ACHIEVING THE HIGHEST DETECTION RATE AND THE SHORTEST TIME IN BUILDING A MODEL FOR INTRUSION DETECTION IN COMPUTER NETWORK INTRUSION DETECTION SYSYTEMS

[1]HADI NAJAFI MOHSENABAD, [2]MEHMET ALI TUT

[1,2]Department of Mathematics, Eastern Mediterranean University TRNC
E-mail: hadi.najafi@emu.edu.tr, mehmet.tut@emu.edu.tr

Abstract - Intrusion detection systems (IDS) have developed into a very important research area among computer network security experts as a result of the expanding use of computer networks on the one hand and the rise in the number of attacks on computer networks on the other. In this study, a method utilizing machine learning algorithms is introduced that can detect the detection rate of secure traffic and the detection rate of Brute Force type attacks with 100% accuracy as well as the detection rate of Bot type attacks using the fewest possible identity features with more than 99% accuracy. The training and testing of the proposed methods utilized the CSE-CIC-IDS2018 data set, one of the most recent standard data sets for evaluating the effectiveness of computer network intrusion detection techniques. The CFSS ubset Eval feature selection method has allowed this method to employ less features while maintaining the same level of intrusion detection accuracy. It also takes less time to develop an intrusion detection model than other methods.

Keywords - Computer Networks, Decision Tree, Feature Selection, Intrusion Detection System.

I. INTRODUCTION

With the birth and growth of network structures, virtually every institution, organization, and business has created its own extensive local area networks (LANs). These networks contain a substantial amount of sensitive data and networks. [1].

New network and device types have also arisen, including 5Gm, cloud computing, and the Internet of Things (IoT). The expansion of these systems and networks makes it essential to protect their cyber security [2]. While traditional machine learning (ML) methods such as Bayesian Belief Networks (BBN), support vector machines (SVM), and others have been made available for cyber security, the generation of large-scale data in the Internet of Things (IoT) required a deep learning-based method that performs better with large data sizes and can learn a representation of the future from raw data in order to be adaptable to different attack scenarios. [3]. There are four distinct feature selection techniques: filter, wrapper, embed, and hybrid. The filter technique makes an effort to classify a portion of the original set that includes a number of selected features based on the assessment criteria. When compared to the wrapper method, which selected features from a variety of learning strategies with great projected accuracy? The embedded technique is used to incorporate the feature selection into the training process.

They introduced a new malware prediction model utilizing a deep learning technique dubbed Mal Generative Adversarial Network (Mal-GN), which might detect potential threats in the future [4]. Results from the KDD Cup '99 dataset challenge showed that the LSTM classifier outperformed other static classifiers for a long time, highlighting the importance of LSTM networks for intrusion detection because of their ability to learn from the past and link connection records consecutively [5]. The work of classifying data has become significantly more difficult recently due to the rapid expansion of data. By lowering the dimensionality of the data and attaining improved accuracy in data classification, feature selection is one method for resolving this problem. [6]. In terms of feature selection, it plays an important role in improving Network Intrusion Detection System (NIDS) performance. This is because a lot of time-consuming features are used in anomaly detection. As a result, the method for selecting features has an impact on the level of accuracy improvement and the amount of time needed to check traffic behavior.

There are four distinct feature selection techniques: filter, wrapper, embed, and hybrid. The filter technique makes an effort to classify a portion of the original set that includes a number of selected features based on the assessment criteria. When compared to the wrapper method, which selected features from a variety of learning strategies with great projected accuracy? The embedded technique is used to incorporate the feature selection into the training process [7]-[8].
The goal of this article is to identify a subset of features that has the fewest features and achieves the maximum accuracy in identifying attacks and traffic of the Brite Force and Bot types while posing the fewest risks and taking the shortest amount of time. In this study, 5 features out of 80 were chosen to use

as part of the CFSSubsetEval feature reduction approach to detect safe traffic and brute force assaults. This method was able to detect 100% of safe traffic and brute force attacks accurately and quickly. In order to identify attacks of this nature, it has also chosen only 4 features out of all the features, and has been successful in detecting traffic with greater than 99.5% accuracy.

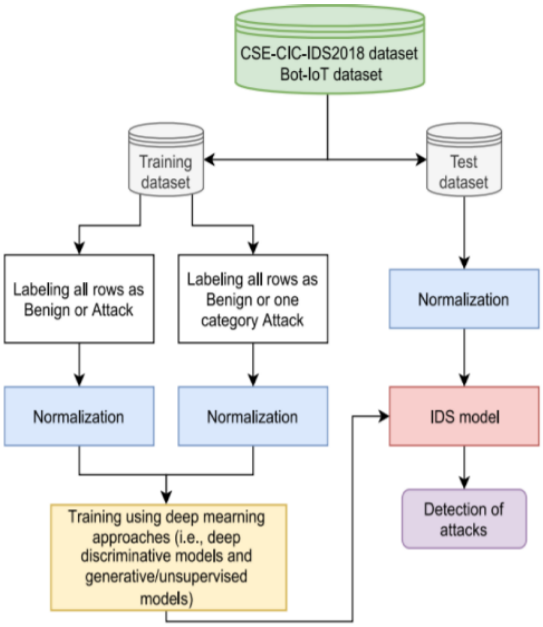


Fig.1. Flowchart of the IDS methodology [9].

II. RELATED WORKS

95 percent accuracy was eventually attained by choosing 55 features out of 80 features using the Binary Particle Swarm Optimization (BPSO) approach and the classification method with Deep Neural Network (DNN) [6]. The Generative Adversarial Network (GAN) algorithm, which has a classification accuracy of 99.34%, outperformed Decision Tree (DT), Artificial Neural Network (ANN), Multi-layer Perceptron (MLP), k-Nearest Neighbors (KNN), Convolutional Neural Networks (CNN), and Auto Encoder algorithms in an evaluation [9]. 29 of the 80 traits were picked for further analysis. Four classification algorithms were tested: baseline, logistic regression, random forest, and gradient boosting. With a 98% accuracy rate, gradient boosting emerged as the winner. Finally, 55 features out of 80 features were selected using the Binary Particle Swarm Optimization (BPSO) method and the classification method with Deep Neural Networks, yielding a final accuracy of 95%. (DNN) [10]. For classification, the artificial neural network (ANN) algorithm has been utilized, and it has been evaluated against other algorithms. The ANN algorithm has the best accuracy, coming in at 99.91% [11]. In the article [12], using 10 features and evaluating four random forest algorithms, ADABoost, Naive Bayes, and artificial neural network (ANN), the best accuracy of 100% is obtained with the random forest algorithm.The best result was obtained when 32 features were taken into consideration for classification, with 98.83% accuracy. 32, 64, and 78 features were employed, respectively [13].Only the identification of DDoS-type attacks utilizing the article's Neural Network technique managed to reach 100% accuracy [14]. The maximum accuracy of 98.20% was attained using the 12 features that were chosen using the CICFlowMeter feature selection method, along with the reviewed algorithms [15]. Convolutional Neural Network (CNN), Long Short-term Memory, and Apache Spark were utilized in [16], which used 19 features and classification techniques. Apache Spark was able to reach 100% accuracy.The Long Short-Term Memory (LSTM) algorithm's maximum accuracy was 99% [17]. Deep Neural Network (DNN), Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), Restricted Boltzmann Machine (RBM), Deep Belief Network (DBN), Deep Boltzmann Machine (DBM), and Denoising Auto Encoder algorithms were considered in [18] and the accuracy of 98.38% was the highest accuracy among these algorithms.

III. FEATURE SELECTION

Feature selection plays and important role in machine learning. The best selection of significant feature set result in better performance of classification.

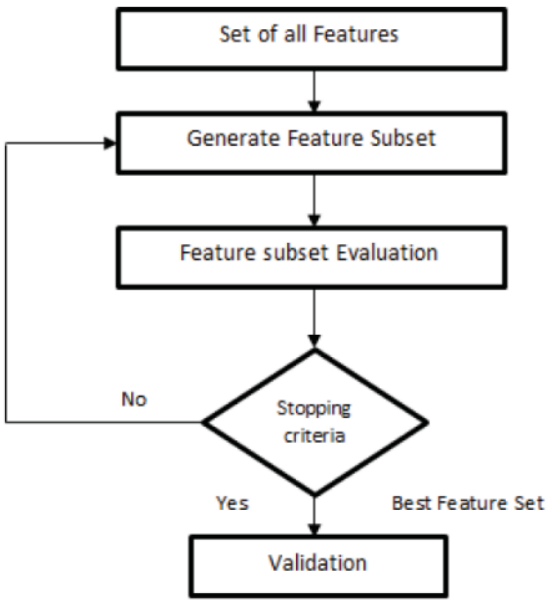


Figure2. Feature selection steps [19].

CFSSubsetEval feature selection method is used in this research, Evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low intercorrelation are preferred.

IV. DATASET

For use in attack detection and prevention, CSE-CIC-IDS2018 was created by the Canadian Institute for Cyber Security (CIC) and the Communications Security Establishment (CSE). It is published as open source and represents the traffic flow and attack kinds of actual network architecture [20]. There are 80 features in total, including tags. The majority of the features are based on statistical information about network traffic that was gathered using the CICFlowMeter and output. [21]. Almost all of the features in the dataset consist of continuous features such as dst_ports, protocol, timestamps, and labels. It includes 14 types of attacks in 6 different scenarios.

Category	Attack Types
Benign	
DDoS	HOIC
	LOIC-http
	LOIC-UDP
Botnet	Bot
Brute Force	FTP
	SSH
Infilteration	Infilteration
Dos	Hulk
	SlowHTTPTest
	GoldenEye
	Slowloris
Web Attacks	Web
	XSS
	SQL Injection

Table1: Types of attacks and safe traffic in data set CSE-CIC-IDS2018

Traffic class	Record count	Share (%)
Benign	13 484 708	83.070%
HOIC[1]	686 012	4.226%
LOIC-HTTP[1]	576 191	3.550%
Hulk[1]	461 912	2.846%
Bot	286 191	1.76%
FTP-BruteForce	193 360	1.191%
SSH-Bruteforce	187 589	1.156%
Infilteration	161 934	0.998%
SlowHTTPTest[1]	139 890	0.862%
GoldenEye[1]	41 508	0.256%
Slowloris[1]	10 990	0.068%
LOIC-UDP[1]	1 730	0.011%
Brute Force-Web	611	0.004%
Brute Force-XSS	230	0.001%
SQL Injection	87	0.0005%

Table 2: The number of records in the CSE-CIC-IDS2018 data set

V. BRUTE FORCE ATTACK

With a Brute Force attack, the attacker attempts to crack a password or username using an “exhaustive search” or trial error approach in cryptography, a Brute Force attack consists of the attack submitting many passwords or passphrases, systematically checking all possibility until the correct one is found. Brute Force attacks are usually aimed at obtaining personal information such as password, passphrases, username and personal identification members (PINS).

VI. BOT ATTACK

An automated web request is used in a bot attack to trick, deceive, or otherwise interfere with a website, application, API, or end users. Bot assaults began as straightforward spamming operations but have since developed into sophisticated, global criminal organizations with independent economies and infrastructures.

VII. EXPERIMENTAL DESIGN

The test process was carried out on Intel(R) Core(TM) i7-9750H CPU @ 2.60GHz 2.59 GHz with 16 GB Ram and Windows 11 Enterprise operating system. It was carried out using the Weka application.

VIII. METHODOLOGY

There are 80 features in all, so we first need a method to choose the ones that will be most helpful to us based on the type of attack and secure network traffic in this research. This study also made use of the Weka software, one of the most effective packages for machine learning algorithms. The method for feature selection employed was CFSSubsetEval. Using the CFSSubsetEval feature selection method, we were able to extract 5 characteristics that can recognize brute force attacks. They are as follows:

1) Dst Port : Destination port number
2) Flow IAT Mean : average time between two flows
3) InitFwd Win Byts : number of bytes sent in initial window in forward direction
4) InitBwd Win Byts : number of bytes sent in initial window in backward direction
5) FwdSeg Size Min : minimum segment size observed in forward

The results of 11 different algorithms using the evaluation criteria of the algorithms used were evaluated, and the results of the evaluation of the algorithms are shown in the table below: 80% of this dataset was used for training the machine using these 5 features, and 20% of this dataset was used for testing the machine:

a) True Positive (TP): These are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also yes..
b) True Negative (TN): These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no.
c) False Positive (FP): When actual class is no and predicted class is yes.
d) False Negatives (FN) – When actual class is yes but predicted class in no.

Accuracy: Accuracy which is another name for positive predictive value, is the fraction of related samples among the retrieved samples.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.

$$Precision = \frac{TP}{TP + FP}$$

Recall (Sensitivity): The sensitivity is calculated mathematically by dividing the true positives by the total of the true positives and false negatives.

$$Recall = \frac{TP}{TP + FN}$$

F-measure (F1Score): In fact, F-measure is a kind of average between parameter p (precision) and parameter r (recall). p is the accuracy of the system among the predicted data. r is the ratio of the number of predicted data to the total number of expected data for prediction.

$$F1Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Time taken to build model: The time required for the algorithm to build the model based on the data given to it for training.

Algorithm	Time taken to build model	Accuracy	Precision	Recall	F-measure
J48	5.39s	1	1	1	1
Random Forest	133.95s	1	1	1	1
BayesNet	5.32s	1	1	1	1
NaiveBayes	1.05s	99.98	1	1	1
SMO	21.8s	99.83	1	0.997	0.999
Logistic	21.33s	99.96	1	0.999	1
Classification via Regression	813.71s	99.99	1	1	1
LogitBoost	14.31s	99.99	1	0.999	1
JRip	8.64s	1	1	1	1
OneR	1.2s	99.96	1	0.999	1
DTNB	32.6s	1	1	1	1

Table3: The result of evaluating different algorithms for Brute Force attack

The BayesNet method, which was able to create the model quickly and achieve 100% accuracy in identifying Brute Force attacks, is the best approach, as you can see in the table above.If speed is more essential to us, the OneR approach, which takes 1.2 seconds to create a model to detect Brute Force attacks, has the fastest detection time for Brute Force attacks.We were able to extract the following four features that are the most useful features to detect Bot type assaults by using the feature selection method indicated in the preceding stage (CFSSubsetEval):

1) Dst Port : destination port number
2) FwdPktlen Min : number of forward packets
3) Fwd PSH Flags : number of time the PSH flag was set in packets in forward direction
4) Active Max : max a flow was idle before becoming active

Using these four features, we again used 80% of the data for training the system and the other 20% for evaluating and testing the system and obtained the results of 12 machine learning algorithms as described in the table below.

Algorithm	Time taken to build model	Accuracy	Precision	Recall	F-measure
Decision Table	8.77 s	99.59	0.994	1	0.997
Bagging	40.31s	99.59	0.994	1	0.997
AdaBoostM1	14.71s	99.57	0.994	1	0.997
REPTree	40.31s	99.59	0.994	1	0.997
NaiveBayes	0.78s	62.49	0.998	0.489	0.654
BayesNet	2.64s	99.57	0.994	1	0.997
JRip	10.25s	19.57	0.994	1	0.997

Classification via Regression	1407.7s	99.57	0.994	1	0.997
DTNB	24.2s	99.59	0.994	1	0.997
Random Forest	184.1s	99.58	0.994	1	0.997
J48	5.23s	99.59	0.994	1	0.997
Random Tree	2.34s	99.58	0.994	1	0.997

Table4: The outcome of comparing various bot attack algorithms

By evaluating the machine learning algorithms that can be seen in the table above, the best accuracy for detecting Bot type attacks is obtained by the J48 algorithm with 99.59% accuracy in 5.23 seconds, on the other hand, if time is more important to us the shortest time to build the model belongs to the NaiveByes algorithm with a time of 0.78 seconds, but the accuracy has obtained is very low 62.49%, so the lowest time with good accuracy can be obtained for the BayesNet algorithm with the time to build the model in 2.64 seconds and with 99.57% accuracy.

IX. CONCLUSION

In this research, it was attempted to use the fewest number of features among the 80 features in the CSE-CIC-IDS2018 data set, the highest level of accuracy in detecting intrusions in the form of Brute Force attacks and Bot attacks, and secure network traffic in the quickest amount of time to build the model. The CFSSubsetEval feature selection method has been used to choose the feature for this purpose, which is to identify attacks. Five features have been chosen to detect Brute Force attacks using the CFSSubsetEval feature selection approach, and four features have been chosen to detect Bot assaults from secure network traffic. Then, using 5 features and 11 machine learning techniques, 80% of the data set was used for training the system and 20% for testing it. The BayesNet algorithm delivered the best outcomes, securing network traffic and detecting Brute Force attacks with a 100% accuracy rate. It's also important to note that 5.32 seconds were needed to build the model for 80% of the data. Additionally, we evaluated the system using 20% of the features after training it with 80% of them. Finally, we compared the outputs of 12 machine learning algorithms, with the best output having an accuracy of 98.59%. With the J48 algorithm, we were able to achieve, and this algorithm takes 5.23 seconds to complete the model construction. Weka, one of the most potent applications in the field of data mining, is the program that we employed for this study.

REFERENCES

[1] Ch. MA, X. Du and L Cao, “Analysis of Multi-Types of Flow Features Based on Hybrid Neural Network for Improving Network Anomaly Detection”, IEEE Access, 7, 1-18, 2019.

[2] L. Liu, P. Wang, J.Lin and L. Liu, “Intrusion Detection of Imbalanced Network Traffic based on Machine Learning and Deep Learning”, IEEE Access, 9, 1-14, 2021.

[3] H. Liu and B. Lang, "Machine Learning and Deep Learning Methods for Intrusion Detection Systems: A Survey", applied science, 9,439, DOI: 10.3390/app9204396, 2019

[4] S. Lu, E. Al., “New Era of Deep learning -Based Malware Intrusion Detection: The Malware Detection and Prediction Based on Deep Learning”, ArXiv, abs/1907.08356, 2019

[5] R. C. Staudemeyer,"Applying long short-term memory recurrent neural networks to Intrusion detection", SACJ, No. 56, July 2015.

[6] R. I. Farhan, A. T. Maolood, N. F. Hassan “Optimized Deep Learning with Binary PSO for intrusion Detection on CSE-CIC-IDS2018 dataset,” Journal of Al-Qadisiyah for Computer Science and Mathematics Vol.12 (3), 2020

[7] N. Chockwanich, V. Visoottiviseth, "Intrusion Detection by Deep Learning with TensorFlow", International Conference on Advanced Communications (ICACT), 2019.

[8] A. Nisioti, A. Mylonas, P. D. Yoo, V. Katos, "From Intrusion Detection to Attacker Attribution: A Comprehensive Survey of Unsupervised Methods", DOI: 10.1109/COMST.2018.2854724, IEEE, 2018

[9] P. Prabakaran, R.S. Mohana “Enhancing the Cyber Security Intrusion Detection based on Generative Adversarial Network” Ilkogretim Online - Elementary Education Online; Vol 20 (Issue 5): pp. 7401-7408, 2021

[10] [10] A. Lava and P. Savant, “Network-Based intrusion detection systems using machine learning algorithm” International Journal of Engineering Applied Sciences and Technology, Vol. 6, Issue 11, ISSN No. 2455-2143, Pages 145-155, 2022

[11] Y. Ayachi, Y. Mellah, J.Berrich and T. Bouchentouf “Increasing the performance of an IDS using ANN model on the realistic cyber dataset CSE-CIC-IDS2018,” International Symposium on Advanced Electrical and Communication Technologies (ISAECT) | 978-1-6654-2222-2/20/$31.00 ©2020 IEEE | DOI: 10.1109/ISAECT50560.2020.9523662, 2020

[12] M. Komisarek, M. Pawlicki, R. Kozik and M. Choras, “How to effectively collect and process network data for intrusion detection,” Entropy 2021, 23(11), 1532; https://doi.org/10.3390/e23111532 - Nov 2021.

[13] M. A. Ferrag, H. Janicke, L. Maglaras and R. Smith “Deep Learning Techniques for Cyber Security Intrusion Detection: A deatiled Analysis,” 6th International Symposium for ICS & SCADA Cyber Security Research 2019 (ICS-CSR), 10th-12th September 2019.

[14] I.l. Kurochkin and S SVolkov “Using GRU based neural network for intrusion detection in software-defined networks,” IOP Conference Series Materials Science and Engineering 927(1):012035”,DOI:10.1088/1757-899X/927/1/012035, 2020.

[15] M. Khan and J. Kim “Toward Developing Efficient Conv-AE-Based intrusion detection system using Heterogeneous,” Electronics 2020, 9(11), 1771; https://doi.org/10.3390/electronics9111771, 2020.

[16] A. A. Hagar and B.W. Gawali “Apache Spark and Deep Learning Models for High-Performance Network Intrusion Detection Using CSE-CIC-IDS2018” Hindawi Computational Intelligence and Neuroscience, Volume 2022, Article ID 3131153, 11 pages, 2022.

[17] A. D. Farhan , B. I. Farhan “Performance analysis of intrusion detection for deep learning model based on CSE-CIC-IDS2018 dataset,” Indonesian Journal of Electrical Engineering and Computer Science 26(2):1165-1172,DOI:10.11591/ijeecs.v26.i2.pp1165-1172, 2022.

[18] B. Budler and R. Ajoodha "Comparative analysis of deep learning models for network intrusion detection system," 2022 IEEE 2nd Conference on Information Technology and Data Science (CITDS), DOI: 10.1109/CITDS54976.2022.9914128, 2022.

[19] K.R. Pushpalatha and A. G. Karegowda "CFS based feature subset selection for enhancing classification of similar looking food grains a-filter approach." 2017 2nd International Conference on Emerging Computation and Information Technologies (ICECIT), DOI: 10.1109/ICECIT.2017.8453403 , 2017.

[20] University of New Brunswick (UNB). A realistic cyber defense dataset (CSE-CIC-IDS2018), https://www.unb.ca/cic/datasets/ids-2018.html , 2018.

[21] S. Arsalan "FastTrafficAnalyzer: An Efficient Method for Intrusion Detection Systems to Analyze Network Traffic," DUJE (Dicle University Journal of Engineering) Page 565-572, 2021.

[22] H. Najafi, A. Ashghari "Intrusion detection in computer network using data mining techniques based on feature selection," 1st international conference on new research achievements in electrical and computer engineering, 2016.

★★★