DECISION TREE BASED NETWORK INTRUSION DETECTION SYSTEM

YISA RHODA NNABA[1], MORUFU OLALERE[2], AFOLORUNSO ADERELE A.[3] & THEOPHILUS ENEM[4]

[1,2] Department of Cyber Security Science, Federal University of Technology Minna, Nigeria.
[3]Department of Computer Science, National Open University of Nigeria, Jabi, Abuja, Nigeria.
[4]Department of Computer Science, Air Force Institute of Technology, Kaduna, Nigeria.
E-mail: danielrhodatsado@gmail.com, lerejide@futminna.edu.ng, aafolorunsho@noun.edu.ng, enemtheophilus@gmail.com

Abstract
Intrusion Detection Systems (IDS) have become crucial parts in computer and network gadgets. With the high increase of network traffic, hackers and malicious users are devising better approaches for network intrusion. In other to address this issue, an intrusion detection system (IDS) is developed which will detect attacks in a computer network. Such attacks like Probing, Denial of Service (DoS), Remote to User (R2U) and User to Root (U2R) attacks are attacks which affect huge number of computers day by day. Detection of these attacks and prevention of computers from it is a major research topic for researchers throughout the world. In this paper, we carry out classification of attacks of network traffic for intrusion detection using Decision tree classifier algorithm. J48 is a supervised and trees machine learning classifier, it shows its best performance in terms of accuracy and classifications. Experimental analysis was conducted on NSL-KDD dataset to judge the implementation of model. The Experimental result shows that our suggested model recorded an improved accuracy of 99.60%.

Keywords: Intrusion detection system, Network devices, Machine learning, Decision tree

Introduction
Information communication network contributes to the improvement of the nature of day-by-day life of individuals, and now considered as fundamental social and financial infrastructure. However, the increase of incidents and threats against this infrastructure has turned out to be a difficult issue. Presently, it is essential to maintain a high-level security to ensure safe and trusted information communication between different organizations. But secured data communication over internet and some other systems is consistently under danger of intrusions and misuses (Abebe & Lalitha, 2013).

Due to the high rate of computer usages at different locations, intrusion detection has become important in the area of network security (Wathiq, Zulaiha & Mohd, 2016). Strategies available for intrusion prevention for instance access control; encryption and firewalls have not given the security level required to protect systems and networks from increasing security attacks (Asma, Alaa & Talaat, 2015). However, these techniques are not enough as each one of the techniques possess significant restrictions. Therefore, it becomes imperative to use other extra defense mechanisms like intrusion detection system (IDS) (Enache & Patriciu, 2014). IDS is a software application or a hardware device that is configured to monitors computer system or network for unusual activities and report or prompt for appropriate action (Amit, Harish & Rahul, 2013). So, Intrusion Detection Systems (IDS) have become essential parts in computer and network security. Intrusion detection systems (IDSs) are arranged into two classes based on the detection method, i.e., signature IDS and anomaly-based IDS. A signature-based IDS matches network traffic patterns with the attack patterns (signatures) already stored in its database Abhizhek & Virender, 2019). Anomaly-based IDS is created based on normal behavior features.

It uses these identified features with normal traffics to pinpoint any action that significantly deviates from the normal features. It uses data taken from normal usage to identify patterns (Shadi, Monthar & Muneer, 2017).

The idea of machine learning has been utilized to create many IDS. Machine learning techniques have yielded better detection performance (Wei-Chao, Shih-Wen & Chih-Fong, 2017). Many machine learning techniques have been introduced to increase the performance of IDS in recent times. One of the popular machines learning techniques in IDS is J48 (Decision tree). The algorithm utilizes a greedy technique to initiate decision trees for classification and uses decreased-error pruning. During the process of building a tree in J48 (Decision tree), there is a particular attribute which is the interior node of the tree and the branches between the inward node holds the details of the values that these attributes can assume. This algorithm has the capability to display or classify both discrete and continuous attributes; and can disregard missing attribute values in a dataset (Chibuzor & Bennett, 2013).
The objectives of this study are as follows:
(i) To identify various intrusion attack types.
(ii) To classify network traffic for intrusion detection using Decision Tree Based Model
(iii) To compare proposed model with different models

In this article we suggest IDS using decision tree classifier to improve accuracy of classifier in classification of different attack types. Section 2 describes about Literature Review and Related Work is narrated in Section. Methodology is explained in Section 3 and in Section 4 we analysis Experimental Results. Finally, in section 5 we conclude.

Literature Review

Reviewing of Concepts

Remote to local (R2L) attack

The attacker wants to send packets to target machine remotely to expose vulnerabilities and obtain access of local target machine (Nikhitha & Jabbar, 2019). The attacker sends packets that are capable of compromising the target system over the network, the machine loopholes or vulnerability are then exploited to gain unauthorized access. Attackers with ability to communicate with their target device but have no account on that device use this kind of attack to exploit weaknesses that exist on the target system to acquire local access on the target device (Jamal, 2014). Attacks like this can be carried out by making use of ports that are open on the target system, using the system loopholes, password guessing [16]. Examples of such attack are multihop, send-mail and Imap (Nikhitha & Jabbar, 2019).

User to Root (U2R) attack

When the attacker starts as a normal account user then slowly exploits vulnerabilities to obtain illegal root access of the computer (Nikhitha & Jabbar, 2019). Unauthorized access to local super user (root) privileges (Kumar, Mishra & Sahu, 2016). The difference is that the attacker here is already a normal user and he/she wants to escalate his/her privilege (Shadi *et al.,* 2017). User to root (U2R) attack simply refer to a situation where a legit or normal user wants to gain higher privilege in other to carryout illegal or unauthorized activities (Jamal, 2014), e.g., various ``buffer overflow'' attacks; Perl, Eject and load module.

Probing attack

This class of attack has to do with reconnaissance, gathering information by scanning systems and networks to find weaknesses that exist with them. The found loopholes are used to exploit the systems and networks (Jamal, 2014). Surveillance and other probing, in these types of attacks, before initiating the attack the attacker will gather all the required information of the target system (Nikhitha & Jabbar, 2019). Examples are Satan, Ipsweep and Nmap, port and scanning.

DoS or DDoS attacks

In this attack, the attacker avoids the authorized user from accessing the network or making the services unavailable to them (Nikhitha & Jabbar, 2019). DDoS attack is one of the attacks that cause menace to the stability of the Internet, affecting services like online applications and procedures, system and network performance, emails and other system resources (Ismaila, Obi, Shafi'i, Morufu, & Baba, 2017). A DDoS attack is a coordinated attack on the availability of services of a given target system or network that is launched indirectly through many compromised computing systems (Olalere, Mohd, Ramlan & Azizol, 2015). Denial of Service (Dos) attack often involves attacker sending traffics that are more than what the victim system can handle making such system deny legitimate users' access to services (Sheetal, Priti & Arundhati, 2017). DoS attack usually originated from a single source. A DoS attack becomes a DDoS attack if the traffics originated from sources more than one (Vani & Munivara, 2016). DDoS attacks are usually carried out by deploying many compromised systems (usually called botnet or zombies) to overwhelm their victim (Yadav, Trivedi & Mehtre, 2016). Dos and DDoS attack are attacks targeted at compromising the availability of computer system, router, network and their resources (Hoque, Kashyap & Bhattacharyya, 2017). The techniques include ICMP flood attack, TCP flood attack, UDP flood attack, http flood attack, SYN flood attack (Yadav *et al.,* 2016).

J48 Tree

As indicated by (Mouhammad & Mohammad, 2018), J48 tree was first presented by Breiman. It is the most well-known classifier used to deal with the database for administered discovering that gives a forecast about new unlabeled information, J48 makes Univariate Decision Trees. J48 based used attribute correlation based on correlation-based feature selection for each attribute (Bhargava, Sharma, Bhargava & Mathunia, 2013). It has been utilized in numerous fields of study, for example, information mining, machine learning, data extraction, design acknowledgment, and text mining. It has numerous points of interest; it is fit for managing diverse info information types: numeric, literary and ostensible. J48 decision tree is an augmentation of the calculation ID3. It has a preferred position over ID3 in that it can construct little trees. It follows a profundity first system, and a separation and overcome approach.

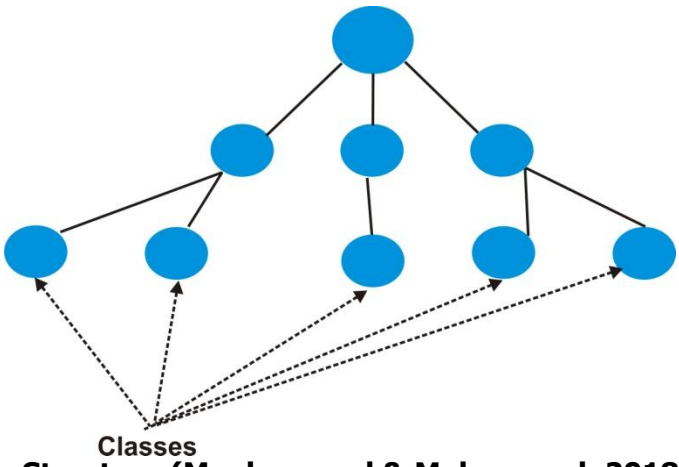

Fig. 1. Decision Tree Structure (Mouhammad & Mohammad, 2018)

A decision tree comprises of a few components: root, internal nodes, and leaves. The internal nodes speak to the conditions where the estimation of the boundaries will be tried. In view of these qualities and the condition, the flow of the tree will be decided (along which branch the decision tree must go). Leaf nodes represent the decision or the class. Fig. 1 shows a typical decision tree structure.

The tree is constructed by following these three principal steps:

(i) Ensure that all of the grouped inputs are of the same class. Then ensure that the tree is labeled with the class.

(ii) Calculate some parameters for each attribute, such as correlation-based feature selection.

(iii) Choose the best split attribute based on the criteria that have been set.

Related Work

In Adiya, Ramanathan and Ramani (2018), proposed an IDS based on lazy learning algorithm. The paper aims to build the overall performance of intrusion detection system. So as to subdue the inherent limitations of high search complexity inherent in lazy learning. K-Nearest Neighbour and Locally Weighted Learning, a respectable heuristic weight-based ordering procedure has been utilized, two well-known lazy learning algorithms were also compared and used on the NSL-KDD dataset for simulating a real world like scenario and comparing their relative performance with K-Nearest Neighbour. The result of the experiment shows that a lazy algorithm is more promising for real-world network intrusion detection.

Alex *et al.,* (2018) proposed an intelligent intrusion detection system using artificial neural network. The proposed approach was experimented using benign network traffic like dynamic link library files, images, word processing documents, music files. An average accuracy of 98% was obtained.

Bahram and Nima (2019) proposed detection system that combines multilayer perception network, artificial bee colony and fuzzy clustering algorithms. The Multilayer Perception was utilized to recognize ordinary and strange traffic while Artificial Bee Colony was utilized to prepare MLP through upgrading the linkages weight and predispositions. The approach was evaluated using NSL KDD on CloudSim simulator. The result obtained shows that the algorithm had 98.6% accuracy rate.

In Chibuzor and Benrett (2018), the authors of the paper proposed an intrusion detection system using machine learning algorithm**.** *Bayes Net, J48, Random Forest, and Random Tree* algorithms were used with the aim of getting high detection value and low false positive value. The result of their approach was evaluated employing KDDCup99datasets. The paper got detection rate of 90.6%, 86.1%, 87.5% and 88.8% for algorithms used.

In the network intrusion detection algorithm developed by Jasmin, Samed and Abudulhamit, (2016), using random tree and Naïve Bayes tree classifiers. The paper aim is to have a hybrid classifier that can classify traffic entering a network into normal or attack with better accuracy compare to the individual classifiers. The study used the NSL-KDD dataset to assess how well their classifiers perform. Detection accuracy of 89.24% was achieved.

Anish and Sundarakantham (2019), deployed machine learning based intrusion detection system. Machine learning techniques like Support Vector Machine (SVM) and Naïve Bayes was applied. The effectiveness of their approach was measured using NSL– KDD dataset. The result indicated that SVM attains accuracy of 93.85% percentages and Naïve Bayes attains accuracy rate of 71.001%**.** The feature work proposed is to a hybrid multi-level model will be constructed to model will be built to improve the precision.

Abhishek and Virender (2019), implemented machine learning based intrusion detection systems for IoT applications using Random Forest (RF), AdaBoost (AB), Gradient Boosted Machine (GBM), Extreme Gradient Boosting (XGB), Extremely Randomized Trees (ETC), Classification and Regression Trees (CART) and Multi-layer Perceptron (MLP). CIDDS-001, UNSWNB15, and NSL-KDD datasets were used. It is observed that RF performs better than other classifiers in terms of accuracy (94.94%) and specificity (91.6%). GBM performs best in terms of sensitivity (99.53%). In terms of AUC metric, XGB performs best by achieving 98.76%. MLP is the worst performer in terms of accuracy (82.76%), whereas AB performs worst in terms of specificity (86.72%) and sensitivity (97.94%).

Rama, Xi-Zhao, Joshua, Haider and Yu-Lin (2017), uses fuzziness based on semi supervised approach for intrusion detection system. To improve the classifier performance for the IDS, samples that are not labeled supported with supervised learning algorithm were used. Results gotten from experiment using this method reveal that samples that are not labeled belonging to the categories of low and high fuzziness groups provide the most input to increase the performance of the classifier compared to classifiers that are already existing examples random forest, naive bayes, support vector machine. They got an accuracy of 84.12%.

Abebe and Lalitha (2013), proposed intrusion detection using Random Forests Classifier with SMOTE and feature reduction. The effectiveness of their approach was measured using NSL–KDD dataset**.** The results show that Random Forests classifier with SMOTE and information gain-based feature selection gives better performance in designing. IDS that is efficient and effective for network intrusion detection. Their experiment yielded accuracy detection rate of 97.2%.

Mouhammad and Mohammad (2018), proposed a combined method of machine learning methods for network intrusion detection. The J48, MLP, and Bayes Network classifiers have been chosen for this study. Among the classification techniques (J48, MLP and Bayes Network), the J48 classifier has achieved the highest accuracy rate of 93.1083 %for detecting and

classifying all KDD dataset attack types (DOS, R2L, U2R, and PROBE). KDD dataset has 41 attributes. Our proposed model show accuracy of 99.60% with 10 folds cross-validation while 99.04% with 70% split using J48 decision tree classifier.

Nabil and Jabba (2016), proposed IDS using Random Forest classifier. Feature selection technique is used to reduce data dimensionality. The experimental result was done on dataset NSL-KDD and the model is efficient for higher classification accuracy and DR. Their proposed random forest model recorded an accuracy of 99.57%.

In Shrivastava 2017), the author of the paper proposed a Supervised Intrusions Detection System. The approach uses KNN algorithm with the aim of obtaining high detection value and low false positive value. The result of the approach was evaluated employing KDDcup 1999 dataset. The paper got an accuracy of 99.33%.

Methodology

Experimental Procedure

In the experiment, our proposed framework was implemented on 2.4GHz Intel core i5-2430M processor with 4GB RAM. We used the WEKA machine learning tool version 3.8.3for our Intrusion Detection model development.

The attack types are grouped into four categories: DoS, Probe, U2R and R2L.

Table (I) shows the attack types in Nsl-Kdd training dataset and their categorization used in the experiments.

Table I: Attack Types in Nsl-Kdd Training Dataset and their Categorization

Attack class	Attack Name
DoS/DDoS	Neptune, Smurf, Pod, Teardrop, Land, Back
Probing	Port-sweep, IP-sweep, Nmap, Satan
R2L	Guess-password, Ftp-write, Imap, Phf, Multihop, Warezmaster, Warezclient
U2R	Buffer-overflow, Load-module, Perl, Rootkit, spy

Table (II) shows the distribution of records in different classes for testing dataset used in the experiments.

Table 1I: Distribution for Test Dataset

Attack types	Number of samples
DoS/DDoS	15,479
Probe	1000
R2L	800
U2R	53
Normal	9,000
Total	

Proposed Model

For the proposed model, we used NSL-KDD full training set and 10-fold cross validation and 70% split test for the testing purposes. In 10-fold cross validation, the available data are randomly divided into10 disjoint subsets of about equivalent size. One of the subsets is then used as the test set and the remaining 9 sets are used for building the classifier. The test set is then used to estimate the accuracy. This is done repeatedly 10 times so that each subset is used as a test subset once. The accuracy prediction is then the mean of the predictions of each of the 10 models. Cross-validation generally works well when sufficient data is available.

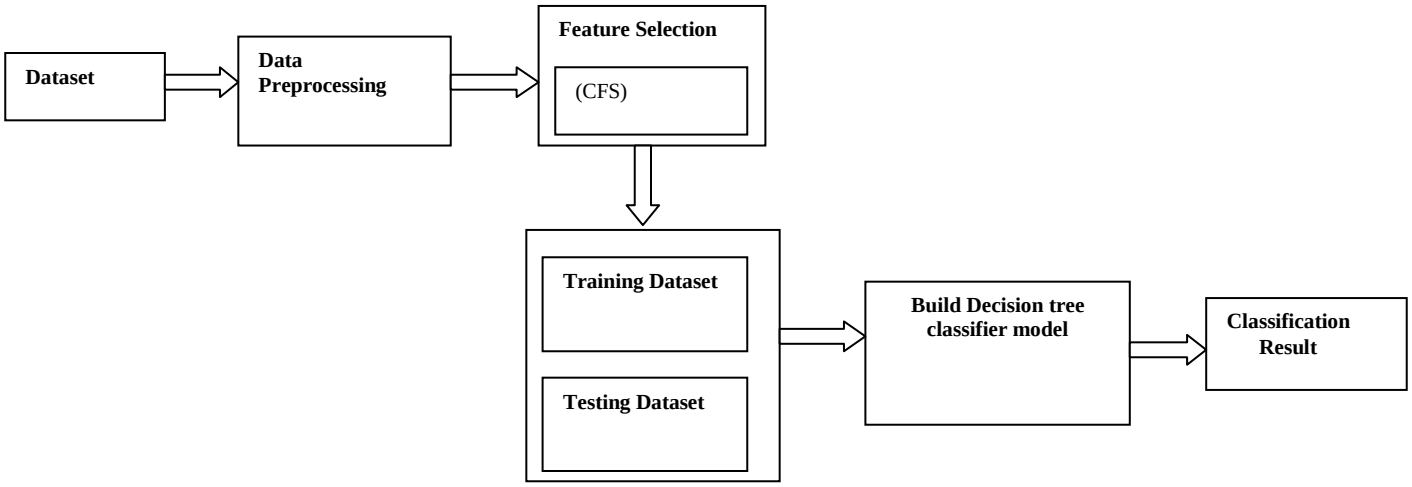

Fig. 2: Proposed IDS Model

Pre-processing

Feature selection step is correlated to data pre-processing step where irrelevant features are removed which increases the accuracy. The collected dataset was preprocessed by removing the instances that are non- numeric in nature, also the data was manually cleaned up by removing blank spaces. The overall process of pre-processing is essential, in which non-numeric or symbolic features are eliminated or replaced, as they do not perform any important participation in intrusion detection. Symbolic attributes like protocol, service and flag get changed or removed. Finally, the instances get labeled under four categories: Normal, DoS, Probe, and R2L.

Feature Selection

Feature Selection refers to identification of features that are strongly correlated to problem which are useful in prediction of class. The experiment was conducted using the correlation based feature selection (CFS) with best first search method in other to remove the irrelevant features from the datasets. The original datasets has 42 attributes including the class, by performing a feature selection on the attributes using the CFS, the attributes on the dataset is now reduced to 14 as seen in table (III).

Table III: Feature Selection of Attributes

Correlation based feature selection (using Best First)
Service
src_bytes
dst_bytes
Hot
num_failed_logins
logged_in
num_file_creations
Count
srv_count
srv_diff_host_rate
dst_host_diff_srv_rate
dst_host_rerror_rate
dst_host_srv_diff_host_rate
Attack

Experimental Results

Performance Evaluation

An intrusion detection system (IDS) is evaluated by the measure of accuracy, precision, recall and F-measure. An intrusion detection system should have a very low false alarm.

Table V: Results Obtained from the Proposed Model Classifier

	Accuracy	FP Rate	Recall	Precision	F-Measure
10 fold	99.60	0.005	0.996	0.996	0.996
70% split test	99.04	0.013	0.990	0.990	0.990

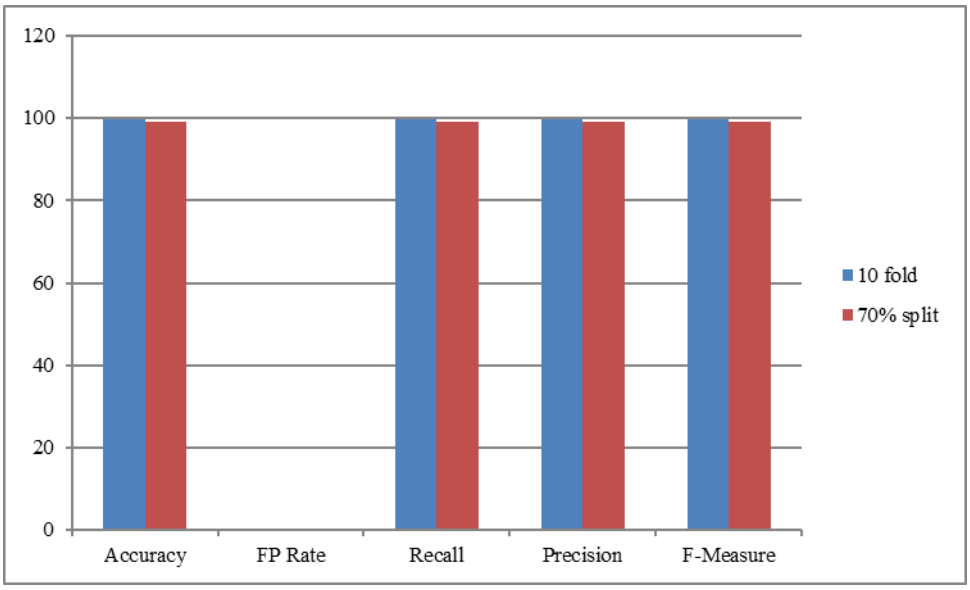

Fig 3: Accuracy, False Positive Rate, Recall, Precision and F-Measure for full training set

Fig: 3, indicate the 99.60%, 0.005, 99.60%, 99.60% and 99.60% Accuracy, FP Rate, Recall, Precision and F-Measure obtained when 10 folds cross-validation was employed on our proposed algorithm and little variation is observed compared to 70% split test 99.04, 0.013, 99.00%, 99.00% and 99.00% Accuracy, FP Rate, Recall, Precision and F-Measure respectively.

Table VI: Differentiating Proposed Model with Others Classifiers

S/N	Approach	Accuracy
1.	SMO	93.92
2.	IBK	99.10
3.	Proposed model	99.60

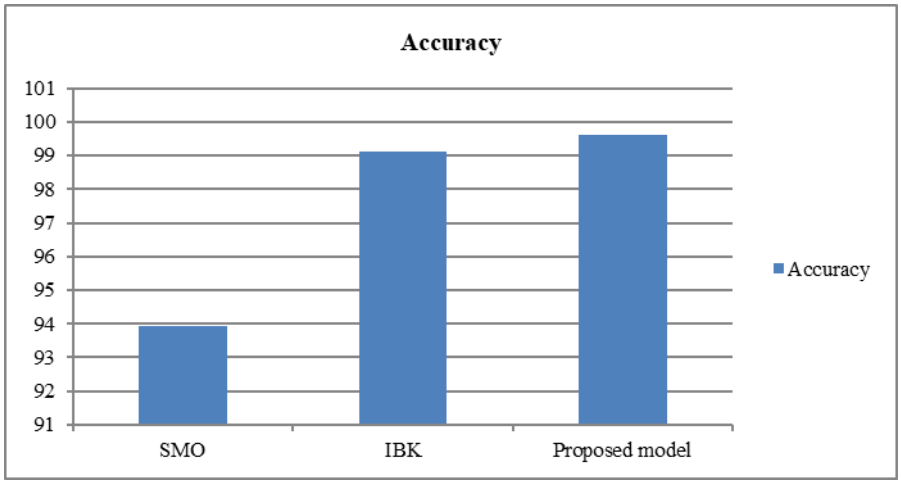

Fig 4: Comparison of our approach with different models.

Figure 4, shows that SMO has the lowest accuracy of 93.92%, IBK has accuracy of 99.10% and the proposed model has the highest accuracy of 99.60%.

Conclusion

In this paper, we applied the Decision tree (J48) algorithm to NSL-KDD data set to classify the type of attacks like Dos, probe, U2R and R2L. Proposed method is validated by 10 cross validation and 70% split test for the classification. The Experimental analysis shows that when compared to other classification methods, proposed model has increased 99.60%, 0.005, 99.60%, 99.60% and 99.60% Accuracy, FP Rate, Recall, Precision and F-Measure values. In future, we intend to apply hybrid techniques for the classification of NSL-KDD dataset to increase the precision of classifiers.

References

Abebe, T. D., & Lalitha, B. (2013). Intrusion detection using random forests classifier with smote and feature reduction. *International* Conference *on Cloud & Ubiquitous Computing & Emerging Technologies.*

Abhishek, V., & Virender, R. (2019). Machine learning based intrusion detection systems for IoT applications. *Wireless Personal Communications*. doi.org/10.1007/s11277-019-06986-8

Aditya, C., Ramanathan, L., & Ramani, S. (2018). Intrusion detection in computer networks using lazy learning algorithm. *International Conference on Computational Intelligence and Data Science.* Procedia Computer Science, (132), 928 – 936.

Alex, S., David, D., & Aladdin, A. (2018). Intelligent intrusion detection systems using artificial neural networks. *ICT Express*, https://doi.org/10.1016/j.icte.2018.04.003

Amit, K., Harish, C. M., & Rahul, M. (2013). A research paper on hybrid intrusion detection system. *International Journal of Engineering and Advanced Technology,* 2(4), 294 – 297.

Anish, H. A., & Sundarakantham, K. (2019). Machine learning based intrusion detection system. *Proceedings of the Third International Conference on Trends in Electronics and Informatics,* ISBN: 978-1-5386-9439-8.

Asma, A. H., Alaa, F. S., & Talaat, M. W. (2015). Intrusion detection system using weka data mining tool. *International Journal of Science and Research*, 6(9), 337–342.

Bahram, H., & Nima, J. N. (2019). Intrusion detection for cloud computing using neural networks and artificial bee colony optimization algorithm. *ICT Express,* (5), 56–59. https://doi.org/10.1016/j.icte. 2018.01.014

Basant, A., & Namita, M. (2012). Hybrid approach for detection of anomaly network traffic using data mining techniques. *2nd International Conference on Communication, Computing & Security*, 6(2), 996–1003.

Bhargava, N., Sharma, G., Bhargava, R., & Mathuria, M. (2013). Decision tree analysis on j48 algorithm for data mining. *Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering,* 3(6).

Chibuzor, J. U., & Bennett, E. O. (2018). An intrusion detection system using machine learning algorithm. *International Journal of Computer Science and Mathematical Theory,* 4(1), 2545-5699.

Enache, A. C., & Patricia, V. V. (2014). Intrusions detection based on support vector machine optimized with swarm intelligence. *9th IEEE International Symposium on Applied Computational Intelligence and Informatics,* 153–158.

Fangjun, K., Weihong, X., & Siyang, Z. (2014). A novel hybrid KPCA and SVM with GA model for intrusion detection. *Applied Soft Computing,* 18, 178–184.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. ACM SIGKDD explorations newsletter 11(1), 10-18.

Hoque, N., Kashyap, H., & Bhattacharyya (2017). Real-time DDoS attack detection using FPGA. *Elsevier journal,* 48–58.

Ismaila, I., Obi, B. F., Shafi'I, M. A., Morufu, O., & Baba, M. (2017). Distributed denial of service detection using multi layered feed forward artificial neural network. *Computer Network and Information Security*, 12, 29-35 DOI: 10.5815/ijcnis.2017.12.04.

Jamal, R. (2014). A survey of cyber-attack detection strategies. *International Journal of Security and Its Applications,* 8(1), 247-256.

Jasmin, K., Samed, J., & Abdulhamit, S. (2016). An effective combining classifier approach using tree algorithms for network intrusion detection. *Neural Computing & Applications*.

Kumar, M., Mishra, S. K., & Sahu, S. S. (2016). Cat swarm optimization based functional link artificial neural network filter for gaussian noise removal from computed tomography images. *Applied Computational Intelligence and Soft Computing, 1 – 5.*

Olalere, M., Mohd, T. A., Ramlan, M., & Azizol, A. (2015). A review of bring your own device on security issues. doi.org/10.1177/2158244015580372.

Mouhammad, A., & Mohammad, A. (2018). Machine learning methods for network intrusion detection. *Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering,* (3),6.

Nabil, F., & Jabbar, M. A. (2016). Random forest modeling for network intrusion detection system, ELSEVIER, Science Direct.

Nikhitha, M., & Jabbar, M. A. (2019). K- nearest neighbor-based model for intrusion detection system. *International Journal of Recent Technology and Engineering*, 8(2).

Rana, A. R. A., Xi-Zhao, W., Joshua, Z. H., Haider, A., & Yu-Lin, H. (2017). Fuzziness based semi-supervised learning approach for intrusion detection system. *Information Sciences*, (378), 484 – 497.

Shadi, A., Monther, A., & Muneer, B. Y. (2017). Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model, *Journal of Computational Science.*

Sheetal, P. D., Priti, R. H., & Arundhati, A. D. (2017). Denial of service attack defense techniques. *International Research Journal of Engineering and Technology*, 4(10), 1532 – 1535.

Shrivastava, U. (2017). Supervised intrusions detection system using KNN. *International Journal on Recent and Innovation Trends in Computing and Communication,* 5(5), 559 – 562.

Vani, N., & Munivara, P. K. (2016). Detection of anomaly-based application layer DDoS attacks using machine learning approaches. *i- manager's Journal on Computer Science,* 4(2).

Wathiq, L. A., Zulaiha, A. O., & Mohd, Z. A. N. (2016). Multi-level hybrid support vector Machine and extreme learning machine based on modiþed k-means for intrusion detection system, *Expert Systems with Applications.*

Wei-Chao, L., Shih-Wen, K., & Chih-Fong, T. (2015). An intrusion detection system based on combining cluster centers and nearest neighbors. *Knowledge based systems.*

Yadav, V. K, Trivedi, M. C., & Mehtre, B. M. (2016). An approach to handle DDoS (ping flood attack. *Proceedings of International Conference on ICT for Sustainable Development, Advances in Intelligent Systems and Computing,* 4(8), 11 – 23.

Yisa, R. N., & Olalere, M. (2019). Detection distributed denial of service (DDoS) using linear support vector machine. *2nd International Conference on Applied ICT,* 2(3).