EVALUATION OF DDOS ATTACKS DETECTION IN A CICIDS2017 DATASET BASED ON CLASSIFICATION ALGORITHMS

Amer A. Abdulrahman[1], Mahmood K. Ibrahem[2]
[1]Informatics Institute for Post Graduate Studies (IIPS), Iraq
[2] College of Information Engineering, Al-Nahrain University, Iraq
amer6567@yahoo.com[1], mahmoodkhalel@coie-nahrain.edu.iq[2]
Received:18/10/2018, Accepted:20/11/2018

***Abstract*–Intrusion detection system is an imperative role in increasing security and decreasing the harm of the computer security system and information system when using of network. It observes different events in a network or system to decide occurring an intrusion or not and it is used to make strategic decision, security purposes and analyzing directions. This paper describes host based intrusion detection system architecture for DDoS attack, which intelligently detects the intrusion periodically and dynamically by evaluating the intruder group respective to the present node with its neighbors. We analyze a dependable dataset named CICIDS 2017 that contains benign and DDoS attack network flows, which meets certifiable criteria and is openly accessible. It evaluates the performance of a complete arrangement of machine learning algorithms and network traffic features to indicate the best features for detecting the assured attack classes.**

Keywords: **Anomaly detection system, DDoS attack, CICIDS 2017 dataset, Feature selection**

I. INTRODUCTION

Intrusion Detection Systems (IDS) are the very significant protection tools against the consistently developing and ever rising network attacks. Due to the need of validation datasets and reliable test and effectiveness datasets, anomaly based intrusion detection methods are experiencing from accurate and consistent performance evolutions [1,2]. Figure 1 illustrates a general structure of intrusion detection system. The anomaly based intrusion detection system (IDS) is widely used dependent on various machine learning algorithms. The IDS is normally evaluated by its ability to make accurate predictions of attacks. There are four possible outcomes in case of the binary classifier IDS. The aim of this research is to describe a new IDS dataset namely CICIDS2017, which contains 225711 samples from DDoS attacks first. Secondly, analyze the normalized dataset to select the best amount of flow packet feature sets to detect attack and also we implemented some common machine learning algorithms to evaluate this dataset.

II. DISTRIBUTED DENIAL OF SERVICE (DDOS)

A distributed denial-of-service (DDoS) attack is a malicious attempt to damage normal traffic of a targeted server, service or network sending them huge packets. DDoS attack achieves effectiveness by using multiple compromised computer systems simultaneously as sources of attack traffic, this will not allow victim to receive the imperative data from the network and this will totally consume the victim bandwidth. Figure 2 shows the structure of DDoS attack [3].

III. DATASET

The Canadian Institute for Cybersecurity Intrusion Detection System dataset (CICIDS 2017) has the latest attributes with new types of attacks. In this section we have described the dataset that contains the DDoS attacks and we have used for training models [4]. This dataset is completely classified with more than eighty network traffic features extracted

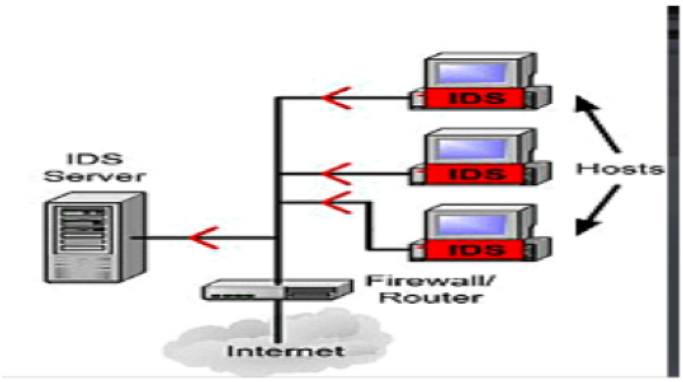

Figure 1: Intrusion detection system

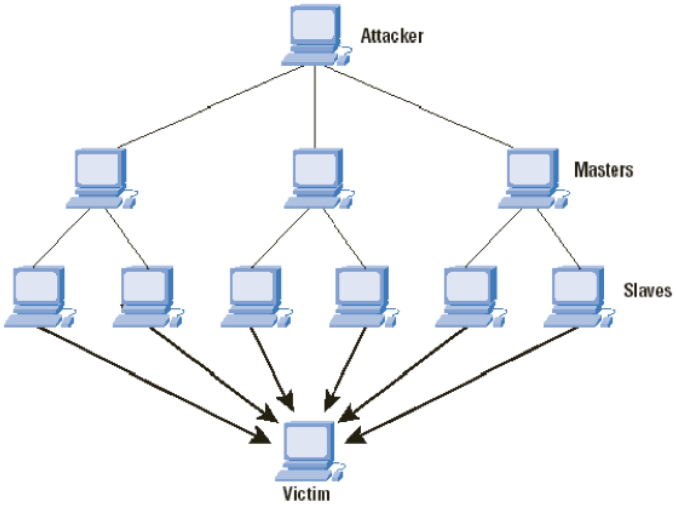

Figure 2: Structure DDoS attack

and computed for all benign and attacks flows by utilizing software named CICFlowMeter which is available publicly in Canadian Institute for Cybersecurity website [5]. It generates Bidirectional Flows, where the first packet determines the forward and backward directions. The 84 statistical features such as Duration, Number of packets, Number of bytes, Length of packets, etc are also calculated separately in the forward and backward direction. The first six columns labeled for each flow, namely FlowID, SourceIP, DestinationIP, SourcePort, DestinationPort, and Protocol with more than 80 network traffic features. We Note that TCP flows are usually terminated upon connection teardown while UDP flows are terminated by a flow timeout. The flow timeout value can be assigned arbitrarily by the individual scheme, e.g. 600 seconds for both TCP and UDP. The output of the application is the CSV file format

IV. PREPROCESSING DATASET

Preprocessing techniques are an important stage use to handle real data into an understandable format, data are inconsistent (having errors, outlier values) and incomplete, before applying data mining techniques there is a need for preprocessing methods to enhance the quality of the data and to improve the accuracy and efficiency of subsequent data mining job. The preprocessing procedures are vital and essential in the analysis of network traffic due to the patterns of this traffic which have different styles of format and dimensions .preprocessing procedures that are used, as Data cleaning ,Data integration , Data reduction ,Data discretization and Data transformation (normalization) techniques [6]. Many techniques for Data Normalization are used like min-max, z-score and decimal scaling normalization. The normalization processing has applied to the numerical features by utilizing several approaches such as Min-Max normalization algorithm. It is very important to enhance the effectiveness and performance of the system by changing all the attribute values within particular scope of [0, 1]. However, it experiences anomaly affectability.

$$Z = ((xi - min(x))/(max(x) - min(x))) \tag{1}$$

Where, xi is the data element, min(x) is the minimum of all data values, and max(x) is the maximum of all data values, Z is a new value [7]. CICIDS 2017 dataset has some missing values, which causes error in normalization process. Missing value has been processed before performing normalization process.

V. FEATURE SELECTION

Feature selection is a procedure to find a subset of significant features from the original set of features and reduces the number of irrelevant redundant features from dataset to enhance the performance of the classification and also decreases storing of memory space [8]. Feature selection helps in understanding data, reducing the effect of curse of dimensionality, reducing calculation requirement, enhancing the accuracy of learning and distinguishing which features may be relevant to a particular issue[9]. There are several methods in supervised feature selection that can be extensively categorized into wrapper, filter and embedded models [9]. "One of the most common filter model methods in feature selection is information gain which measures the information gain of each attribute by evaluating the worth of an attribute based on entropy with respect to the class. The attributes which have higher entropy are the more information content." Table I shows the best ten important attribute from 80 features of CICIDS2017 dataset that we extracted to perform our evaluation based on information gain method.

VI. MACHINE LEARNING ALGORITHMS

Machine learning provides a set of algorithms that transform data into actionable knowledge". It is best when it expands as opposed to replaces the specific knowledge of a of a topic master". "A predictive model is utilized for tasks that include, as the name implies, "the prediction of one value using other values in the dataset". "The learning algorithm attempts to discover and model the relationship between the objective feature and the other features." The processing of training predictive model is known as supervised learning or classification" [10]. There are several algorithms of supervised learning

TABLE I
FEATURE SELECTION

Feature name	Weight
Fwd.IAT.Total	46.083171
Flow.IAT.Max	39.047967
Active.Max	38.372911
Active.Min	37.004728
Fwd.IAT.Max	36.595626
Active.Mean	35.621885
Idle.Min	33.588032
Idle.Max	32.288567
Flow.IAT.Std	29.902196
Fwd.IAT.Mean	28.631780

such as "Decision Trees (DT), NaÃ¯ve Bayes (NB), Neural Networks (NN), Support Vector Machine (SVM), Random Forests (RF), etc". In this paper we construct four machine learning models by using C5.0, Naive Bayes, SVM and Random Forests algorithms, then compared between them to choose the best model. 6.1 C5.0 decision tree algorithm: This algorithm is an enhanced version of his prior algorithm C4.5 (j48) which itself is an enhancement over his Iterative Dichotomiser 3 (ID3). "The benefits of the C5.0 algorithm are that it is opinionated about pruning it takes care of many decisions automatically using fairly reasonable defaults". "C5.0 algorithm depends on the concept of Information entropy. The algorithm requires a set of training pairs inputs and output where the output is the relating class". Both numerical and categorical data are supported and the outcome is presented as a tree, making it readable for humans". It has many features like [10]:"

- C5.0 algorithm gives recognize on noise and missing data
- C5.0 algorithm can be viewing the large decision tree as a set of rules which is easy to understan
- C5.0 classifier can anticipate which attributes are relevant and which are not relevant in classification
- It solved the problem of over fitting and error pruning

1) Random Forests Algorithm: This algorithm is an ensemble learning classifier "(learning algorithms that build a set of classifiers and then classify new data points by taking a (weighted) vote of their predictions) that operate by building a huge number of decision trees at training time and outputting the class that is the mode of the classes of the individual trees. Random forests correct for decision trees propensity of overfitting to their training set" [11].

2) Naive Bayes Algorithm (NB): This algorithm is based on Bayes theorem which it applied on classification problems. Although it is not the only machine learning method that uses Bayesian methods, it is the most common one. This is particularly true for text classification, where it has become the defacto standard. The Bayesian classifiers methods use training data to compute an observed probability of each outcome based on the evidence provided by feature values. When the classifier is later applied to unlabeled data, it utilizes the observed probabilities to predict the most probable class for the new features[10].

3) Support Vector Machines (SVM): This training algorithm" constructs a model that allocate new examples to one classification or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of

TABLE II
CONFUSION MATRIX

		Predicted class	
		Normal	Attack
Actual class	Normal	TP	FP
	Attack	FN	TN

TABLE III
THE CONFUSION MATRIX FOR THE FOUR ALGORITHMS

C5.0		Predicted class		RF		Predicted class	
Actual class		BENIGN	DDoS	Actual class		BENIGN	DDoS
	BENIGN	36667	108		BENIGN	36639	136
	DDoS	6006	2360		DDoS	5821	2545
SVM		Predicted class		NB		Predicted class	
Actual class		BENIGN	DDoS	Actual class		BENIGN	DDoS
	BENIGN	33997	6301		BENIGN	33123	5378
	DDoS	2778	2065		DDoS	3652	2988

the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that equivalent space and predicted to have a place with a classification dependent on which side of the gap they fall" [1].

VII. EXPERIMENTAL ANALYSES

Our evaluation has been implemented on 225711 samples of CICIDS 2017 dataset which divide them on 80% for training and 20% for testing then classify the BENIGN and DDoS attack. The experiments are implemented using R studio software that is containing different libraries for machine learning algorithms. A confusion matrix is a technique for summarizing the performance of a classification algorithm. It represents true and false classification results. Calculating a confusion matrix gives a better idea of what your classification model is getting right and what types of errors it is making. The followings are the possibilities to classify events and depicted in Table II :

Where: The True Positives (TP) and True Negatives (TN) are correct classifications. False positive (FP) it occurs when the out-come is incorrectly predicted as yes (or positive) when it "is actually no (negative). False negative (FN) it occurs when the outcome is incorrectly predicted as negative when it is actually positive [12]". The performance and accuracy have been checked for the selected features in Table I with four machine learning algorithms using cross-validation 5 folds to enhancement the results. Table III shows the confusion matrix results for C5.0, RF, NB and SVM algorithms.

VIII. PERFORMANCE EVALUATION

The performance of the IDS can be evaluated using the confusion matrixes in Table III. From these confusion matrixes we have use several following common information retrieval evaluation metrics: Precision (Pr) or Positive Predictive value is the ratio of correctly classified attacks flows (TP), in front of all the classified flows (TP+FP). Recall (Rc) or Sensitivity is the ratio of correctly classified attack flows (TP), in front of all generated flows (TP+FN). Detection rate is the rate of true events also predicted to be events. The accuracy, recall, precision, detection rate and false alarm rate were calculated by using the following equations as follow: [12]

$$Accuracy = TP + TN/(TP + TN + FP + FN) \tag{2}$$

$$Recall(Rc) = TP/(TP + FN) \tag{3}$$

$$Precision(Pr) = TP/(TP + FP) \tag{4}$$

$$Detectionrate = TP/(TP + TN + FP + FN) \tag{5}$$

$$Falsealarmrate = FP/(TN + FP) \tag{6}$$

Table IV shows the performance examination results in terms of the weighted average of our evaluation metrics for the four selected common machine learning algorithms, namely Random Forest (RF), C5.0, Naive-Bayes (NB), and SVM. These results are based on the confusion matrices of Table III with performance metric equations 2,3,4,5 and 6.

TABLE IV
THE PERFORMANCE EXAMINATION RESULTS

Model	Accuracy	Recall	Precision (Pr)	Detection Rate	False alarm
C5.0	0.86457	0.85925	0.99706	0.81227	0.04637
RF	0.86803	0.86290	0.99630	0.81165	0.05072
NB	0.79996	0.90069	0.86031	0.73376	0.64284
SVM	0.79887	0.92445	0.84364	0.75312	0.75316

Of four classification algorithms for handling numerical data that were evaluated, the Random Forest (RF) and C5.0 classifiers surpasses the others with average accuracy of 86.80%, 86.45% respectively and for them the probability of success (Precision) is about 99%.. The false positive rate of RF and C5.0 are 0.050%, 0.046% respectively which means the probability of falsely rejecting the null hypothesis for the test is acceptable (less than 10%) The maximum of false positive rate is 75% in SVM algorithm which means the number of incorrectly classified instances is very high.

IX. CONCLUSION

A dependable "publicly available IDS evaluation datasets is one of the essential concerns of researchers and producers in this domain. In this paper," we have described the latest intrusion detection dataset and we presented the evaluation of its using common machine learning algorithms performance. The complexity of classification algorithms depends on the number of features and the number of training data samples. If the number of features increases, then the amounts of training data which are required are increasing.

REFERENCES

[1] Bhavsar, Y. B., and Waghmare, K. C. (2013). Intrusion detection system using data mining technique: Support vector machine. International Journal of Emerging Technology and Advanced Engineering, 3(3), 581 − 586.Nani, S. V. (2008). Computer security: A machine learning approach. Royal Holloway, University of London.

[2] Madni H., Javed M. and M.J. Arshad, "An Overview of Intrusion Detection System (IDS) along with its Commonly Used Techniques and Classifications", International Journal of Computer Science and Telecommunications Vol. 5, Issue 2, 2014
[3] Adrien Bonguet and Martine Bellaiche (2017) , A Survey of Denial-of-Service and Distributed Denial of Service Attacks and Defenses in Cloud Computing
[4] http://www.unb.ca/datasets/ids-2107.html
[5] Habibi L, A., Draper Gil, G., Mamun, M. S. I., and Ghorbani, A. A. (2017). "Characterization of tor traffic using time based features". In Proceedings of the 3rd International Conference on Information Systems Security and Privacy (ICISSP), pages 253 − 262.
[6] Duda, R. O., Hart, P. E., and Stork, D. G. (2012). Pattern classification. John Wiley and Sons..
[7] Jaina Patel J. and Mr. Panchal K., "Effective Intrusion Detection System using Data Mining Technique", Journal of Emerging Technologies and Innovative Research (JETIR) Vol. 2, Issue 6, 2015.
[8] Ahmad, I.(2015) 'Feature selection using particle swarm optimization in intrusion detection', International Journal of Distributed Sensor Networks, 11(10), pp: 806 − 954.
[9] Shardlow M., (2014) "An Analysis of Feature Selection Techniques", Journal of Machine Learning Research,.
[10] Brett Lantz, (2015) , Machine Learning with R Second Edition
[11] Tesfahun, A., and Bhaskari, D. L. (2013, November). Intrusion detection using random forests classifier with SMOTE and feature reduction. In Cloud and Ubiquitous Computing and Emerging Technologies (CUBE), 2013 International Conference on (pp. 127 − 132). IEEE.
[12] Santra A. K. and Christy C. J.," Genetic Algorithm and Confusion Matrix for Document Clustering ", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 2, January 2012.