Adaboost and random forest. Then the adaptive voting model was generated and applied to the NSL-KDD dataset [22]. The diagram of their model is shown in figure 2. And after training, they got 81.6% accuracy on the KDDTest+ dataset.

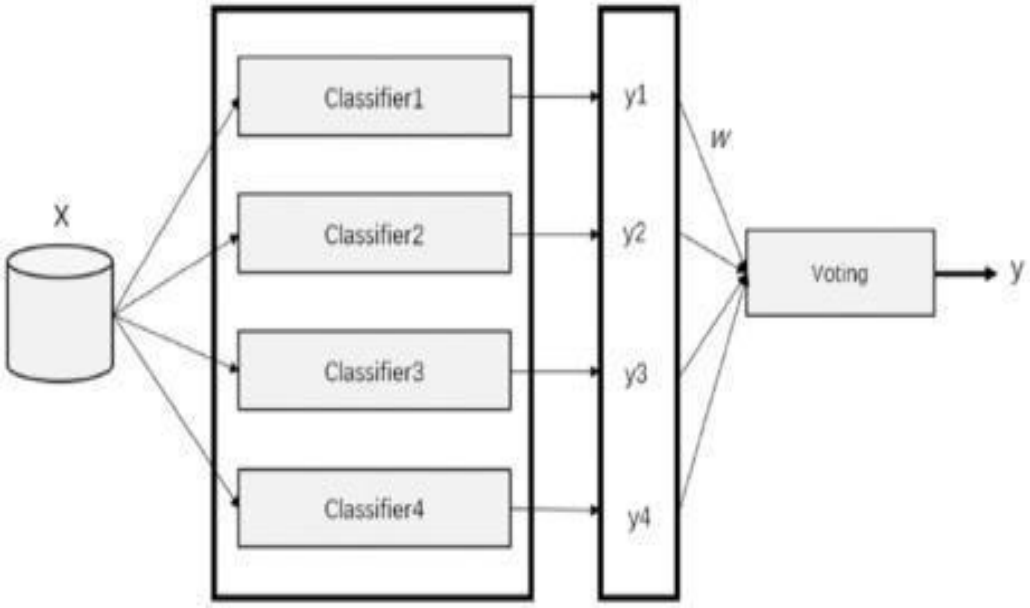

Fig. 2: IDS using Adaptive Voting Algorithm [5].

3 Proposed Method

Our proposed system for an intelligent intrusion detection system is based on the Naive Bayes algorithm and support vector machine. As we know the supervised learning techniques are having various types of classification algorithms but a few of them are easy to use and give a faster & reliable result. We used the SVM and Naive Bayes method over other algorithms because the Intrusion Detection System is a classification problem. As we can see many researchers are using SVM with the KDD-NSL dataset for their research and evaluation purposes. But in

In our research we used a forward selective wrapper selection method and labelled the dataset in different classes, then we used our normalized SVM and SVM algorithm. We divide our system into two parts, So, let's see that one by one.

3.1 Prepossessing

We use NSL-KDD [14] and UNSW-NB15 [13] dataset. So, pre-processing the datasets and evaluation of results are the challenging phases of our proposed model. In our system, every phase is important and essential to evaluate the performance of the system. Also, SVM and Naive Bayes classifiers are playing a vital role. An overview of our proposed system shown in figure 3.

In the pre-processing steps, we processed the symbolic features, like protocol, service, flags, etc that are present in the datasets and we remove all symbolic and non-numeric features because they are not required active participation for our intrusion detection system. We converted the nominal attributes to numerical attributes for improving the performance. After that, we use a forward selected wrapper selection method that helps us to improve performance and reduction of features.

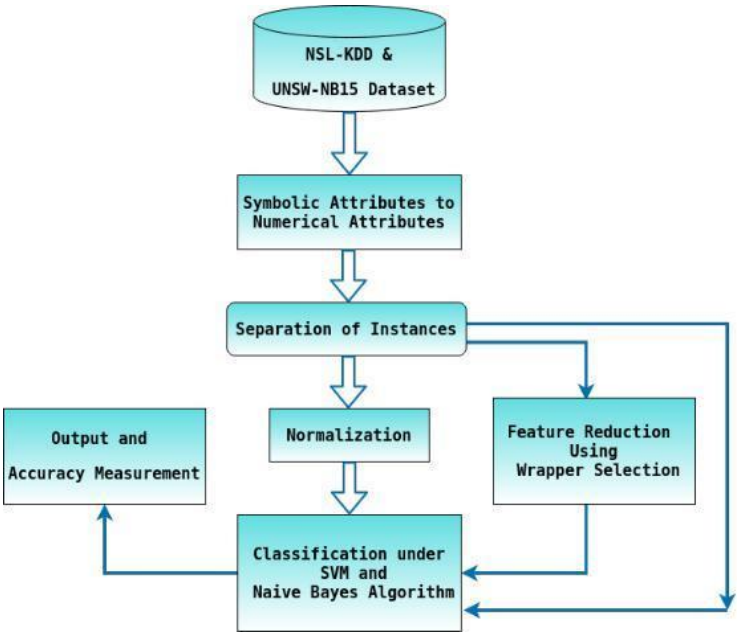

Fig. 3: Overview of Our Proposed System

Proceedings DOI: 10.21467/proceedings.115
ISBN: 978-81-947843-5-7

Series: AIJR Proceedings
ISSN: 2582-3922

3.2 Feature Selection and Methodology

After pre-processing, we labelled NSL-KDD [14] data into four categories Remote to Local (R2L) attacks, Probe attacks, User to Root(U2R) attacks, and Denial of Service (DoS) attacks. Where R2L attack means, an attacker gains access to a local machine via sending a malicious packet using a remote machine, for example, IMAP [18], file write attack, etc. In the Probe attack, the attacker gained information about the whole network before exploiting the network, like Nmap, IP sweep attacks. When normal users exploit the system and gain the root privilege known as U2R. Attacks like load module, command injections, eject attacks, and many more. And extensive usage of internet bandwidth known as DoS [19] attack, for example, Smurf, TearDrop, etc. Total 24 types of attacks are present in those four categories. We also used the UNSW-NB15 [13] dataset that has nearly 100 GBs of TCP dump captured with two million records with 45 attributes. But in the CSV file of the customized dataset having 540,044 records.For the UNSW-NB15 [6–8] dataset we select smean, service, ct src dport ltm, ct dst ltm, ct dst sport ltm, tcprtt, dwin as features. And the training set having 175,341 records where testing set with 82,332 records are used.After the feature selection step, we randomize the data and collect 21,000 instances of the NSL-KDD dataset. These collected instances are used for comparative analysis and performance measures. Then we applied the SVM [16] and Naive Bayes classification method to classify our dataset. But we also applied
various normalization methods and forward selected wrapper method for better evaluation and measurement purpose. The algorithm we followed for SVM is shown in Algorithm 1 and the Naive Bayes classifier is shown in Algorithm 2.

We use sci-kit-learn, NumPy and Pandas libraries, and python for implementing our method on a modern computer.

Algorithm 1 Support Vector Machine (SVM)

IN: Pre-Processed Dataset.
OUT: Right/Wrong Classi ed Classes.
System Initialization
x_i = Input Sample; y_i = Labeled Output;
w_v = Weight Vector;
λ= Regularized Parameter; F = Object Function

$F = (\min w_v \lambda) \; || \; w_v \; || \; (2 + (1 y_i(x_i, w_v)))$

Applying Gradient Descent with respect to Weight(w_v)
$w_v = w_v + \eta \; (y_i x_i - 2 \lambda w_v)$ Used for Updating Wrong Classification
$w_v = w_v + \eta \; (i - 2 \lambda w_v)$ Used for Updating Right Classification
Return F

Algorithm 2 Naive Bayes Algorithm

IN: T = Train Dataset; P = Prediction Variable;

OUT: Testing Set.

System Initialization

Reading the Training Dataset T.

Calculate the Probabilistic function P for each class

A = Dependency Classes; B = Class Variables

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Sort Classes by its Maximum Probabilistic Value

Generating the Confusion Matrix

Finding Accuracy & False Positive Rate (FPR)

4 Performance Measurement

Measuring the result of our trained model using a test dataset we use TP, TN, FP and FN. Those terms are described as follows, True Positive (TP): This represents the values that are correctly classified as attacks. True Negative (TN) : This represents the values that are correctly classified as normal or general packets. False Positive (FP) : FP represents the error values that are classified as attack but in reality those are normal packets. False Negative (FN) : FN used for incorrect
classified packets that's actually attacked but system labeled as a normal packet.

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

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

After the analysis, using accuracy rate and false positive rate (FPR) of our 21,000 instances by Support Vector Machine [16] and Naive Bayes algorithm we got the data shown in Table 1 and Fig.4. And the analysis of the UNSW-NB15 data-set shown in Table 2 and Fig. 5.

Proceedings DOI: 10.21467/proceedings.115
ISBN: 978-81-947843-5-7

Series: AIJR Proceedings
ISSN: 2582-3922

Table 1: COMPARISON OF ACCURACY AND FPR FROM DIFFERENT METHODS ON KDD DATASET.

No.	Method	Accuracy	FPR
1	Naive Bayes	74.64	24.96
2	SVM	98.31	1.26
3	Naive Bayes - Wrapper selection	68.57	31.23
4	SVM - Wrapper selection	97.61	2.29
5	Naive Bayes - Normalization	82.49	17.51
6	SVM - Normalization	97.14	2.78

Table 2: COMPARISON OF ACCURACY AND FPR FROM DIFFERENT METHODS ON UNSW-NB15 DATASET

No.	Method	Accuracy	FPR
1	Naive Bayes	69.47	29.16
2	SVM	96.54	3.13
3	Naive Bayes - Wrapper selection	74.62	25.17
4	SVM - Wrapper selection	96.21	3.64
5	Naive Bayes - Normalization	76.83	22.92
6	SVM - Normalization	96.43	3.20

For the NSL-KDD dataset, we got a 74.64 accuracy rate and 24.96 FPR using the Naive Bayes method. Then we got an accuracy of 98.31 and 1.26 FPR using the Support Vector Machine. Then we applied a customized wrapper selection method on both Naive Bayes and SVM and we got a 68.57 & 97.61 accuracy rate with 31.23 & 2.29 of false-positive Rate respectively. Then we also evaluate our system using normalized Naive Bayes and SVM. After applying the normalized method, we got 82.49 & 97.14 accuracy rate and 17.51 & 2.78 false-positive rate respectively for Naive Bayes and SVM.

For the UNSW-NB15 data-set, together we got a 69.47 accuracy rate and 29.15 FPR on the Naive Bayes method. After that we got an accuracy of 96.54 and 3.13 FPR using the Support Vector Machine. Then we applied a customized wrapper selection method on Naive Bayes and SVM and we got a 74.62 & 96.21 accuracy rate with 25.17 & 3.64 of false-positive rate respectively. Then we also evaluate our system using normalized Naive Bayes [17] and SVM. After applying the normalized method we got 76.83 & 96.43 accuracy rate and 22.92 & 3.20 false-positive rate respectively for Naive Bayes and SVM [16].

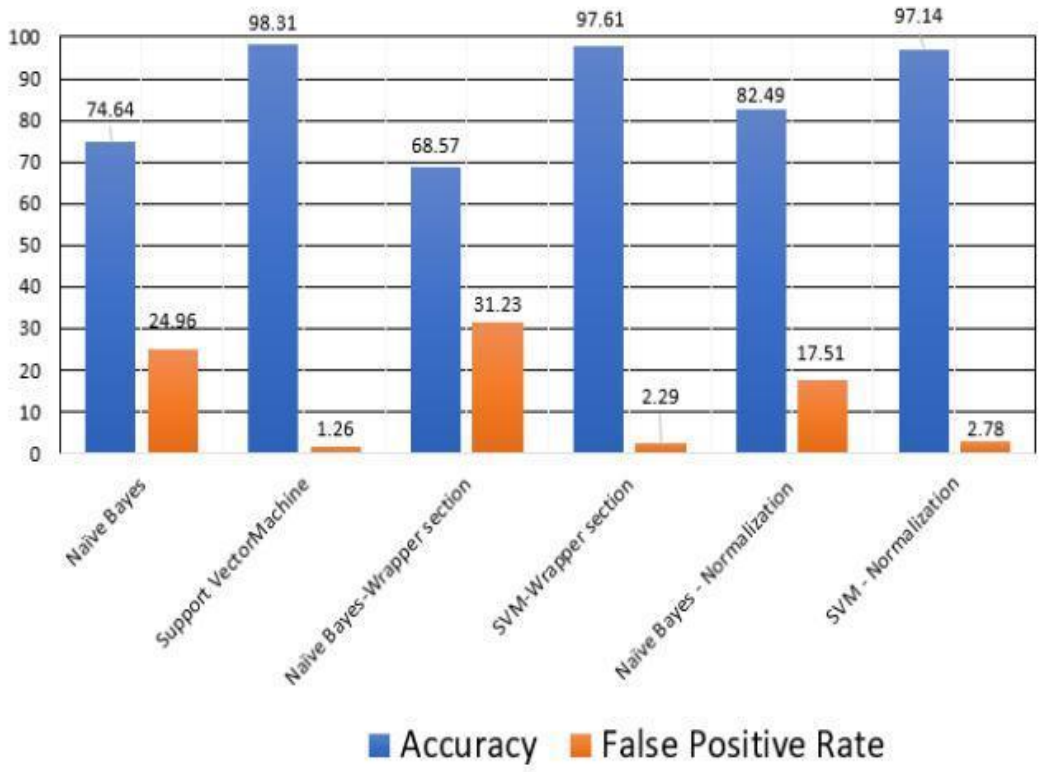

Fig. 4: Accuracy Rate and False Positive Rate (FPR) for KDDTest++ Dataset.

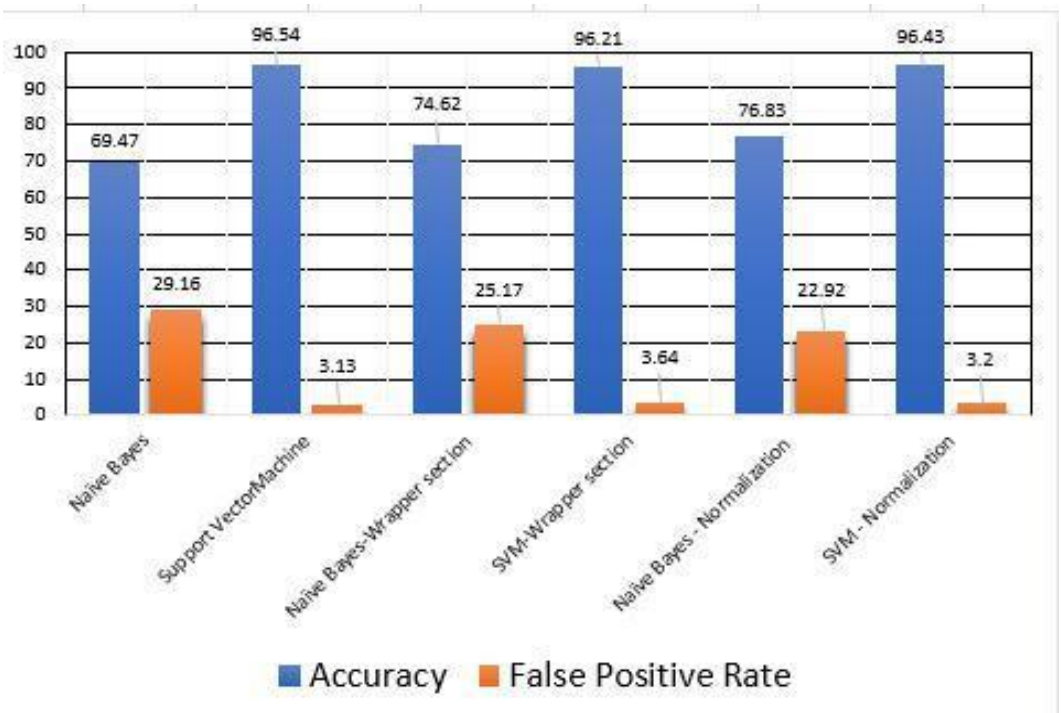

Fig. 5: Accuracy Rate and False Positive Rate(FPR) for UNSW-NB15 Dataset.

Proceedings DOI: 10.21467/proceedings.115
ISBN: 978-81-947843-5-7

Series: AIJR Proceedings
ISSN: 2582-3922

5 Conclusion and Future Work

In this, we used the NSL-KDD dataset that is pretty old and used for many research work but this dataset can be used for comparative results. By our proposed system, this dataset performs using the SVM algorithm. And we see that the Support Vector Machine [16] method gives fewer false alarms with higher accuracy than Naive Bayes. Also, we use a new, big dataset named UNSW-NB15. In that dataset we also see that our SVM stands out and gives us good accuracy. But this dataset having so much data and some TCP dumps are not labeled correctly, so we got low accuracy compared with the NSL-KDD dataset. But in the future, this dataset can be used on deep neural networks to get a better result and fewer false alerts.

6 Conflicts of Interests

The authors report no conflicts of interest.

References

[1] Kelton A.P. da Costa, Jo~ao P. Papa, Celso O. Lisboa,Roberto Munoz, Victor Hugo C. de Albuquerque,Internet of Things: A survey on machine learning-based in-trusion detection approaches, Computer Networks,2019,Pages 147-157,ISSN 1389-1286,https://doi.org/10.1016/j.comnet.2019.01.023.

[2] M. Almseidin, M. Alzubi, S. Kovacs and M. Alkasassbeh, "Evaluation of machine learning algorithms for intrusion detection system," 2017 IEEE 15th International Symposium on Intelligent Systems and Informatics (SISY), Subotica, 2017, pp. 000277-000282, doi: 10.1109/SISY.2017.8080566.

[3] Phurivit Sangkatsanee, Naruemon Wattanapongsakorn, Chalermpol Charn-sripinyo, Practical real-time intrusion detection using machine learning ap-proaches,Computer Communications, Volume 34, Issue 18,2011,Pages 2227-2235,ISSN 0140-3664,https://doi.org/10.1016/j.comcom.2011.07.001.

[4] Manjula C. Belavagi, Balachandra Muniyal,Performance Evaluation of Supervised Machine Learning Algorithms for Intrusion Detec-tion,Procedia Computer Science,Volume 89,2016,Pages 117-123,ISSN 1877-0509,https://doi.org/10.1016/j.procs.2016.06.016.

[5] X. Gao, C. Shan, C. Hu, Z. Niu and Z. Liu, "An Adaptive Ensemble Machine Learning Model for Intrusion Detection," in IEEE Access, vol. 7, pp. 82512-82521,2019,doi:10.1109/ACCESS.2019.2923640.

[6] N. Moustafa and J. Slay, "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," 2015 Military Com-munications and Information Systems Conference (MilCIS), Canberra, ACT, Aus-tralia, 2015, pp. 1-6, doi: 10.1109/MilCIS.2015.7348942.

[7] Nour Moustafa & Jill Slay (2016) The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set, Information Security Journal: A Global Perspective, 25:1-3, 18-31, DOI: 10.1080/19393555.2015.1125974

[8] N. Moustafa, J. Slay and G. Creech, "Novel Geometric Area Analysis Technique for Anomaly Detection Using Trapezoidal Area Estimation on Large-Scale Networks, in IEEE Transactions on Big Data, vol. 5, no. 4, pp. 481-494, 1 Dec. 2019, doi:10.1109/TBDATA.2017.2715166.

[9] Tan, Z., Jamdagni, A., He, X., Nanda, P., Liu, R.P.: Denial-of-service attack de-tection based on multivariate correlation analysis. In: International Conference on Neural Information Processing, pp. 756{765. Springer (2011)

[10] Tan, Z., Jamdagni, A., He, X., Nanda, P., Liu, R.P.: A system for denial-of-service attack detection based on multivariate correlation analysis. IEEE transactions on parallel and distributed systems 25(2), 447{456 (2014)

[11] Zuech, R., Khoshgoftaar, T.M., Wald, R.: Intrusion detection and big heterogeneous data: a survey. Journal of Big Data 2(1), 1 (2015)

[12] Moustafa N., Creech G., Slay J. (2017) Big Data Analytics for Intrusion Detec-tion System: Statistical Decision-Making Using Finite Dirichlet Mixture Mod-els. In: Palomares Carrascosa I., Kalutarage H., Huang Y. (eds) Data Analyt-ics and Decision Support for Cybersecurity. Data Analytics. Springer, Cham. https://doi.org/10.1007/978-3-319-59439-2 5

[13] Moustafa, N., Slay, J.: The evaluation of network anomaly detection systems: Sta-tistical analysis of the unsw-nb15 data set and the comparison with the kdd99 data set. Information Security Journal: A Global Perspective (2016)

[14] M. Tavallaee, E. Bagheri, W. Lu, and A. Ghorbani, \A Detailed Analysis of the KDD CUP 99 Data Set," Submitted to Second IEEE Symposium on Computa-tional Intelligence for Security and Defense Applications (CISDA), 2009

[15] M.A. Jabbar, Rajanikanth Aluvalu, Sai Satyanarayana Reddy S,RFAODE: A Novel Ensemble Intrusion Detection System,Procedia Computer Science,Volume 115,2017,Pages 226-234,ISSN 1877-0509,https://doi.org/10.1016/j.procs.2017.09.129.

[16] Suthaharan S. (2016) Support Vector Machine. In: Machine Learning Models and Algorithms for Big Data Classi cation. Integrated Series in Information Systems, vol 36. Springer, Boston, MA.https://doi.org/10.1007/978-1-4899-7641-3 9

[17] Shenglei Chen, Geo rey I. Webb, Linyuan Liu, Xin Ma,A novel selective naive Bayes algorithm,Knowledge-Based Systems,Volume 192,2020,105361,ISSN 0950-7051,https://doi.org/10.1016/j.knosys.2019.105361.

[18] Robin Dhamankar, Yoonkyong Lee, AnHai Doan, Alon Halevy, Pedro Domingos, SIGMOD '04: Proceedings of the 2004 ACM SIGMOD in-ternational conference on Management of dataJune 2004 Pages 383{394 https://doi.org/10.1145/1007568.1007612

[19] Nikita Borisov, George Danezis, Prateek Mittal, Parisa Tabriz, CCS '07: Proceed-ings of the 14th ACM conference on Computer and communications security Oc-tober 2007 Pages 92{102 https://doi.org/10.1145/1315245.1315258

[20] Pajouh, H.H., Dastghaibyfard, G. & Hashemi, S. Two-tier network anomaly detection model: a machine learning approach. J Intell Inf Syst 48, 61{74 (2017). https://doi.org/10.1007/s10844-015-0388-x

[21] Sasan, Harvinder Pal Singh, and Meenakshi Sharma. "Intrusion detection using feature selection and machine learning algorithm with misuse detection." International Journal of Computer Science and Information Technologies 8.1 (2016): 17-25.

[22] L. Haripriya, M.A. Jabbar. "Role of Machine Learning in Intrusion Detection Sys-tem:Review", 2018 Second International Conference on Electronics, Communica-tion and Aerospace Technology (ICECA), 2018

Proceedings DOI: 10.21467/proceedings.115
ISBN: 978-81-947843-5-7

Series: AIJR Proceedings
ISSN: 2582-3922