2021 5th International Conference on Computing Methodologies and Communication (ICCMC) | 978-1-6654-0360-3/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCMC51019.2021.9418291

An Effective Intrusion Detection System using Supervised Machine Learning Techniques.

Aamir S Ahanger[1], Sajad M Khan[2], Faheem Masoodi[3]*

[1]Research Scholar, Department of Computer Science, University of Kashmir
[2]Scientist D, Department of Computer Science, University of Kashmir
[3]Assistant Professor, Department of Computer Science, University of Kashmir
*Corresponding author email: masoodifahim@uok.edu.in

Abstract – **With the increased use of Internet resources, cyber attackers are using novel ways to attack the services of network. Thus network security is becoming inevitable part of the network system. In order to detect such attacks efficiently and effectively, robust IDS (Intrusion Detection System) is needed. An IDS is a tool that analyzes each and every packet deeply to detects malicious activity by monitoring a network or a system. The main purpose of IDS is to identify unwanted or abnormal action and to inform the network administrator about such actions. Thus, IDS is important tool for the network administrator to prevent the network from both known and unknown attacks that make the network resources more vulnerable. Machine learning methods can be used to employ efficient intrusion detection system (IDS). In this research work four machine learning methods were used namely RF (Random Forest), DT (Decision Tree), MLP (Multilayer perceptron) and SVM (Support Vector Machine) for classification of the data. NSL-KDD dataset was used for training and testing these various machine learning models. Feature selections were used to eliminate the irrelevant and unwanted features from the dataset. Therefore feature selection reduces the dimensionality of the dataset which in turn reduces the computational complexity. The proposed model's output was evaluated using three feature subsets, randomly selected from the NSL-KDD dataset. The proposed method has a classification accuracy of more than 99 percent.**

Keywords – IDS ; RF ; DT ; SVM ; MLP ; Feature Selection; Machine learning.

1. Introduction

Cyber-attack is a malicious action intended to target the network and its resources, in order to destroy, disable, change or to gain unauthorized access to the network resources or data they have [1]. The rise in Cyber-attacks has increased the threats to the network resources which lead to new challenges for the cyber-security. With the advent of newer technologies like Internet of things, Cloud computing and big data, the organizations are more vulnerable to these attacks. Therefore, it is extremely urgent for the organizations to take necessary steps to protect their asserts from the damage[2]. It very important for the network administer to take necessary security mechanisms to protect the sensitive data and resources from unauthorized attempt. The primary goal of network security is to protect the network from the unwanted code that changes the data, logic or computer code that could harm the network resources and to preserve integrity, enhance the availability and protect the confidentiality of the network. There are certain types of attacks which the conventional security mechanisms cannot detect because the intruders use several different techniques, approaches to skip and penetrate the security system of the network. Firewalls and Encryption techniques cannot provide complete security solution to all types of network attacks (e.g. DoS). Automatically firewalls are unable to defend the network against the new or unknown attacks that may arise. Firewalls cannot guard the network against malicious attacks performed by the insiders and these attacks are considered to be the most damaging [3]. Internet is continuously changing attackers are discovering new vulnerabilities, ways to attack the networks. Thus novel security mechanisms are needed to deal with all types attacks in an efficient manner to preserve the security of the computer networks. So there is a need to deploy another new enhanced device which can detect all type attacks intrusions with maximum accuracy [4]. Thus IDS play important role in detecting such attacks. If such security mechanisms are not implemented within the network the attackers will misuse or destroy the whole network.

Intrusion detection is a process of finding suspicious patterns from the network data that may damage the network infrastructure[5]. When performing intrusion detection it is built on the fact that the malicious traffic looks different from the normal traffic. An IDS is considered as second line of security mechanism, especially designed to check all network traffic and computer system, senses the incoming and outgoing traffic continuously to find the hidden anomalies within the data and immediately produces an alert if something unusual found in the traffic before intruders can damage the network infrastructure[6]. Therefore the main function of the intrusion detection is to inspect the network traffic (both incoming and outgoing) and taking appropriate actions while malicious traffic is identified/encountered i.e. raising an alert or other actions are also possible such as dropping the packet. The main component of IDS is the detection engine and its primary function is to find the malicious traffic. When some of the samples of malicious traffic have been identified, the IDS locked these sample in other component called log for the later use and decides a possible action against the selected attack to guard the network. An IDS can be described based on the several characteristics which will define the taxonomy. Based on the detection methods used by the detection engine IDS can be divided into two categories [7]: Anomaly based and Misuse based. These categories define the internal functioning of the IDS. In MIDS (or signature or knowledge based) the detection is based on the reality of the model ie our model describes how attack and normal traffic looks like. AIDS or behavior based the internal model defines the structure of the normal and attack traffic. Further based on the deployment or position of the IDS in the network architecture these are classified as Network-based (NIDS) protects the whole network for which it is implemented or Host-based (HIDS) protects a single host. Before deploying the IDS in real world the performance of the IDS must be evaluated. Thus for evaluation purposes, researchers require decent quality of the dataset to train and test the model. To evaluate the performance of machine learning classifiers on the given dataset the metrics used for the evaluation purpose are accuracy, recall, precision etc. Some of the known publically available datasets are DARPA, KDD-CUP'99, NSL-KDD and ADFA-LD. Machine learning algorithms has been used successfully in many areas like Image processing, natural language processing and computer vision etc. Machine learning algorithms depends heavily on large data to find the hidden patterns by using set of procedures, complex transformation functions[8][9]. Machine learning algorithms uses two learning approach to understand the data clearly: Supervised learning and Unsupervised learning methods. In supervised learning the data used for learning the model contains labeled data (data with output) but in unsupervised learning training data does not contain labels the model itself lurks inside the data to find the natural patterns. During the training phase, the training data is used to set the parameters of the complex functions, so that the model can classify the data efficiently. Intruders are changing their behavior by using latest techniques and tools. Intruders use such techniques to change their network behavior patterns to bypass traditional intrusion detection system. Thus becoming necessary for research community to switch to new latest and dynamic approaches to detect and prevent these intrusions. Therefore Implementing an effective IDS which can detect such novel attacks is challenging task. The rapid enhancements in machine learning techniques has increased the predictions and computing power of machines. Thus these techniques can be used to build robust Intrusion Detection Systems. Recently researchers are using Intrusion Detection Systems and feature selection based on Machine Learning tools that are showing promising results in detecting intrusions like Random Forest[10], Decision Tree[11] , Multilayer Perceptron[12] and Support vector Machine[13].

2. *Literature Survey :*

Soodeh et al[14] introduced a new machine learning algorithm consists of Genetic Algorithm, Logistic Regression and ANN for the intrusion detection system. In the first stage Logistic Regression and Genetic algorithm were used to extract the correlated feature subset from the dataset. Then in the second stage Artificial Neural Network is trained using PSO and GS algorithm to detect the intrusions and. Two datasets were used to assess the performance of the proposed model namely NSL-KDD and KDD cup'99. The proposed model gets lower accuracy rate but the model detects attacks faster than the other ANN based methods.

Faezah et al [15] reduced the features of the data by using wrapper method based on Differential evolution technique for IDS. The number of features has been reduced because irrelevant features influence the accuracy of IDS. The idea is to select some of the feature from NSL-KDD dataset using differential evolution and using ETM to determine the performance of the given model. The proposed model attained classification rate of 80.15% for five classes and 87.3% for two classes.

Iram et al [16] a empirical research on machine learning classifiers based on SVM, KNN, LR, NB, MLP, RF, DT and ETC for the classification of network data as abnormal and normal was employed and the performance of the research was evaluated on four different subsets derived from NSL-KDD dataset. Before the training the model the training data was preprocessed based on significant features. The results reveal that the machine learning classifiers produce better results for Denial of Service attack and low results were achieved for U2R attacks and in general accuracy of the model is 99%.

Miranal et al[17] designed an IDS based on deep learning technique using NSL-KDD dataset to detect the intrusion within the network. The model learns as well as have adaptive ability to find the new patterns that are not interpreted earlier. The proposed model uses NSL-KDD dataset for training and incorporates auto-encoder along with Logistic Regression. The precision score achieved by the model is more than 84%.

Yuyang et al [18] proposed an efficient IDS based on heuristic feature selection called CFS-BA to reduce the dimensionality of the data based on correlation between the attributes. Then for the detection purpose an ensemble approach consists of Random forest, Forest by penalizing Attribute and C4.5 algorithms were employed. Finally by voting process the probability distribution of the base learners were combined to recognize the attacks. The results showed an accuracy of 99.8% with the subset having 10 features selected from NSL-KDD dataset.

3. *Methodology*

This section discusses the results of the proposed work, in which four classifiers, RF, DT, SVM, and MLP, were used to mark packets as normal or malicious based on the data they contained. The model's output was evaluated using three separate feature subsets taken from the NSL-KDD dataset.

The steps used in this work will be described and summarized in the following parts. The NSL-KDD dataset is preprocessed in the first phase to optimize and delete unnecessary features from the raw data. In the second stage, three datasets were chosen at random to test the models' accuracy. Machine learning classifiers were used for training and testing in the third phase.

The effects of the four classifiers were evaluated in the final stage.

A. Dataset and Preprocessing

Machine learning algorithms relies on huge amounts of data to train the model before they can provide better results. Data is typically stored in storage device like files, databases etc this data can not directly used to training purpose. We have to preprocess or refine the data to achieve better results before it can be passed to the machine learning model for the training purpose. Training data enables the machine learning classifier to understand how given values are related to the class. So that machine learning model can easily understand the training data and provide better results. Data preprocessing step consists of multiple processes. It starts from loading the data to the machine learning algorithm to handling the missing variable of the dataset , scaling the data with the help of standardization and normalization, splitting the dataset into training and testing dataset. So that we can

pass the training set to the learning classifier for the training purpose and use test set to evaluate the performance of machine learning classifiers. Table 1 summarizes the details of the three randomly selected feature subsets from the NSL-KDD dataset.

Table 1. shows feature subsets selected and number of instances selected randomly from NSL-KDD Dataset for training & testing.

SELECTED FEATURE SUBSET	TOTAL NO. OF ROWS	NO. OF ROWS IN TRAINING SET	NO. OF ROWS IN TEST SET	SELECTED FEATURES	NO.OF FEATURES SELECTED
FEATURE SET 1ST	*1,25,971*	*81,882*	*44,089*	*7-9, 12, 20-24, 26, 28-34, 36-41.*	*23*
FEATURE SET 2^{ND}	*1,25,971*	*81,882*	*44,089*	*23, 24, 26, 28-34, 37-41.*	*15*
FEATURE SET 3^{RD}	*1,25,971*	*88,180*	*37,791*	*23, 24, 29-31, 34, 36-41.*	*12*

B.. Classification

The primary challenges that IDS face are high false alarm rate (false negative and false positive) and lack of real time response. Machine learning algorithms have the power to tackle such challenges. Machine learning techniques can be used to built intelligent IDS which can detect both known and unknown attacks with high speed, maximum accuracy and minimum false alarm rate [19][20]. Thus machine learning algorithms can be used to boost the IDS with enhanced capabilities. Machine learning classifiers, namely RF (Random Forest), DT (Decision Tree), MLP (Multi-layer Perceptron), and SVM (Support Vector Machine) were used as Intrusion monitoring engine to find the intrusions by classifying the data. Random forests are the powerful supervised learning algorithms. Random forests are ensemble classifiers (consists of multiple decision tress) that are used to increase the performance of the system[21]. The output of multiple trees are selected to categorize the data in appropriate class. Random Forest the most used supervised machine learning algorithm for grouping, classification of the data based on similar features they share. Random Forests are built from finite trees. Each tree behaves like a single decision tree where each tree selects features randomly from dataset. Therefore, Using Random Tree for classification purpose, the number of trees should be fixed before implementation. DT is machine learning classification algorithm to classify the data. The decision tree algorithm is predicatively learned to built a model from the categorized set of data to mapping an instance based on set of selected features to a specific class[22]. Every sample is defined by their values of respective features. The main function is to detect the features, which best divides the data into their respective classes. Nodes can be split by using entropy. Entropy measures the purity of split of a sample in node. For the purpose of split entropy was used in this research to choose the best node. MLP (Multilayer Perceptron) is a neural network consists of single or more than one hidden layers. The layers in the MLP should be a minimum of three layers i.e. input, output and hidden layer which maps the input variable to the final output[5]. Rectified linear unit function was used to train and test the model and 10 neurons were used in only hidden layer. SVM is supervised learning algorithm used for binary categorization of the non linear and linear data. SVM (Support Vector Machine) classifies the data by constructing N-dimensional hyperplane by divides a group of positive samples from a group of negative samples with highest margin [4][5]. For training radial basis kernel function was applied to maximize the predictive results for non-linear data. The performance of the proposed model was analyzed on three different feature subsets extracted from the NSL-KDD dataset. Preprocessing is an essential task to remove/replace the irrelevant feature to increase the accuracy and improve the robustness of the detection process. The performance, computational cost of the IDS is based on the selected features/dimensionality of the dataset, therefore, dataset was preprocessed to remove unimportant attributes. In this research work two datasets were used. Those features which contribute the most for the classification were selected randomly.

4. Results

The experiments were carried on NSL-KDD dataset. NSL-KDD dataset has 41 columns, makes it difficult to work as they increase the computational cost. Therefore, the dataset is reduced to meet the requirement for the experiments. Three datasets were selected randomly from the original dataset to reduce the computation cost. All the experiments were carried out in Google Colab and effectiveness of all the classifiers in classifying the NSL-KDD data set is studied. The RF classifier produces the highest accuracy of above 99% using 1st feature subsets. Random forest is ensemble classification technique (multiple classifier are combined to increase the prediction) which uses finite (two or more) number of decision trees to perform the classification. Thus the model reduces the computational cost by removing some irrelevant features and enhances the accuracy of the model especially for RF and DT. Table 2 shows the results of different classifiers in classifying the data using three randomly selected feature subsets and figure 1 shows a graphical representation of the findings in terms of accuracy on various feature subsets.

Table 2. Results generated by four classifiers (RF, DT, SVM, MLP) on three datasets with different feature set

SR.NO.	CLASSIFIER	ACCURACY OF CLASSIFERS ON DIFFERENT SELECTED DATASETS / ATTRIBUTES		
		DATASET 1^{ST} WITH 23 FEATURES	DATASET 2^{ND} WITH 15 FEATURES	DATASET 3^{RD} WITH 12 ATTRIBUTES
1.	RF	99.1%	98.77%	98.43%
2.	DT	97.14%	96.81%	96.66%
3.	SVM	98.79%	98.24%	96.85%
4.	MLP	97.97%	97.43%	97.85%

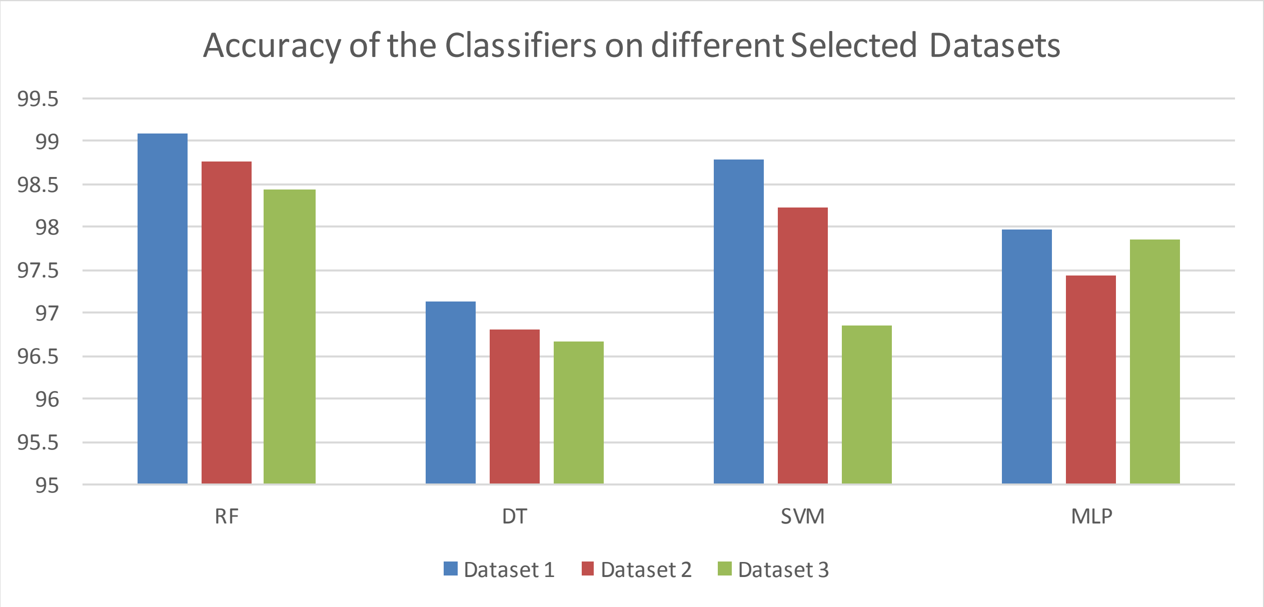

Figure.1: Graphical representation of the results generated.

5. *Conclusion*

In this paper, empirical experiments were performed using four machine learning classifiers namely RF, DT, MLP and SVM to tested and evaluate the efficiency and performance. The training and testing were carried on the three feature subsets extracted from NSL-KDD intrusion detection dataset. Initially NSL-KDD dataset was preprocessed to select relevant features to increase the efficiency and reduce the training time. In the experiments 81,882 instances of rows were used to train the selected machine learning models. For testing purpose 44,089 random instances were used. Based on the results achieved random forest produced the highest classification rate of more than 99%, and decision tree produced the lowest accuracy rate of 96.60% among the four classifiers.

The researchers should focus on false positive and false negative performance metrics that degrade the performance of the intrusion detection model. The empirical study has reveal that there is no single machine learning algorithm which can detect all the types of attacks effectively. In future, relevant features can be extracted from the original dataset to decrease the computation time and increase the accuracy rate of machine learning classifiers. Ensemble- based methods can be used to test and evaluate the performance, these methods may predict the attacks efficiently.

REFERENCES:

[1] O. S. Bechhofer, “web ontology language, in Encyclopedia of Database Systems,” *Springer Publ., New York*, 2009.

[2] S. T. Masoodi, F., Alam, S., & Siddiqui, “Security and privacy threats, attacks and countermeasures in Internet of Things.,”. *J. Netw. Secur. Appl*, pp. 67–77, 2019.

[3] A. M. Bamhdi, I. Abrar, and F. Masoodi, “An ensemble based approach for effective intrusion detection using majority voting,” *Telkomnika (Telecommunication Comput. Electron. Control.*, vol. 19, no. 2, pp. 664–671, 2021, doi: 10.12928/TELKOMNIKA.v19i2.18325.

[4] C. Yang, G. N. Odvody, C. J. Fernandez, J. A. Landivar, R. R. Minzenmayer, and R. L. Nichols, “Evaluating unsupervised and supervised image classification methods for mapping cotton root rot,” *Precis. Agric.*, vol. 16, no. 2, pp. 201–215, 2015, doi: 10.1007/s11119-014-9370-9.

[5] S. R. Sain, “The Nature of Statistical Learning Theory,” *Technometrics*, vol. 38, no. 4, pp. 409–409, 1996, doi: 10.1080/00401706.1996.10484565.

[6] M. U. (2019) Masoodi, F. S., & Bokhari, “Symmetric Algorithms I,” *Emerg. Secur. Algorithms Tech.*, no. 79, 2019.

[7] A. S. Ashoor and S. Gore, “Difference between Intrusion Detection System (IDS) and Intrusion Prevention System (IPS),” *Commun. Comput. Inf. Sci.*, vol. 196 CCIS, pp. 497–501, 2011, doi: 10.1007/978-3-642-22540-6_48.

[8] H. Rajadurai and U. D. Gandhi, “A stacked ensemble learning model for intrusion detection in wireless network,” *Neural Comput. Appl.*, vol. 5, 2020, doi: 10.1007/s00521-020-04986-5.

[9] I. Sumaiya Thaseen, B. Poorva, and P. S. Ushasree, “Network Intrusion Detection using Machine Learning Techniques,” *Int. Conf. Emerg. Trends Inf. Technol. Eng. ic-ETITE 2020*, pp. 1–7, 2020, doi: 10.1109/ic-ETITE47903.2020.148.

[10] C. Ambikavathi and S. K. Srivatsa, “Predictor Selection and Attack Classification using Random Forest for Intrusion Detection,” *J. Sci. Ind. Res.*, vol. 79, no. 05, pp. 365–368, 2020.

[11] I. H. Sarker, Y. B. Abushark, F. Alsolami, and A. I. Khan, “IntruDTree: A machine learning based cyber security intrusion detection model,” *Symmetry (Basel).*, vol. 12, no. 5, pp. 1–15, 2020, doi: 10.3390/SYM12050754.

[12] M. Moukhafi, K. El Yassini, and S. Bri, “Intelligent intrusion detection system using multilayer perceptron optimised by genetic algorithm,” *Int. J. Comput. Intell. Stud.*, vol. 9, no. 3, p. 190, 2020, doi: 10.1504/ijcistudies.2020.109602.

[13] M. Safaldin, M. Otair, and L. Abualigah, “Improved binary gray wolf optimizer and SVM for intrusion detection system in wireless sensor networks,” *J. Ambient Intell. Humaniz. Comput.*, 2020, doi: 10.1007/s12652-020-02228-z.

[14] S. Hosseini, “A new machine learning method consisting of GA-LR and ANN for attack detection,” *Wirel. Networks*, vol. 26, no. 6, pp. 4149–4162, 2020, doi: 10.1007/s11276-020-02321-3.

[15] F. H. Almasoudy, W. L. Al-yaseen, and A. K. Idrees, “ScienceDirect ScienceDirect Differential Evolution Wrapper Feature Selection for Intrusion Differential Evolution Wrapper Feature Selection for Intrusion Detection System Detection System,” *Procedia Comput. Sci.*, vol. 167, no. 2019, pp. 1230–1239, 2020, doi: 10.1016/j.procs.2020.03.438.

[16] I. Abrar, Z. Ayub, F. Masoodi, and A. M. Bamhdi, “A Machine Learning Approach for Intrusion Detection System on NSL-KDD Dataset,” *Proc. - Int. Conf. Smart Electron. Commun. ICOSEC 2020*, no. Icosec, pp. 919–924, 2020, doi: 10.1109/ICOSEC49089.2020.9215232.

[17] S. Gurung, M. Kanti Ghose, and A. Subedi, “Deep Learning Approach on Network Intrusion Detection System using NSL-KDD Dataset,” *Int. J. Comput. Netw. Inf. Secur.*, vol. 11, no. 3, pp. 8–14, 2019, doi: 10.5815/ijcnis.2019.03.02.

[18] Y. Zhou, G. Cheng, S. Jiang, and M. Dai, “Building an efficient intrusion detection system based on feature selection and ensemble classifier,” vol. 174, no. April, 2020, doi: 10.1016/j.comnet.2020.107247.

[19] F. S. Masoodi and M. U. Bokhari, “Symmetric Algorithms I,” *Emerg. Secur. Algorithms Tech.*, no. January, pp. 79–95, 2019, doi: 10.1201/9781351021708-6.

[20] M. A. Jabbar, R. Aluvalu, and S. S. Reddy, “RFAODE: A Novel Ensemble Intrusion Detection System,” *Procedia Comput. Sci.*, vol. 115, pp. 226–234, 2017, doi: 10.1016/j.procs.2017.09.129.

[21] L. Rutkowski, L. Pietruczuk, P. Duda, and M. Jaworski, “Decision trees for mining data streams based on the mcdiarmid’s bound,” *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 6, pp. 1272–1279, 2013, doi: 10.1109/TKDE.2012.66.

[22] I. M. A. and A. Aleroud, “SDN-Based Real-Time IDS / IPS Alerting System,” 2017.