Journal of Theoretical and Applied Information Technology
15th March 2020. Vol.98. No 05

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN EFFICIENT INTRUSION DETECTION APPROACH USING LIGHT GRADIENT BOOSTING

HAYEL KHAFAJEH

Faculty of Information Technology, Zarqa University, Zarqa, Jordan

E-mail: hayelkh@zu.edu.jo

ABSTRACT

Nowadays, network security has been received more attention from researchers. Intrusion detection systems (IDSs) serves as an essential element of network security. In order to increase the network's security, machine-learning algorithms may be utilized for the detection and prevention of the attacks that launched against the network. The researcher of this study used LightGBM's algorithm for training a model in order to detect several types of network attacks. The proposed approach was compared with classical machine learning in terms of performance on the same dataset. The experimental results show that the proposed approach achieves a detection rate of 97.4% with a false-positive rate of 0.9%.

Keywords: *Network security, IDS, Machine learning, LGBM.*

1. INTRODUCTION

Regarding the Network Intrusion Detection Systems (NIDS) [1-9], it plays an important role in protecting computer systems by preventing the malicious network-based attacks. Such attacks shall lead to the disruption of the services provided by the system. The powerful provision of robust NIDS is considered a difficult mission. That is attributed to several factors. Much growth is achieved through internet traffic. Such traffic consists of many data types that traverse the network. In addition, the NIDS must be capable of analyzing those enormous volumes of traffic. The NIDS must be capable of differentiating between malicious and legitimate behaviours with showing an acceptable degree of accuracy. The NIDS systems have been categorized into two major kinds. These kinds are 1)The misuse-based system (maybe named signature-based); 2)The anomaly-based system[10, 11].

The first kind relies much on an extensive database. This database consists of attack signatures. Each one of the signatures involves a set of rules that corresponds to the types of attacks which have happened earlier. With the latest version of the NIDS signature, a system that is misused-based is effective in detecting of previous attacks. Despite that, this system shows the vulnerability of the 0-day attack. It's deemed as requiring much time for processing. The anomaly-based IDS system aims at detecting the computer system attacks. That is done through the observation of abnormal traffic patterns or statistics. Also, the system can be employed for the detection of a 0-day attack.

Regarding the results obtained from discovering the attack, they can be updated the database for detection in the future by using a system that's misused-based. On the other hand, an anomaly-based system has a weakness which is showing a high rate value of false alarm. That is because much regular traffic that shows an unusual behaviour can trigger the system's alarm. In a practical system, a hybrid of anomaly-based and misused-based systems are usually utilized for mitigating the influence of misused-based and the influence of the 0-day attack.

Recently, the anomaly-based NIDS utilizes machine learning methods has received much attention. Many classification models are used for differentiating between suspicious traffic and the normal one. Despite that, the feasibility of that approach is considered low due to showing a low level of performance in terms of detection. That may be attributed to several reasons. Such reasons may include: 1) the nature of traffic is diverse; 2) the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

imbalance traffic classes; 3) having feature selection processes that are ineffective.

For overcoming these limitations, several studies employ the Deep Neural Network (DNN) methods. Such methods may include the recurrent neural network (RNN) [12]. Such methods show improved performance in detection. However, it still requires a large volume of train dataset that is effective. It also requires a significant training time

In this study, the researcher focused on the problem faced by the NIDS binary classification at which the system seeks differentiating between the normal activities and the ones that attack. The researcher of this study aimed to develop an intrusion detection system based on anomaly behavior. This system is based on a LightGBM classification model. The reason behind selecting this model is attributed to show high-performance level, being accessible, various selections and fast implementation of hyper-parameters.

Section 2 of this paper includes a review of the literature related to the NSL-KDD dataset. The proposed system is described in the third section. The fourth section presents the discussion. The fifth section presents the conclusion

2. LITERATURE REVIEW

Many scholars employed a system for intrusion detection that is based on a supervised learning approach. Such approaches may include: 1) the neural network (NN)-based approach and the support vector machine (SVM)-based approach.

Ibrahim et al.[13] conducted a comparison between KDD99 dataset from one hand and NSL-KDD dataset based on Self Organization Map (SOM) from another hand. They utilized an artificial neural unsupervised network in a hierarchical anomaly intrusion detection system. SOM neural nets were employed for detection. They were also employed for the separation of the attacking traffic from the normal one. The latter study assessed the effectiveness of SOM in detecting anomaly intrusion.

The study of Lakhina et al. [14] focused on feature reduction through conducting a component analysis for the anomaly-based applicable NSL-KDD dataset. Those researchers decreased the number of the features that are in the NSL-KDD dataset and deemed as redundant and irrelevant for carrying out the operation of the anomaly detection. Those researchers implemented a principal component analysis that is hybrid and used a neural network algorithm for the detection of attacks effectively.

Revathi et al. [15] conducted an analysis for the NSL-KDD dataset. That was done through using several machine learning methods for the system of detecting intrusion. The latter analysis concentrated on specific NSL-KDD datasets for analyzing several techniques of machine learning. The algorithm of the RF classification shows an accuracy rate that's the highest in comparison to other targeted algorithms of classification.

In [16] the authors investigated the NSL-KDD dataset for the system of detecting intrusion. That is done by applying various algorithms of classification. They conducted an analysis of the NSL-KDD dataset. That was done for examining the effectiveness of algorithms of classification in the detection of anomaly network traffic. In addition, the authors investigated the relationship between the protocols which are within the network protocol stack from one hand and the attacking traffic from another hand. That was done for generating anomalous network traffic.

Chae et al. [17] investigated feature selection in order to detect the intrusions. That was done by employing the NSL-KDD dataset. Besides, they identified several input features that are important in the building of IDS. The performance level of the standard methods of feature selection was assessed.

Sadek et al. [18] investigated an anomaly IDS. That was done based on a new hybrid algorithm. This algorithm is called (the neural network) with utilizing an Indicator Variable and employing a rough set for reducing attributes (NNIV-RS). As for the results of the latter study, they indicate that the proposed algorithm of the NNIV-RS shows robust and better data representation. The proposed algorithm is capable of reducing unnecessary features for improving the IDS reliability level and efficiency level.

Ingre et al. [19] aimed to analyses the performance of the NSL-KDD dataset through the use of (ANN). The results that were obtained are based on various measures for assessing the performance of the binary class and the 5-class classification on attacking types. Regarding the accuracy level of ANN, it is provided.

Ray. [20] explored the impacts of the neural network structure on the level of performance of the (IDSs). The researcher formed an equation to find the best of hidden neurons number an MLFFNN network IDS. The latter equation could be utilized in determining how many hidden neurons are there. That

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

is done for eliminating the error calculations and lengthy trials in the MLFFNN case.

In 2016, Hussain & Lalmuanawma [21] The JRip and OneR can be listed as algorithms for association rule mining. The processing speed of JRiP is high and accurate, while OneR produces a single rule on all features and selects a minimum error rule. The combination of SVM with JRip, thereby achieves high accuracy and low false-positive values. 97.2% and 2.7% respectively, whereas the OneR indicates 91.7% and 8.2% with accuracy and low false positives, respectively.

Chauhan et al. [22] have compared the performance of the ten best classification method including C4.5, Bayesnet, Random Forest, Random Tree and REPTree to Stochastic Gradient Descent (SGD), IBK, JRip, PART. The studies of the authors have shown that the Random Forest, which is a collection of decision tree classifiers, has surpassed other methods for Accuracy, Sensitivity, and Specificity.

Garg and Khurana[23] carried out comparative analyzes using WEKA with 41 attributes, using various classification algorithms for the NSL-KDD dataset. 94,000 instances have been used from the KDD dataset for training data, and over 48,000 have been used as a test dataset. In addition, Garrett's ranking technique has been utilized for classifying various classification according to their performance. The methodology of rotation forest classification was better than the rest. They tested 45 classification methods for the dataset and achieved the best possible accuracy from the Rotation Forest technique with 96.4%.

In the present study, the researcher employed the NSL-KDD dataset for assessing the proposed models' accuracy. The NSL-KDD dataset is the enhanced version of the popular IDS benchmarking dataset KDDCUP'99. It has been employed in several researchers since the time it was introduced at [24-26]. In the research, an anomaly-based network system for detecting the intrusion is proposed by employing the LightGBM algorithm.

3. THE PROPOSED APPROACH

The main goals of the proposed method are 1) The detection of malicious activities that are insensible; 2) The detection of malicious activities without having to carry out a deep packet inspection; 3) the primary elements of the proposed detection scheme are displayed through Figure 1.

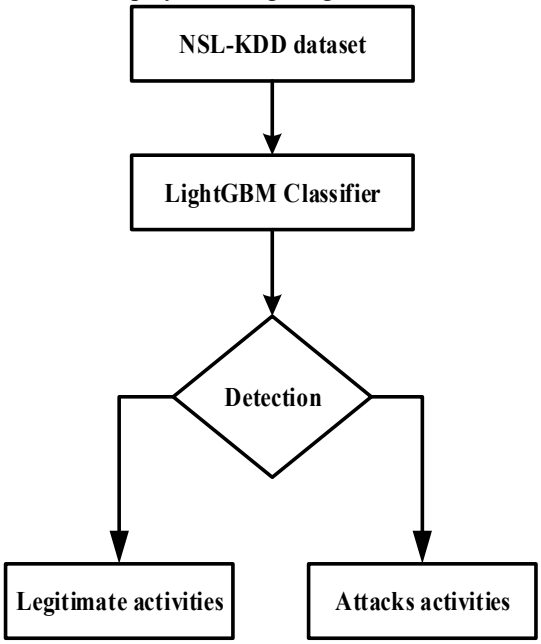


Figure 1. The proposed Intrusion detection approach

The Gradient Boosting Decision Tree (GBDT) [27] showed much success in many applications. Regarding the GBDT, it is an ensemble model of decision trees. Within the iteration, GBDT learns the trees of decision. That is done by fitting the negative gradients. Despite that, with the development of data, that's big, the efficiency level and accuracy level of GBDT are faced with difficulties and challenges. For instance, the computational complexities of GBDT are considered as proportionate to the number of instances and features. That leads to having several calculations that are time-consuming. For solving those challenges, the method of LightGBM was proposed [28]. This method is a framework for boosting the gradient, which is distributed and based on a tree of decision for the implementation of the GBM.

In comparison with other GBMs, the method of LightGBM has made some optimization. It is an algorithm that itis based on a histogram-based tree of the decision and uses the subtraction of histogram for the purpose of acceleration. It contributed to the optimization of sparse features through the use of the

method that is based on the histogram.

Leaf-wise leaf growth strategy with depth limitation has been adopted. That shall reduce the number of errors. It shall raise the level of accuracy. Regarding the depth limitation of the Leaf-wise, it can ensure having a high-efficiency level. It is capable of preventing the over-fitting at the same time. The rate of the cache hit was optimized, and the multi-threaded was optimized.

LightGBM has added the rules of decision to the features of category. That is done to avoid additional computational and memory overhead. It's done through the conversion of features into a one-hot multi-dimensional feature.

LightGBM is a new Gradient Boosting Decision Tree algorithm, introduced by Ke and colleagues in 2017, which has been used in many filed of data mining problem, such as classification, regression and ordering [29]. Two new techniques are included in the LightGBM algorithm, which includes the one-sided gradient analysis and the exclusive features bundling.

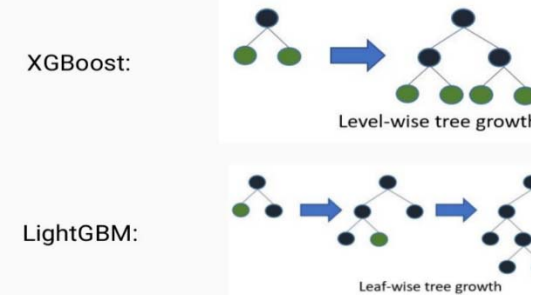
Given the supervised training set$X = \{(xi, yi)\}_i^n = 1$, the target of LightGBM is to find an approximation f(x) to a certain function $\hat{f}(x)$ which reduces the expected loss function value$L(y, f(x))$ as follows:

$$\hat{f} = \arg\min Ey, xL(y, f(x)) \quad (1)$$

LightGBM integrates a number of T regression trees $\sum_{t=2}^{T} f_t(x)$ to approximate the final model, which is

$$f_T(X) \sum_{t=2}^{T} f_t(x) \quad (2)$$

The regression trees could be represented as



$q(x), q \in \{1, 2, \ldots, J\}$, where J denotes the leaves number, q represents the rule of decision tree, and w is a vector of leaf nodes weighs. Hence, LightGBM would be additively trained at step t as follows:

$$T_t = \sum_{i=1}^{n} L(yi, f_{t-1}(xi) + f_t(xi)) \quad (3)$$

In LightGBM, Newton's method easily approximates the objective function.

Where gi and hi indicate the first- and second-order gradient statistics of the loss function, let Ij show the example set of leaf$j$.

$$T_t = \sum_{i=1}^{n}((\sum_{i\in Ij} g_i) + \frac{1}{2}\left(\sum_{i\in Ij} h_i + \lambda\right) w_j^2)) \quad (4)$$

For the tree structure$q(x)$, the optimum leaf weight score of each leaf node w^* and the extreme value of Tt could be solved as follows:

$$w_j^* = -\frac{\sum_{i\in Ij} g_i}{\sum_{i\in Ij} h_i + \lambda) w_j^2} \quad (5)$$

LightGBM employs a one-sides-sampling (GOSS) approach to detect split value in data instances while XGBoost utilizes pre-sorted algorithms & histogram-based algorithms to compute the best split point.

Simply speaking, the histogram-based algorithm separates all data points in discrete cases for an element and uses them to identify the splitting point of the histogram. Although it is efficient than the pre-sorted speed algorithm, which enumerates all possible split points on the pre-sorted feature value, in terms of speed, it remains behind GOSS. Figure 2 shows a comparison between the procedure work of XGBoost and LighGBM.

Figure 2. Leaf wise tree growth in XGBoost and LightGBM.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4. RESULTS AND EXPERIMENT

4.1. Dataset

The dataset of KDD-Cup 99 [30] is used much for testing the performance level of a system for the detection of anomaly-based intrusion. [30] shed light on two critical issues. That was done by relying on a statistical analysis that is conducted for the dataset. That led to having prediction results that are over-simplistic. For circumventing that problem, the latter scholars proposed the dataset of NSL-KDD. The latter dataset possesses the advantages listed below. These advantages are listed in comparison with KDD-Cup 99:

1. In NSL-KDD, many duplicates and redundant data that is faced in the KDD-Cup 99 are eliminated from the concerned datasets.
2. For providing an assessment -that is more accurate- for various techniques of learning, the number of the selected records from each group of difficulty-level is deemed inversely proportionate to the proportion of records within the original KDD dataset. Such a dataset has been categorized into four datasets. These 4 datasets are:

A. KDDTrain+: It's the overall train dataset. It consists of 125, 973 records.
B. KDDTrain+ 20Percent: It is the training dataset that consists of twenty percent of the overall train dataset. In fact, it possesses 25,192 records.
C. KDDTest+: It is a test dataset that consists of 22,544 records.
D. KDDTest-21: It consists from 11,850 records. It can be obtained through the application of the 21 models of machine learning on the KDDTest+ dataset. That is done for the prediction of the label of the dataset, which may be normal or attack label. These labels are predicted accurately by the 21 models. They get excluded from the dataset.

The NSL-KDD seeks to categorize the attacks into 4 types. These types are Denial-of-Service,

Probe, Root to Local, and Unauthorized to Root. As showing in Table 1.

Each kind of attack is displayed below:

1. (DoS): It overwhelms the concerned resources (Network, CPU or Memory). Thus, the typical operation shall not be operated as it is expected. Regarding the attack samples, they may include: transmitting a significant number of the packets to the server that is targeted. In this case, normal users shall not be capable of accessing it.
2. Probe: It may include port scanning for the identification of the vulnerabilities that are within the computer systems for other attacks.
3. R2L: The ones launching an attack attempt accessing the computer resources that are unauthorized for modifying or destroying the operations carried out by the computer systems that are targeted
4. U2R: The ones launching an attack attempt gaining access to unauthorized resources through the use of root privileges.

The normal\attack ratio of each kind of dataset within NSL-KDD is listed in Table 2. The NSL-KDD seeks to categorize the attacks into four types. These types are Denial-of-Service, Probe, Root to Local and Unauthorized to Root. That is displayed through the first table. Each kind of attack may be illustrated below in Table 2.

Table 1. Type of Attacks [30]

Category	Attacks
DoS (Denial of Service)	Neptune, pod, smurf, teardrop, prosess table, warezmaster, aoache2, mail bomb,back
Probe	Multihop,http tunnel, ftp_write, root kit, ps buffer overflow, xterm
R2L (Root to Local)	Named, snmpgetattack, xlock, send mail, guess_passwd
U2R (Unauthorized to Root)	Ipsweep, nmap, port sweep, satam, mscan, saint

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 2. List the percentage of legitimate /attacks for each dataset in NSL-KDD. [30]

Dataset	Data type	legitimate	malicious
KDDTrain+	Number	67343	58630
KDDTest+	Number	9711	12833
KDDTest-21	Number	4342	7508

5. ASSESSING THE METRICS AND THE RESULTS

In the present study, the researcher assessed the performance level of the introduced model. That was done in terms of Accuracy, True Positive Rate (TPR), or Recall, False Positive Rate (FPR). The researcher employed (AUC) for the total measure of the performance level across all possible thresholds of classification.

1. Accuracy metric equation is presented below

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}} \quad (6)$$

Where TP, TN, FP, and FN represent True Positive, True Negative, False Positive and False Negative, respectively.

2. Rate of Detection (DR) (also named recall). It indicates the percentage of the malicious instances which have been predicted as malicious instances.

$$\text{DR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (7)$$

3. False-positive rate (FPR): It refers to the ratio of the items which have been classified in an incorrect manner as an attack to all the items that belong to normal. The equation of this rate is

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \quad (8)$$

4. Precision: It stands for the percentage of instances that are classified correctly as being a positive instance.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (9)$$

5. The area under the Curve (AUC): It is a performance measurement for the problems of classification in various settings of thresholds. The receiver operating characteristic (ROC) is a probability curve. AUC stands for measure or degree of separability. It informs one about the capability of the model to distinguish between classes. A higher value of the AUC, better the model is at distinguishing between the patients that are with the disease and the ones that are without.

In our experiment, LightGBM algorithm has several parameters that used to tune the algorithm, such as type of boosting, maximum depth, rate of learning, leaves number fraction of features, maximum depth and number of iterations. As a type of boosting, we selected the Gradient boosting decision tree (GBDT). The value of parameters used to tune the LightGBM is listed in Table 3.

Table 3. Parameters of LightGBM algorithm

Parameter	Value
Boosting_type	Gbdt
Objective	Binary
Evaluation_Metric	'binary_logloss', 'auc'
Learning_rate	0.1
Number_of_leaves	200
Feature_fraction	0.64
'bagging_fraction'	0.8
'bagging_freq'	1
Maximum_depth	5
Num_of_ boosting_iterations	100

The results show that the proposed technique archives the highest rate of accuracy. The accuracy rate is 98.3 % approximately. It is presented in Table 2; it can be noticed that the proposed technique shows

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

the highest DR, AUC, and precision. That is represented in 0.983, 0.974 and 0.981, respectively. The proposed technique shows the lowest FPR, which is represented in 0.009.

Table 4. LightGBM Results

Evaluation Matrices	LightGBM
ACC	0.983
AUC	0.981
DR (recall)	0.974
FPR	0.009
F1-Score	0.982
Precision	0.989

We conduct our experiments using machine learning algorithms that used for IDS Random Forest Classifier [31] , Bagging Classifier [32], Decision Tree Classifier [33], Extra Tree Classifier [35 ,34], Gradient Boosting Classifier [36], SVC[37], K-Neighbors Classifier [38], AdaBoost Classifier[39], Linear SVC[40], Logistic Regression CV [41], Ridge Classifier CV [42], Perceptron [43], BernoulliNB[44] , Passive Aggressive Classifier [45], SGD Classifier [46], GaussianNB [47].

Table 5 shows the comparison results of several machine-learning algorithms. The accuracy, Precision, detection rate (DR) and area under roc (AUC) are comparison based on NSL-KDD dataset. The result shows that highest classification accuracy, DR and AUC is obtained by Decision Tree Classifier; which achieved 0.7994, 0.6684 and 0.8205 respectively.

Table 5. Comparison between several algorithms of classification

MLA Name	Accuracy (Test dataset)	Precision	DR	AUC
Decision Tree Classifier	0.7994	0.9699	0.6684	0.8205
Bagging Classifier	0.7980	0.9709	0.6652	0.8194
Random Forest Classifier	0.7856	0.9688	0.6440	0.8083
AdaBoost Classifier	0.7795	0.9328	0.6602	0.7987
KNeighbors Classifier	0.7731	0.9639	0.6249	0.7970
Bernoulli NB	0.7678	0.9668	0.6132	0.7927
Linear SVC	0.7100	0.8016	0.6518	0.7193
Logistic Regression CV	0.7013	0.8776	0.5524	0.7253
Ridge Classifier CV	0.7011	0.9593	0.4960	0.7341
Passive Aggressive Classifier	0.6832	0.8482	0.5401	0.7062
Gaussian NB	0.4503	0.9366	0.0369	0.5168
Perceptron	0.3799	0.4680	0.6529	0.3360
SGD Classifier	0.3609	0.4549	0.6192	0.3193
Proposed approach	**0.983**	0.989	0.974	0.981

Table 6. A Comparison with other works

Research work	Classifiers	Accuracy on NSL-KDD Dataset (%)
Hussain et al.[21]	Decision Table RBF Classifier SVM	97.5 96.7 92.9
Chauhan et al [22]	Logistic regression SGD	97.269 97.443
Garg et al [23]	Random Forest	96.12
Proposed approach	**LightGBM**	**98.3**

The results presented through Table 6 are the appropriate comparisons with other research works. Constant comparison is always made between several methods of classification. Scholars have been struggling much for finding the optimum method of classification. Recently, scholars contested the application of LightGBM and the method of the RF classification on datasets. They did that for finding a better model of classification. The results give an advantage to the LightGBM over the RF classification method [48]. LightGBM can, in parallel, classify several systems that are operating [49], and by setting the parameters, the desired algorithm that is based on decision tree may be run on LightGBM, which can be the algorithm that's desired in addition to the boosting made to it. Thus, it stands out. It should be deemed as a significant player in the classification field.

In comparison with traditional Boosted models, LightGBM raises the speed ten times. Hence, the mode of classification shall become better. In a network model, speed plays a role that's important [49]. LightGBM is very flexible. It does not rely on the kind of data that is entering a network. It is capable of converting the entering data quickly into a numeric type. From there, it is capable of classifying the data with showing proper levels of speed and good levels of accuracy.

The kind of entries of data in the IDS is related in a direct manner to the kind of environment that the IDS is deployed in. That's where LightGBM obtains an advantage as it shows flexibility in its implementation. LightGBM can take several types of data as the entries of input and offers the choice of running simultaneously on different systems that are operating. Such systems may include Windows and Linux. In addition, data overfitting problems are well handled in the LightGBM. Also, the LightGBM offers much flexibility in the application of the Algorithms based on three decisions and Model Solvers that are linear.

6. CONCLUSION

LightGBM algorithm was designed for performance and speed learning. This algorithm was based on gradient-boosted decision trees algorithms. Moreover, LightGBM may be deployed in various environments of computing. The proposed system utilizes the LighGBM algorithm for achieving high rates of accuracy with 98.3% rates. Besides, we provide a comparison of several machine-learning algorithms with the LightGBM classifier in the context of intrusion detection. The researcher made a plan for extending the approach through the application of a deep learning approach with unsupervised learning. That was done to enhance the levels of performance and accuracy throughout time.

ACKNOWLEDGEMENTS

This research has been supported by Zarqa University, Jordan.

REFERENCES

[1] R. KUMAR, RITURAJ, S. M. R, R. KUMAR, RITURAJ, And S. M. R, "Using Data Mining And Machine Learning Methods For Cyber Security Intrusion Detection," *INTERNATIONAL JOURNAL OF RECENT TRENDS IN ENGINEERING & RESEARCH,* Vol. 3, No. 4, 2017, Doi: Https://Doi.Org/10.23883/IJRTER.2017.3117.9NWQV.

[2] M. Alauthman, N. Aslam, M. Al-Kasassbeh, S. Khan, A. Al-Qerem, And K.-K. Raymond Choo, "An Efficient Reinforcement Learning-Based Botnet Detection Approach," *Journal Of Network And Computer Applications,* Vol. 150, P.

102479, 2020/01/15/ 2020, Doi: Https://Doi.Org/10.1016/J.Jnca.2019.102479.

[3] M. Alauthaman, N. Aslam, L. Zhang, R. Alasem, And M. A. Hossain, "A P2P Botnet Detection Scheme Based On Decision Tree And Adaptive Multilayer Neural Networks," *Neural Computing And Applications,* Vol. 29, No. 11, Pp. 991-1004, 2018/06/01 2018, Doi: 10.1007/S00521-016-2564-5.

[4] A. Ammar, A. Mohammad, A. Firas, O. Dorgham, And O. Atef, "An Online Intrusion Detection System To Cloud Computing Based On Neucube Algorithms," *International Journal Of Cloud Applications And Computing (IJCAC),* Vol. 8, No. 2, Pp. 96-112, 2018, Doi: 10.4018/IJCAC.2018040105.

[5] K. Alieyan, A. Almomani, M. Anbar, M. Alauthman, R. Abdullah, And B. Gupta, "DNS Rule-Based Schema To Botnet Detection," *Enterprise Information Systems,* Pp. 1-20, 2019.

[6] A. Alnawasrah, A. Alnomani, F. Meziane, And M. Alauthman, "Fast Flux Botnet Detection Framework Using Adaptive Dynamic Evolving Spiking Neural Network Algorithm," In *2018 9th International Conference On Information And Communication Systems (ICICS)*, 3-5 April 2018 2018.

[7] M. Almseidin, A. A. Zuraiq, M. Al-Kasassbeh, And N. Alnidami, "Phishing Detection Based On Machine Learning And Feature Selection Methods," *International Journal Of Interactive Mobile Technologies (Ijim),* Vol. 13, No. 12, Pp. 171-183, 2019.

[8] M. Alkasassbeh, "An Empirical Evaluation For The Intrusion Detection Features Based On Machine Learning And Feature Selection Methods," *Arxiv Preprint Arxiv:1712.09623,* 2017.

[9] M. Almseidin, I. Piller, M. Al-Kasassbeh, And S. Kovacs, "Fuzzy Automaton As A Detection Mechanism For The Multi-Step Attack," *International Journal On Advanced Science, Engineering And Information Technology,* Vol. 9, No. 2, Pp. 575-586, 2019.

[10] M. Alkasassbeh, "A Novel Hybrid Method For Network Anomaly Detection Based On Traffic Prediction And Change Point Detection," *Arxiv Preprint Arxiv:1801.05309,* 2018.

[11] M. Almseidin, M. Al-Kasassbeh, And S. Kovacs, "Detecting Slow Port Scan Using Fuzzy Rule Interpolation," In *2019 2nd International Conference On New Trends In Computing Sciences (ICTCS)*, 2019: IEEE, Pp. 1-6.

[12] C. Yin, Y. Zhu, J. Fei, And X. He, "A Deep Learning Approach For Intrusion Detection Using Recurrent Neural Networks," *Ieee Access,* Vol. 5, Pp. 21954-21961, 2017.

[13] L. M. IBRAHIM, D. T. BASHEER, And M. S. MAHMOD, "A COMPARISON STUDY FOR INTRUSION DATABASE (KDD99, NSL-KDD) BASED ON SELF ORGANIZATION MAP (SOM) ARTIFICIAL NEURAL NETWORK," *Journal Of Engineering Science And Technology,* Vol. 8, No. 1, Pp. 107-119, 2013.

[14] S. Lakhina, S. Joseph, And B. Verma, "Feature Reduction Using Principal Component Analysis For Effective Anomaly–Based Intrusion Detection On NSL-KDD," 2010.

[15] S. Revathi And A. Malathi, "A Detailed Analysis On NSL-KDD Dataset Using Various Machine Learning Techniques For Intrusion Detection," *International Journal Of Engineering Research & Technology (IJERT),* Vol. 2, No. 12, Pp. 1848-1853, 2013.

[16] L. Dhanabal And S. 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.

[17] H.-S. Chae, B.-O. Jo, S.-H. Choi, And T.-K. Park, "Feature Selection For Intrusion Detection Using Nsl-Kdd," *Recent Advances In Computer Science,* Pp. 184-187, 2013.

[18] R. A. Sadek, M. S. Soliman, And H. S. Elsayed, "Effective Anomaly Intrusion Detection System Based On Neural Network With Indicator Variable And Rough Set Reduction," *International Journal Of Computer Science Issues (IJCSI),* Vol. 10, No. 6, Pp. 227-233, 2013.

[19] B. Ingre And A. Yadav, "Performance

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Analysis Of NSL-KDD Dataset Using ANN," In *2015 International Conference On Signal Processing And Communication Engineering Systems*, 2015: IEEE, Pp. 92-96.

[20] L. Ray, "Determining The Number Of Hidden Neurons In A Multi Layer Feed Forward Neural Network," *Journal Of Information Security Research,* Vol. 4, No. 2, Pp. 63-70, 2013.

[21] J. Hussain And S. Lalmuanawma, "Feature Analysis, Evaluation And Comparisons Of Classification Algorithms Based On Noisy Intrusion Dataset," *Procedia Computer Science,* Vol. 92, Pp. 188-198, 2016.

[22] H. Chauhan, V. Kumar, S. Pundir, And E. S. Pilli, "A Comparative Study Of Classification Techniques For Intrusion Detection," In *2013 International Symposium On Computational And Business Intelligence*, 2013: IEEE, Pp. 40-43.

[23] T. Garg And S. S. Khurana, "Comparison Of Classification Techniques For Intrusion Detection Dataset Using WEKA," In *International Conference On Recent Advances And Innovations In Engineering (ICRAIE-2014)*, 2014: IEEE, Pp. 1-5.

[24] J. Kevric, S. Jukic, And A. Subasi, "An Effective Combining Classifier Approach Using Tree Algorithms For Network Intrusion Detection," *Neural Computing And Applications,* Vol. 28, No. 1, Pp. 1051-1058, 2017/12/01 2017, Doi: 10.1007/S00521-016-2418-1.

[25] M. Panda, A. Abraham, And M. R. Patra, "Discriminative Multinomial Naïve Bayes For Network Intrusion Detection," In *2010 Sixth International Conference On Information Assurance And Security*, 23-25 Aug. 2010 2010, Pp. 5-10, Doi: 10.1109/ISIAS.2010.5604193.

[26] I. Obeidat, N. Hamadneh, M. Alkasassbeh, M. Almseidin, And M. Alzubi, "Intensive Pre-Processing Of KDD Cup 99 For Network Intrusion Classification Using Machine Learning Techniques," 2019.

[27] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals Of Statistics,* Vol. 29, No. 5, Pp. 1189-1232, 2001. [Online]. Available: Http://Www.Jstor.Org/Stable/2699986.

[28] G. Ke *Et Al.*, "Lightgbm: A Highly Efficient Gradient Boosting Decision Tree," Presented At The Proceedings Of The 31st International Conference On Neural Information Processing Systems, Long Beach, California, USA, 2017.

[29] G. Ke *Et Al.*, "Lightgbm: A Highly Efficient Gradient Boosting Decision Tree," In *Advances In Neural Information Processing Systems*, 2017, Pp. 3146-3154.

[30] A. Pattawaro And C. Polprasert, "Anomaly-Based Network Intrusion Detection System Through Feature Selection And Hybrid Machine Learning Technique," In *2018 16th International Conference On ICT And Knowledge Engineering (ICT&KE)*, 21-23 Nov. 2018 2018, Pp. 1-6, Doi: 10.1109/ICTKE.2018.8612331.

[31] A. Subasi, E. Molah, F. Almkallawi, And T. J. Chaudhery, "Intelligent Phishing Website Detection Using Random Forest Classifier," In *2017 International Conference On Electrical And Computing Technologies And Applications (ICECTA)*, 2017: IEEE, Pp. 1-5.

[32] W. Gad And S. Rady, "Email Filtering Based On Supervised Learning And Mutual Information Feature Selection," In *2015 Tenth International Conference On Computer Engineering & Systems (ICCES)*, 2015: IEEE, Pp. 147-152.

[33] S. R. Safavian And D. Landgrebe, "A Survey Of Decision Tree Classifier Methodology," *IEEE Transactions On Systems, Man, And Cybernetics,* Vol. 21, No. 3, Pp. 660-674, 1991.

[34] A. Sharaff And H. Gupta, "Extra-Tree Classifier With Metaheuristics Approach For Email Classification," In *Advances In Computer Communication And Computational Sciences*: Springer, 2019, Pp. 189-197.

[35] P. Geurts, D. Ernst, And L. Wehenkel, "Extremely Randomized Trees," *Machine Learning,* Vol. 63, No. 1, Pp. 3-42, 2006.

[36] A. Bansal And S. Kaur, "Extreme Gradient Boosting Based Tuning For Classification In Intrusion Detection Systems," In *International Conference On Advances In Computing And Data Sciences*, 2018: Springer, Pp. 372-380.

[37] T. Nagunwa, S. Naqvi, S. Fouad, And H.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Shah, "A Framework Of New Hybrid Features For Intelligent Detection Of Zero Hour Phishing Websites," In *International Joint Conference: 12th International Conference On Computational Intelligence In Security For Information Systems (CISIS 2019) And 10th International Conference On European Transnational Education (ICEUTE 2019)*, 2019: Springer, Pp. 36-46.

[38] M. Sarkar And T.-Y. Leong, "Application Of K-Nearest Neighbors Algorithm On Breast Cancer Diagnosis Problem," In *Proceedings Of The AMIA Symposium*, 2000: American Medical Informatics Association, P. 759.

[39] T. Hastie, S. Rosset, J. Zhu, And H. Zou, "Multi-Class Adaboost," *Statistics And Its Interface,* Vol. 2, No. 3, Pp. 349-360, 2009.

[40] J. Platt, "Probabilistic Outputs For Support Vector Machines And Comparisons To Regularized Likelihood Methods," *Advances In Large Margin Classifiers,* Vol. 10, No. 3, Pp. 61-74, 1999.

[41] V. V. De Melo And W. Banzhaf, "Improving Logistic Regression Classification Of Credit Approval With Features Constructed By Kaizen Programming," In *Proceedings Of The 2016 On Genetic And Evolutionary Computation Conference Companion*, 2016: ACM, Pp. 61-62.

[42] A. N. Tikhonov, "On The Stability Of Inverse Problems," In *Dokl. Akad. Nauk SSSR*, 1943, Vol. 39, Pp. 195-198.

[43] I. STEPHEN, "Perceptron-Based Learning Algorithms," *IEEE Transactions On Neural Networks,* Vol. 50, No. 2, P. 179, 1990.

[44] A. C. Müller And S. Guido, *Introduction To Machine Learning With Python: A Guide For Data Scientists*. " O'Reilly Media, Inc.", 2016.

[45] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, And Y. Singer, "Online Passive-Aggressive Algorithms," *Journal Of Machine Learning Research,* Vol. 7, No. Mar, Pp. 551-585, 2006.

[46] A. Dongari And M. S. Reddy, "Suspicious URL Detection System Using SGD Algorithm For Twitter Stream," *Int. J. Comput. Sci. Inf. Eng., Technol.,* Vol. 2, No. 4, Pp. 1-6.

[47] K. Keshari. "Naive Bayes Tutorial: Naive Bayes Classifier In Python " Dzone. Https://Dzone.Com/Articles/Naive-Bayes-Tutorial-Naive-Bayes-Classifier-In-Pyt (Accessed 22-10-2019, 2019).

[48] X. Ma, J. Sha, D. Wang, Y. Yu, Q. Yang, And X. Niu, "Study On A Prediction Of P2P Network Loan Default Based On The Machine Learning Lightgbm And Xgboost Algorithms According To Different High Dimensional Data Cleaning," *Electronic Commerce Research And Applications,* Vol. 31, Pp. 24-39, 2018.

[49] X. Shi, Y. Cheng, And D. Xue, "Classification Algorithm Of Urban Point Cloud Data Based On Lightgbm," In *IOP Conference Series: Materials Science And Engineering*, 2019, Vol. 631, No. 5: IOP Publishing, P. 052041.