IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 21, Issue 1, Ser. III (Jan - Feb 2019), PP 41-47
www.iosrjournals.org

Comparative study for Networks Intrusion Detection (NID) By Using Various Classification Algorithms to protection the Networks Systems from attacks (Comparison Studies)

Dr. Sefer Kurnaz[1],Omer Fawzi Awad[2]
[1]*Electrical and Computer Engineering Altinbas university Istanbul, turkey*
[2]*Electrical and Computer Engineering Altinbas university Istanbul, turkey*
Corresponding Author: Dr. Sefer Kurnaz

Abstract: *Network traffic anomalies could refer to a potential intrusion in the network, so detecting anomalies is important for the detection and prevention of security attacks. Earlier researches in this part and commercially obtainable Intrusion Detection Systems (IDS) are, for the most part based on signature. The signature-based task is the need to update the signature of the datasets when there are new attack signatures and thus, they are not appropriate for detect real-time network anomalies. The recent trend in detecting anomalies is based on classification approaches of automated learning. We apply three different techniques for automated learning into the NSL_KDD dataset and evaluating the performance for those methods. Our findings have shown that for this specific dataset, the majority of automatic learning approaches porvide a higher accuracy than 90%, memory and accuracy. On the other hand, the Random Forest model achieved the best performance accuracy among the three algorithms that have been researched in this study.*

Keywords: *Machine learning, Network Intrusion Detection System, Data mining*

Date of Submission: 13-02-2019 Date of acceptance:28-02-2019

I. Introduction

Due to developments in the Internet, cyber security has gained increasing attention [1], [2], which has encouraged many researchers to design effective protection system called Network Intrusion Detection Systems (NIDSs). Intrusion defined as “attempts to breach the confidentiality, integrity or availability a Computer or Network, or to bypass the security mechanisms of a computer or network.” Using this concept, IDS can be defined as programs or devices that are supposed to disclose attempts to breach the availability, confidentiality, and integrity of a computer or network, or to bypass a computer or network’s security mechanisms. Differentiated models have be suggest for the network behavior model and the detection of Abnormal flows [3]. At present, Intrusion Detection methods are divide for two main categories: misuse-based detection and detecting anomalies. Systems to detect list based on random naming are extracted from the known characteristics and patterns of known attacks and are manually symbolized in the system. These rules are compared with traffic to detect attacks. They are effective and effective to detect a known type of attack and have a so lower false alarm. Therefore, misuse-based detection systems are currently prevalent in NIDS, some of which have been filed in real scenarios, such as snort [4]. However, abuse detection systems require frequent update of basics and signatures and are incapable of identifying any new or unknown attacks. In late years, Anomaly-based Network Intrusion Detection Systems (ANIDSs) have attract many attention to their ability to detect zero-day attacks. They rely on data mining algorithms or machine learning algorithms to sample the Normal behavior of the network and detect deviations as deviations from regular behavior. There are three types of algorithms that presents in this paper and experiments at the NSL_KDD [5], these algorithms are Random forest (RF), K-Nearest Neighbors (K-NN), and Naive Bayes (NB). Those algorithms are all initially evaluated with the use of 3 well-known metrics which are (i) precision, (ii) recall (iii) accuracy and next using the ROC metric. It has been found that the ROC metric is the most suitable to rank the results of the algorithms.

Intrusion Detection Systems (Idss)

In this part we present the aim of IDS is to detected harmful traffic. In ordering to fulfill this, IDS monitors all traffic that comes in or goes out. There are many ways to do this around implementing IDS. Among these, the two most popular:

Anomaly detection: That technicality is depend on the detect of traffic anomaly. The observed traffic aberration of the Normal measured. Various different applications has be suggest for that technique, based on metrics used to measure the deviation of the traffic profile.

Misuse/Signature detection: That technicality look for patterns and signatures of attacks already known in network traffic. Typically a data-base is utilized to update constantly to store signatures of known attacks. The way this technology handles intrusion detection is same to the way antivirus programming works.

Drawbacks Of Idss

IDS systems has be a standard component of security infrastructure, allowing network director to detection the politics violations. Some of the policy violations that maybe do the sets of the exterior attackers attempting to unauthorized access to online insiders who misuse them.

NIDS With Data Mining

Data mining (DM) is a highly automatic search of data volumes for patterns that use correlation rules. It is a modern subject in computer science but uses many old computer techniques of statistics, information retrieval, automatic learning and pattern recognition.

Here are some particular things which could play the role of extracting data in an Infiltration works:

(i) Removing normal activity from alert data in order to allow analysts to be focused on real attacks. (ii) Identifying the wrong alarm generators and the “bad” sensor signatures. (iii) The search for abnormal activity reveals a real attack. (iv) Define long and continuous patterns (different IP address, same activity) [6].

Data extraction techniques can be distinguished by their different functions in representation, representation, preference criteria, and algorithms [6].

The main function of the model to which we are concerned is classification, as Normal or Abnormal [7], [8].

Supervised And Unsupervised Learning

Intrusion Detection approaches may be divided of two types supervised and unsupervised learning, we explain this in this section.

Supervised Learning

Learning to supervise is the method of learning the machine, supervised by the word, as there is some sorting of assistant in the way it is used, there be a path to assist them with the labels that determine which data will be used. From another concept, the main objective of the supervised approach is to create a predictive model (classifier) for classifying or labeling incoming patterns. The classifier must be training with lablled styles to be fit for the classification new unlablled styles. The named training patterns are used for learning the classes and descriptions. Some of the supervised approaches include support vector machines (SVM), neural networks (NNs) and genetic algorithms (GAs) among others [9].

Unsupervised Learning

Unsupervised learning is the opposite of supervised learning, rather than merely inspecting one of the labels, which is performed in supervised, it looks at the whole picture. It collects data then processes the entire input, for the sake of making decisions based on the entire input. When utilizing unsupervised learning the system isn’t aware of the correct “answers”. This is one of the main reasons of the use of unsupervised learning, it should be capable of taking in the input which has been sent in/given and come up with some structured answer on the basis of patterns that it was trained to recognize or to collect sets of data. It could, for instance, be utilized by stores for the analysis of patterns in shopping habits of people and group the people with identical habits and send out offers on that basis [10], [11].

From another concept, unsupervised methods follow a different method via grouping unlabeled patterns into clusters according to similarities. Patterns that belong to the same clusters are more identical to one another than they are to patterns pertinence to several of the clusters. Data clustering is highly beneficial when little priori information about the data can be obtained [9].

Data Mining Algorithms

In this study we presents 3 algorithms of data mining to choose the best algorithm gives low false positive rate (FPR) and high accuracy.

K-Nearest Neighbors (KNN)

This approach is a simple algorithm that saves each available instance and classifies new situations according to a similarity measure. This algorithm is an instance based learning that performs its classification according to a measure of similarity, such as distance functions in Manhattan, Euclidian and Minkowski, equations for this distance as follows:

Manhattan equation (1):

Euclidean equation (2):

Minkowski equation (3):

Naïve Bayes (Nb)

The NaïveBayes matrix provides a simple approach, with clear connotations, to represent and learn probabilistic knowledge. It is referred to as naive due to the fact that it is based on two important abstract assumptions that predictive attributes are conditioned conditional on class, and that there are no hidden or latent features that affect the prediction process[12].

Random Forest (RF)

Random forest technique [82] is a group of regression trees (or classification). These groups are effective if specific members are not alike, while random forests obtain a difference between separate trees using random suppliers: first, each tree is created on individual examples of training data; secondly, only a selected subset of features The data under consideration in each decade of the contract in the construction of separate trees.

Literature Review

Intrusion detection to the network is a "binary" classification task where infiltration detection is described as normal or abnormal. Data extraction techniques are useful for this type of intrusion detection because of their ability to identify small anomalies in large data sets. In this part of the literature review, we have reviewed some previous research relevant to our study.

Chand et al. [13] Comparing SVM collections with nine other classifiers, and evaluating the results across the NSL-KDD dataset, and concluded that the SVM + RF group is the better amongst the compare options. The results found for SVM , RF were: 97.5% accuracy, detection rate 93.49%, privacy 98.38%, accuracy 97.6%, and 97.6% recall.

McElwee [14] an effective method of detecting learning intrusion is proposed based on a random forest workbook and k-Means assembly. Daily events are presented to a Random Forest workbook and outputs that receive more than 95% of the KNN algorithm.

Garcia-Teodoro et al. [15] It is proposed that an anomaly-based network intrusion detection infrastructure consists of stages: parametrization, train, and detection. The survey focuses on commonly used intrusion detection techniques and systems that apply these technologies.

The questionnaire refers to the challenges faced by high-cost-time-swap questions and the use of such methods in high data rates.

Buczak and Guven [16] point to confusion in the literature of intrusion detection with respect to the use of data extraction concepts, automated learning and knowledge discovery in databases (KDD). They provide summary information on comparison metrics, data sets, and inputs (NetFlows, IP, TCP, UDP, ICMP headers), and pproaches based on data extraction and automated navigation techniques.

It also presents the time complexities and algorithm flow capabilities of some techniques commonly used to detect intrusion.

Folino and Sabatino [17] provides a brief overview of IDS-based methods, they focus on distributed and collaborative approaches.

Ahmed et al. [18] A survey was conducted on the techniques of detecting anomalies in 2016. The authors classified methods of detecting anomalies in: (1) statistics, (2) classification, (3) compilation, and (4) information theory. The fundamental role in this paper is to highlight the challenges and issues that are associated with the use of datasets which are commonly used for research on traffic abnormality.

Clustering approaches are often utilized for anomaly detection. Syarif et al. [18] He presented and discussed 5 approaches for detecting anomalies. The authors have utilized the NSL-KDD dataset to evaluate clustering algorithms in Detection Network Anomaly.

Lakhina [19] the suggest of the Network anomalies detecting technique using the entropy measures and comparing them with the results that have been obtained with the use of the scale. The research indicates that feature distributions of the network traffic are enriched with information which might become utilized to detection network anomaly.

Automated learning is a distinct section of AI in which the system learns how to classify data instance (note with a number of features) via training. Supervised learning is carried out through the training of an automated learning system with the use of historical data in which the expected output is known. Many supervised learning algorithms exist, such as Neuroscience, Gauss process regression, Bayesian statistics, Lazy learning, Vector support machine, Nearest neighborhood algorithm, Hidden Markov model, Bayes networks, Decision trees, Closest relative, Promotion of classification of works , Linear discrimination by Fisher, Naïve Bayesian, Perceptron, SVMs, quadratic works are several of the most widely used approaches of supervised learning.

Li and Wu [20] Provide an enhanced clustering approach based on entropy information and a sensitive frequency measurement. Those measurements are utilized for establishing the primary centers of groups. The authors used the KDD Cup 1999 dataset. The enhanced algorithm produces a 98.3% detection rate when tested with DOS attacks.

Bao et al. [21] Suggested approach to SVM in the network intrusion detection. They have stated that this algorithm is effective with a more sufficient learning capability for small samples. They put forward an NIDS model and principle based on vector support.

Kim and Kim [22] the use of randomized forest and an uncontrolled self-regulating map (SOM), Detection of attacks on networks. The Random Forest model is use as an abuse detection module and SOM model is used as a module for detecting anomalies. Except that the external values detected in the anomaly detection module are not sent to the module training to detect abuse. Network traffic is first categorized by using the Random Forest form. Then known attacks are detected at this stage, this approach achieved a better detection rate on NSL-KDD (a subset of KDD99) than the method suggested by Zhang and Zulkernine [23] (96.1% vs. 94.7%), but false positive rates were much higher (8.6% versus 2%).

Shafieian et al. [24] the suggestion of a method depends on a random forest workbook to detect a Slow Service Denial (DoS) attack on networks. The slow-read attack is launched by opening many HTTP connections and keeping them open by reading large contents slowly. To enforce a slow connection, an attacker can, for example, set specific TCP header values, like the size of the window. They used a custom data set with three basic features and could achieve AU values of more than 0.995, 99.6% accuracy, and false positive rates for AUC 0%.

Gupta and Kularya [25] suggest using Apache Spark to detect network penetration quickly and efficiently. On this structure, the authors compared the fulfillment of five works on KDD99 and NSL-KDD datasets. Random Forest Method achieve the best accuracy on both sets of data. The best overall resolution (92.13%) was found in the KDD990% data set.

Masarat et al. [26] Suggest a modified Random Forest method to detect sniffing, the differences are the following: (1) On each node, the selected variable is chosen to make the division of the weighted roulette wheel using the information gain values. (ii) Fuzzy logic is applied to Random Forest outputs that combine the result of classification and detection rates Error cost matrix. In the evaluation provided on the KDD99 data set, the approach provides a 94.4% accuracy.

Farnaaz and Jabbar [27] suggest using SU to select feature and Random Forest as a type of network intrusion detection. The tests on the NSL-KDD dataset indicated a 99.63% accuracy of the method.

Stefanova and Ramachandran [28] proposed a two-stage catalog for the detection of intrusions across the network, The first stage classifies traffic to "normal" and "attacks" classes. The attack movement is then sent to a second stage, which is classified into attack types. In the proposed method, the first classifier is a random forest and the second is a partial decision tree (Frank and Witten [29]).

Xun-Yi Ren et al. [30], introduced an intrusion detection system model with the advantage of multiclassification integration based on hadoop. They used a new classification map according to the rating centers. They then remove duplicate values to repair the new detection form. They used the KDD CUP99 data sets and their results from the huge data set test that showed that the molten classifier had more accuracy than just a classifier.

II. Datasets And Experiments

Datasets description

In this study, we use the NSLKDD dataset. NSL-KDD is an enhanced version of the KD99 Cup99 dataset, that has the issue of a large number of repetitive records [5], (42) attributes of NSL-KDD datasets include can show simple of attributes in table (1).

As listed in Table 3, NSL-KDD includes 41 packages where every one of those packages is categorized into one of the following 4 classes. Moreover, only a group of those attacks is presented in the test set.

Table 1: simple of datasets attributes

type	duration	protocol type	service	flag	src bytes
normal.	0	tcp	http	SF	181
normal.	0	tcp	http	SF	217
smurf.	0	icmp	ecr_i	SF	1032
neptune.	0	tcp	private	S0	0
normal.	2	tcp	smtp	SF	1572
normal.	2	udp	domain_u	SF	93
type	duration	protocol type	service	flag	src bytes

III. Data Sets Attacks Types

The NSL_KDD data-set can be classified to 4 types of attack:

Denial-of-service attack (DoS): Is a category of attack in which the attacker makes some computer or memory resources busy or extremely filled to not respond to requests, for example, neptune, smurf, teardrop, back, pod and land.

User to Root Attacks (R2L): Is a class of attacks in which the attacker gets arrivel to a plain user account on the system for obtaining root user arrivel to the system later than, for example. Guess_passwd, warezclient, ftp_write, warezmaster, phf, spy, multihop, and imap.

Probing (Probe): Is a category of attacks in which an attacker is look about a network for some data about potential network vulnerabilities, for example. satan, ipsweep, portsweep and nmap.

Remote to User Attacks (U2R): Is a category of attack in which the attacker sends some packets to a system via network, and after that get several information about possibility secured vulnerabilities in this system, for example. buffer_overflow, rootkit, loadmodule and perl [31].

Weka

In this section we presents the tool that can be utilized WEKA, is a collecting of machine learning models for tasks of data mining. It include tools that are used for data elaboration, visualization, classification, regress, cluster, and associate rules mining; establish only on New Zealand islands. The name is clear like this, WEKA is an open source software which has been issued under the License of GNU, [32].

Experiments

In this section, we determine the experiments of the algorithms in the datasets, in this study we using WEKA to explain how the Networks Intrusion Detection Systems that can protection the networks from attacks and send alarms to administrator that attacks may be occur by using data mining algorithms.

Evaluation Metrics

The efficiency of every workbook is evaluated based on accuracy, false positive rate and region under the ROC curve (AUC) and execution time. A good IDS system must accomplish a high level of precision with a low false positive rate. Accuracy is calculated by equation (1):

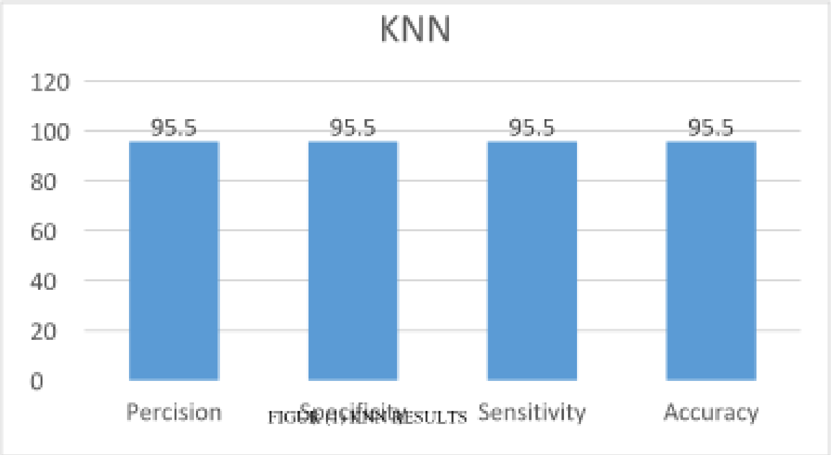
The true positive (TP) is the number of properly categorized attack records; (TN) true negative is the number of normal traffic records that are correctly categorized; false positives (FP) is the number of normal traffic records that are wrongly classified and False Negatives (FN) Cases of attack logs which are classified as false. A false positive rate (FPR) is calculated by equation (2):

IV. Experimental Results

In this part, we presents the result from the experiments that can be show in figure below, from figure (1) and table (2) that explain the confusion matrix, we can show the accuracy of the KNN algorithm the accuracy (95.5) and the FPR (0.004).

From figure (2) and table (3) we can show the precision of NB algorithm the accuracy (90.2) and the FPR (0.08).

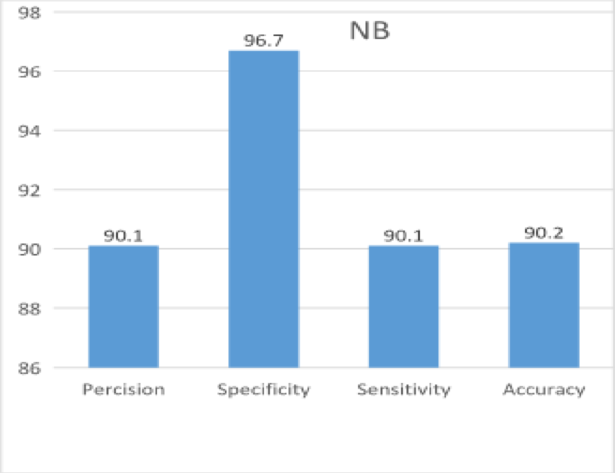
From figure (3) and table (4) we can show the precision of NB algorithm the accuracy (99.8) and the FPR (0.0002).



Figur (1) KNN results

Table 2: confusion matrix of KNN

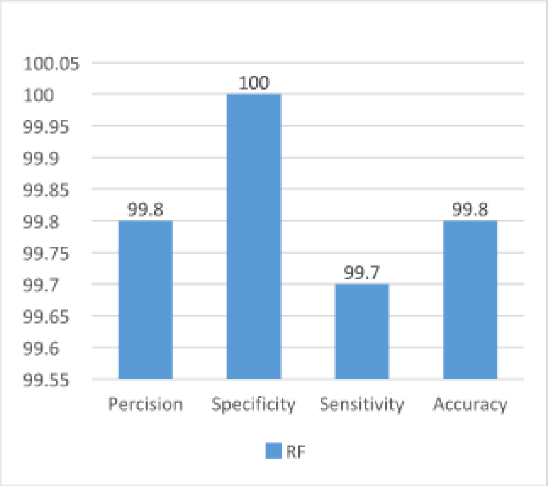
		Predicated	
		0	1
Actual	0	4041	18
	1	21	3478



Figur (2) NB results

Table 3: confusion matrix of NB

		Predicated	
		0	1
Actual	0	3718	341
	1	409	3090



Figur (3) RF results

Table 4: confusion matrix of RF

		Predicated	
		0	1
Actual	0	4054	5
	1	10	3489

V. Conculsion

In order to contribute to this paper, we first conducted a recent survey of recent studies on the detection of sniffing on the net, which was evaluated using the NSL_KDD data set. We then use WEKA to show a comparison of the wide performance between the ultimate popular classification algorithms. Finally, the Random Forest (RF) is the best models are proposed for algorithms selection with high accuracy and low FPR.

References

[1]. J. Cui, Y. Zhang, Z. Cai, A. Liu, and Y. Li, "Securing Display Path for Security-Sensitive Applications on Mobile Devices," *C. Comput. Mater. Contin*, vol. 55, pp. 17–35, 2018.

[2]. A. Pradeep, S. Mridula, and P. Mohanan, "High security identity tags using spiral resonators," *Comput. Mater. Contin.*, vol. 52, no. 3, pp. 185–195, 2016.

[3]. R. Bace and P. Mell, "NIST special publication on intrusion detection systems," BOOZ-ALLEN AND HAMILTON INC MCLEAN VA, 2001.

[4]. M. Roesch, "Snort: Lightweight intrusion detection for networks.," in *Lisa*, 1999, vol. 99, no. 1, pp. 229–238.

[5]. M. Lichman, "Datasets | Research | Canadian Institute for Cybersecurity | UNB," 2017. [Online]. Available: https://www.unb.ca/cic/datasets/index.html. [Accessed: 29-Dec-2018].

[6]. U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The KDD process for extracting useful knowledge from volumes of data," *Commun. ACM*, vol. 39, no. 11, pp. 27–34, 1996.

[7]. A. K. Ghosh, A. Schwartzbard, and M. Schatz, "Learning Program Behavior Profiles for Intrusion Detection.," in *Workshop on Intrusion Detection and Network Monitoring*, 1999, vol. 51462, pp. 1–13.

[8]. S. Kumar, "Classification and detection of computer intrusions." PhD thesis, Purdue University, 1995.

[9]. A. Demiriz, K. P. Bennett, and J. Shawe-Taylor, "Linear programming boosting via column generation," *Mach. Learn.*, vol. 46, no. 1–3, pp. 225–254, 2002.

[10]. Thomas H. Davenport, "Machine Learning What it is and why it matters," *sas*. .

[11]. Andrew Ng, "Unsupervised Feature Learning and Deep Learning." [Online]. Available: http://openclassroom.stanford.edu/MainFolder/CoursePage.php?course=ufldl. [Accessed: 05-Jan-2019].

[12]. G. H. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," in *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, 1995, pp. 338–345.

[13]. A. Chakrabarti and G. Manimaran, "Internet infrastructure security: A taxonomy," *IEEE Netw.*, vol. 16, no. 6, pp. 13–21, 2002.

[14]. S. McElwee, "Active learning intrusion detection using k-means clustering selection," in *SoutheastCon, 2017*, 2017, pp. 1–7.

[15]. P. Garcia-Teodoro, J. Diaz-Verdejo, G. Maciá-Fernández, and E. Vázquez, "Anomaly-based network intrusion detection: Techniques, systems and challenges," *Comput. Secur.*, vol. 28, no. 1–2, pp. 18–28, 2009.

[16]. A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Commun. Surv. Tutorials*, vol. 18, no. 2, pp. 1153–1176, 2016.

[17]. G. Folino and P. Sabatino, "Ensemble based collaborative and distributed intrusion detection systems: A survey," *J. Netw. Comput. Appl.*, vol. 66, pp. 1–16, 2016.

[18]. M. Ahmed, A. N. Mahmood, and J. Hu, "A survey of network anomaly detection techniques," *J. Netw. Comput. Appl.*, vol. 60, pp. 19–31, 2016.

[19]. A. Lakhina, M. Crovella, and C. Diot, "Mining anomalies using traffic feature distributions," in *ACM SIGCOMM Computer Communication Review*, 2005, vol. 35, no. 4, pp. 217–228.

[20]. H. Li and Q. Wu, "Research of clustering algorithm based on information entropy and frequency sensitive discrepancy metric in anomaly detection," in *Information Science and Cloud Computing Companion (ISCC-C), 2013 International Conference on*, 2013, pp. 799–805.

[21]. J. Song, H. Takakura, Y. Okabe, M. Eto, D. Inoue, and K. Nakao, "Statistical analysis of honeypot data and building of Kyoto 2006+ dataset for NIDS evaluation," in *Proceedings of the First Workshop on Building Analysis Datasets and Gathering Experience Returns for Security*, 2011, pp. 29–36.

[22]. E. Kim and S. Kim, "A novel hierarchical detection method for enhancing anomaly detection efficiency," in *Computational Intelligence and Communication Networks (CICN), 2015 International Conference on*, 2015, pp. 1018–1022.

[23]. J. Zhang and M. Zulkernine, "A hybrid network intrusion detection technique using random forests," in *null*, 2006, pp. 262–269.

[24]. S. Shafieian, M. Zulkernine, and A. Haque, "CloudZombie: Launching and detecting slow-read distributed denial of service attacks from the cloud," in *Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing (CIT/IUCC/DASC/PICOM), 2015 IEEE International Conference on*, 2015, pp. 1733–1740.

[25]. G. P. Gupta and M. Kulariya, "A framework for fast and efficient cyber security network intrusion detection using apache spark," *Procedia Comput. Sci.*, vol. 93, pp. 824–831, 2016.

[26]. S. Masarat, S. Sharifian, and H. Taheri, "Modified parallel random forest for intrusion detection systems," *J. Supercomput.*, vol. 72, no. 6, pp. 2235–2258, 2016.

[27]. N. Farnaaz and M. A. Jabbar, "Random forest modeling for network intrusion detection system," *Procedia Comput. Sci.*, vol. 89, pp. 213–217, 2016.

[28]. Z. Stefanova and K. Ramachandran, "Network attribute selection, classification and accuracy (NASCA) procedure for intrusion detection systems," in *Technologies for Homeland Security (HST), 2017 IEEE International Symposium on*, 2017, pp. 1–7.

[29]. E. Frank and I. H. Witten, "Generating accurate rule sets without global optimization," 1998.

[30]. X.-Y. Ren and Y.-Z. Qi, "Hadoop-based Multi-classification Fusion for Intrusion Detection," *J. ofApplied Sci.*, vol. 13, no. 12, pp. 2178–2181, 2013.

[31]. P. G. Jeya, M. Ravichandran, and C. S. Ravichandran, "Efficient classifier for R2L and U2R attacks," *Int. J. Comput. Appl.*, vol. 45, no. 21, p. 29, 2012.

[32]. "Weka 3 - Data Mining with Open Source Machine Learning Software in Java." [Online]. Available: https://www.cs.waikato.ac.nz/ml/weka/. [Accessed: 10-Jan-2019].

Dr. Sefer Kurnaz. " Comparative study for Networks Intrusion Detection (NID) By Using Various Classification Algorithms to protection the Networks Systems from attacks (Comparison Studies)" IOSR Journal of Computer Engineering (IOSR-JCE) 21.1 (2019): 41-47