İSTANBUL UNIVERSITY
CERRAHPAŞA

RESEARCH ARTICLE

Electrica 2020; 20(2): 159-167

Detection DDOS Attacks Using Machine Learning Methods

Tuğba Aytaç[1], Muhammed Ali Aydın[2], Abdül Halim Zaim[1]

[1]Department of Computer Engineering İstanbul Commerce University, İstanbul, Turkey
[2]Department of Computer Engineering, İstanbul University-Cerrahpaşa, İstanbul, Turkey

Cite this article as: Aytaç T, Aydın MA, Zaim AH. Detection DDOS Attacks Using Machine Learning Methods. *Electrica,* 2020; 20(2): 159-167.

ABSTRACT

Wishing to communicate with each other of people contributes to improving technology, and it has made the internet concept an indispensable part of our daily life. Cyber attacks from extranets to enterprise networks or intranets, which are used as personal, can cause pecuniary loss and intangible damage. It is critical to take due precautions for minimizing the losses by early detection of attacks. This study aims to analyze the rate of success in the intrusion detection system by using different methods. In this study, the CICDDoS2019 data set has been used, and DDOS attacks in this data set were compared. The success rates of threat determination were analyzed as using Artificial Neural Networks (ANN), Support Vector Machine (SVM), Gaussian Naive Bayes, Multinomial Naive Bayes, Bernoulli Naive Bayes, Logistic Regression, K-nearest neighbor (KNN), Decision Tree (entropy-gini) and Random Forest algorithms. It has been seen that the highest of the success rate is the models that ensure almost 100% success that was made by using K-nearest neighbor, Logistic Regression, Naive Bayes, (Multinomial – Bernoulli algorithms).
Keywords: CICDDoS2019, intrusion detection system, machine learning methods

Corresponding Author:
Tuğba Aytaç

E-mail:
tugba.aytac@istanbulticaret.edu.tr

Received: 15.05.2020

Accepted: 25.05.2020

DOI: 10.5152/electrica.2020.20049

Introduction

The internet, which is an indispensable factor in our daily life with improved technology, takes a great space in banking, health, many other industries, and our social life. It is foreseen that almost 50 billion objects will be dependent on the internet, a network system until 2020 with ARPANET (Advanced Research Projects Agency Network), which appeared in 1969, ensures communicating the tiny devices, and provides a basis of internet [1].

Many systems or software such as antivirus and firewalls have been used for protecting the enterprise networks from malicious people. However, these measures can remain incapable of averting the attacks. Because some people who want to lose our reputation or financial gain and work as a team always to find the weaknesses for detecting the deficits in systems.

It is essential to early detecting the attacks for systems and preventing them. Intrusion Detection Systems (IDS) are software or hardware constituents that qualify an "alarm" for protecting the information systems toward the attacks from the network. IDS can prevent attempts from entering the systems by detecting the misusing and unauthorized access to systems [2].

There are two different methods of IDS design [3]. These two approaches are a signature recognition based system that determines with the special characters for every behavior and the system which determines the attacks by examining the abnormal network traffic on networks. The primary aim of these two approaches is to keep definite attacks in quarantine to determine attacks as close to real-time and eliminate the damage caused by attacks [4].

The study aims is to design the establisher system in the overachievement rate as earliest as possible with the machine learning methods of DOS attacks. Machine learning methods from Anomaly Based Attack Determiner Systems data set were used in this study, and the education had been done in those data sets. It was aimed to find the fastest determination values with the overachievement rate by composing quadruplet, sestet, octuplicate, denary, and duodenary data sets. It is essential to determine the attack as keeping narrow of attribute values in terms of the IDS performance.

Intrusion Detection Systems

Gradually increasing of using the internet with improving the information technologies and taking place in a depended network together of systems raise the needs of making provi-

sion against the attacks which can come from the outside. Attacks can be identified as being precluded of a system's privacy, integrity, or accessibility. Intrusion detection systems are designed to recognize attacks on smart systems in the same network. They are used by corporate firms, companies, and governments (computer, tablet, mobile phone, etc.) to stall the systems in the network coming from the Internet or local area network, which are composed of various packets and unnecessary traffic density. Moreover it is a security technology that avertssuch attacks [5].

Attacks can be determined as examined by two approaches. Anomaly Based Attack Determiner Systems (ABADS) and Signature Based Attack Determiner Systems (SBADS) are two approaches. While ABADS determines the traffic which has harmful content as a result of the analysis of network traffic, SBADS examines based the previous data, which were recorded on the system and confirms the attack.

Machine Learning Methods and Studies On IDS

There are many studies about intrusion detection systems made by machine learning methods that have the property of decision-making and occur in the subassemblies of artificial intelligence. Firstly the anomaly-based intrusion detection system, which also machine learning methods take place in, identifies the average values then decides the incoming data are normal or abnormal, and it can make a classification.

The reason for being more preferred to our abnormal-based intrusion detection system is that the signature-based intrusion detection system can be immobilized according to previously recorded values and lower the success rate and the determination time. In this part, the machine learning methods which were used in the work and studies about these methods were discussed.

Artificial Neural Networks

Artificial neural networks in subassemblies of artificial intelligence are based on a smart algorithm that stimulates the learning property and reflectiveness of people. It is composed of multilayer or single-layer constructs created by neural networks methods such as neural networks in the human brain, which ensure learnability and reflectivity. There are five parameters that produce artificial neural networks. These parameters are inputs, weights, additional functions, activation functions, and outputs [6].

The construct of artificial neural networks is seen in Figure 1.

The highest success rate, 97.92 %, was attained in dos attacks (Smurf, Neptune, Back, Teardrop) rl2, and probe attacks by using artificial neural networks KDD'99 data set. We could attain exact results with quite a low error rate by using test sets and different education, thanks to the YSA method in the study [2].

Murat H. SAZLI and Haluk TANRIKULU trained DARPA data sets as generating multilayer artificial neural networks with "Neural Network Tools Box" in MATLAB program on their studies. They procured to immobilize Dos attacks with a high success rate via a well-designed attack determination system [7].

The reason for being preferred YSA methods in IDS works is having many advantages such as the high categorizing ability of the subject method, making inference for new data from previous learning, completing the incomplete knowledge data, and learning new cases. Absence of composing artificial neural network rules, taking a long time for learning due to making a much trying are its disadvantages.

Naive Bayes Algorithm

Naive Bayes, which is a classification algorithm according to the data category, takes its name from Thomas Bayes (1701 - 7 April 1761). Its success rate is quite high since the Naive Bayes algorithm transcludes the highest rate value into categorization by calculating the whole probabilities. If the user data have multiple class Multinomial Naïve Bayes, if there is a normal distribution in data Gauss Naïve Bayes, if making forecasts is being wanted dually Bernoulli Naïve Bayes can be preferred. Besides its advantages such as fast resulting, working with data which can show high rate reality and variability, inability to model the relations between variables is its disadvantage [8].

In the attack determination system, which made via the Naïve Bayes algorithm by using the KDDCup'99 data set, detecting the attacks system in 1.89 seconds with a 95 % success rate achieved, and it was shown that KNN (K-means clustering algorithm) and YSA methods are better [9].

It was provided to determine the DDOS attacks on the web of Shital K. Ajagekar, and Vaishali Jadhav in the range of 79% - 99,5% by categorizing with Multinomial NB algorithm and the highest rate success was gained in determining HTTP-flooding attack [10].

K The Nearest Neighbor Algorithm (KNN)

It is an algorithm that calculates the proximity of attribute of the previous k quantity of the new attributes on the phase of the classifying of attributes which data have. By choosing different attributes with NSL-KDD data set, kNN-1, kNN-2, J-48, and Naive Bayes algorithms had been tried. It is seen that the highest success rate obtained from on qualified and original data set PCA 21 with kNN-1, J-48 algorithms [11].

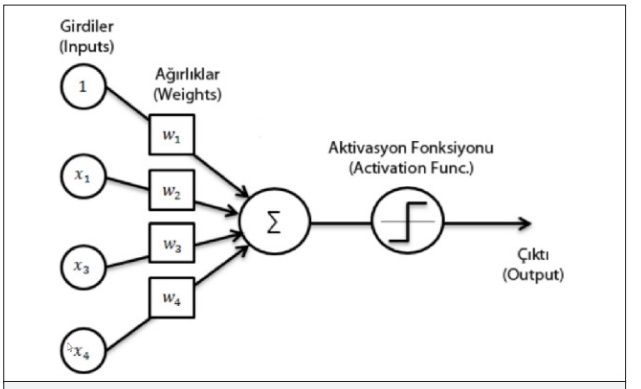

Figure 1. The Construct of Artificial Neural Set Cell

Table 1. The best attributes and success rates among the results were stated in

Property	Explanation
Best Values with 4 Property (Success rate: 0,993%)	
Tot Fwd Pkts	Total number of packets in the forward direction
Tot Len Bwd Pkts	Total length of packets in the backward direction
Bwd IAT Mean	Average time between two packets sent backwards
Fwd Seg Size Min	Minimum size observed forward
Best Values with 6 Property (Success rate: 0,994%)	
Bwd Pkt Len Std	Standard deviation of packet lengths in the backward direction
Flow IAT Max	Maximum arrival time of the packets
Pkt Len Min	Minimum length of a stream
Down/Up Ratio	Download and upload rate
Fwd Byts/b Avg	Average number of bytes forwarded
Fwd Seg Size Min	Minimum size observed forward
Best Values with 8 Property (Success rate: 0.973)	
Bwd Pkt Len Max	Maximum length of packets in the backward direction
Flow IAT Std	Standard deviation of arrival times of packets
Flow IAT Min	Minimum arrival time of packets
Fwd IAT Tot	Total time between two forward packets
Bwd IAT Std	Standard deviation of time between two packets sent backwards
Fwd Pkts/s	Number of forward packets per second
Pkt Len Std	Standard deviation of a stream
SYN Flag Cnt	Number of packets containing SYN
Best Value with 10 Property (Success rate: 0.979)	
Fwd Pkt Len Mean	Average length of forward packets
Bwd Pkt Len Min	Minimum length of forward packets
Bwd Pkt Len Mean	Average length of backward packets
Fwd IAT Max	Maximum time between two forward packets
Bwd PSH Flags	Number of active PSH flag in packets moving backwards (0 for UDP)
Fwd Pkts/s	Number of forward packets per second
FIN Flag Cnt	Number of packages containing FIN
SYN Flag Cnt	Number of packages containing SYN
PSH Flag Cnt	Number of packages containing PUSH
Down/Up Ratio	Download and upload rate
Best Values with 12 Property (Success rate: 0,998)	
Bwd Pkt Len Min	Minimum length of back packets
Flow Pkts/s	Number of packets flowing per second
Flow IAT Min	Minimum arrival time of packets
Fwd IAT Tot	Total time between two forward packets
Bwd IAT Std	Standard deviation of time between two packets sent backwards
Fwd URG Flags	Number of active URG flag in packs moving forward (0 for UDP)
Fwd Header Len	Total bytes used for forward header
Bwd Header Len	Total bytes used for backward headers
Bwd Pkts/s	Number of backward packets per second
Subflow Bwd Pkts	Average number of packets in a backward downstream
Fwd Act Data Pkts	Number of packets with at least 1 byte of TCP data carrying capacity
Fwd Seg Size Min	Minimum size observed forward

(Decison Tree) Algorithm

Classifying method decision nodes made by a using decision tree algorithm is composed of cut sheet and axis. Decision nodes or cut sheets occur depending on whether the classifying materialized or not. Attack determination had been done via the decision tree algorithm with the 99% high success rate in a study made by using the CICIDS2017 data set. It was seen that Denial of Service (Dos), Spanned Service Rejection (DDOS), and Port Scanning (Port Scan) in data set are seen as abnormal, and 78 traffic properties were used to determine it [12].

(Random Forest) Classifying Algorithm

Random forest classifying is a community classifier that produces a multi-decision tree using a subset of variables and selection education [13]. LGBM, CNN, and Random Forest methods, which were achieved by using CSE-CIC-IDS2018 data set, had been tried, and it was seen that bi-level hybrid construct created by random forest model has the highest success with the 0.86 F-score macro avg [14].

Support Vector Machines (SVM)

Support Vector Machines algorithm is an algorithm used to distinguish two classes, ideally on the base. Wani and his friends determined DDOS attacks on the cloud computing context by using SVM, Random Forest, and Naïve Bayes methods on their works in 2019 [15].The most successful result was taken from the SVM algorithm with 0.998 F-score using 9 attributes within the data cluster created for the study [15].

Data Set

Many data sets are using in studies that are made using different algorithms in Intrusion Detection System designs. CICDDoS2019 data set is the latest designed data set which was shared by Canada Cyber Security Institute, was prepared in a proper test context, was designed by having regard to imperfections in previous data sets [16].

Along with good natured and the latest DDOS attacks which are similar to real data (PCAP), CICDDoS2019, also include the result of network traffic analyses. Network traffic analyses use CICFlowMeter-V3 that has labelled traffic justified to attack (CSV documents), time stamping, originator and target IPs, originator and target ports, protocols. There are some different DDOS attack kinds such as Port Map, NetBIOS, LDAP, MSSQL,

Table 2. Algorithms used for ddos attack detection, success rates and detection times

Algorithm used for training	Success Rate (%)	Detection Time (sn)
Logistic Regrasyon	99,8	788
Gaussian Naive Bayes	98,7	1041
k-nearest neighbor	99,9	1040
Multinomial Naive Bayes	99,1	1041
Bernoulli Naive Bayes	99,8	1042
Decision Tree(entropy)	99,12	1043
Decision Tree(gini)	99,34	1043
Random Forest	98,4	1047
SVM	99,7	1074

Table 3. Algorithms detection times and success rates

Property	Explanation
Tot Bwd Pkts	Total number of packets in the backward direction
Fwd Pkt Len Min	Minimum length of forward packets
Bwd Pkt Len Min	Minimum length of backward packets
Bwd Pkt Len Mean	Average length of backward packets
Flow Byts/s	Number of bytes flowing per second
Flow Pkts/s	Number of packets flowing per second
Flow IAT Std	Standard deviation of arrival times of packets
Bwd IAT Mean	Average time between two packets sent backwards
Fwd Header Len	Total bytes used for forward headers
Pkt Len Std	Standard deviation of a flow
Pkt Len Var	Length variance of a flow
CWE Flag Count	Number of packets containing CWE

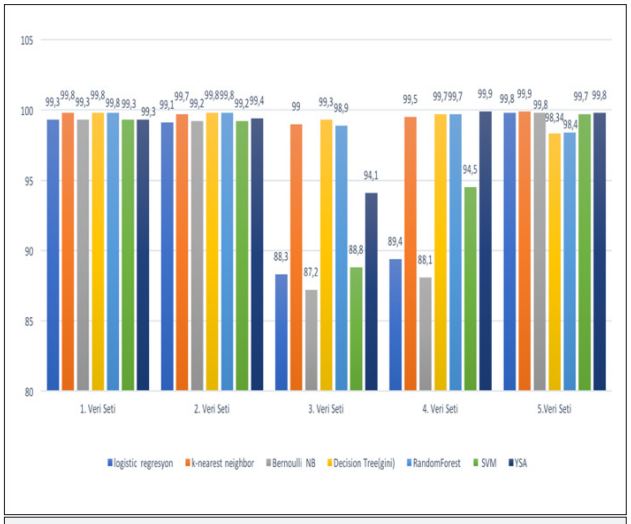

Figure 2. Graph for comparing the accuracy rates of algorithms

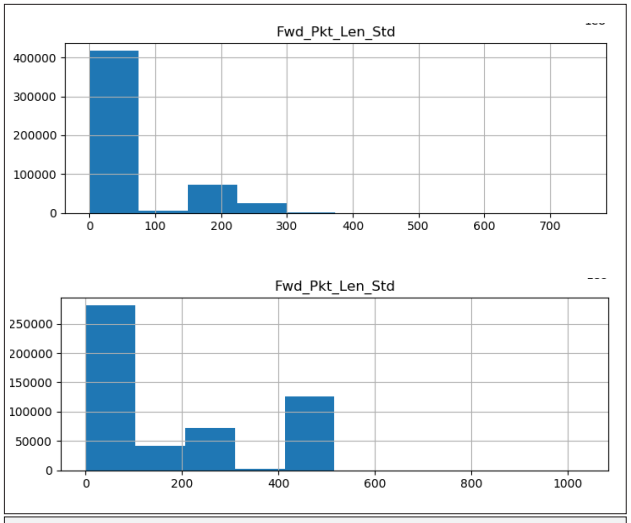

Figure 3. Graph of the best feature in the First Data Set

UDP, UDP-Lag, SYN, NTP, DNS, and SNMP in this data cluster [16].

Studies That Made By Using Data Set

Properties that subtrahend for choosing the best detection property from 80 properties in every DDOS attack in the first study data set with producing the CICDDoS2019 data set was tested using Random Forest Regressor. Finding the best properties to determine DDOS attacks with the 11 kinds of new data sets in this study [17].

Adhibition

Good natured (benign) values and DDOS attacks that have 83 attributes were examined using the CICDDoS2019 data set in this study. Since many values in data set such as space, string, port numbers were unavailable, they had been ejected via being subject to preprocessing and had been optimized for trainable with C and C++ programs. The ready data set is created in

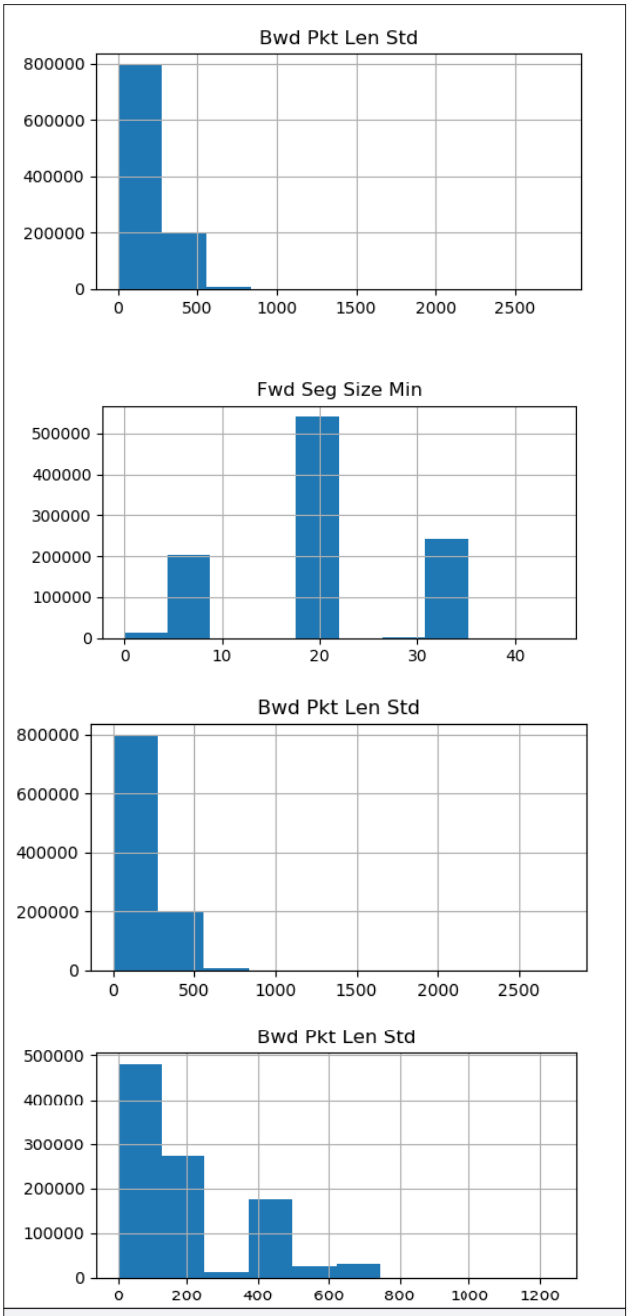

Figure 4. Graph of the best features in the second Data Set

such a way that 77 attributes with one label evaluate, if there was an attack or not, and 76 attributes which will be subjected to training within 83 attributes and in data set.

Some data had 3 million housekeeping that belongs to DDOS attack and nondestructive data in Data Set. In total, having attribute values Fifty for each quad, octet, decimal, and duodenary data sets chosen by Brute Force method, based upon the subject data set, was created. Subject attribute values were subject to train firstly by the Artificial Neural Network method created in Pyhon Pycharm program.

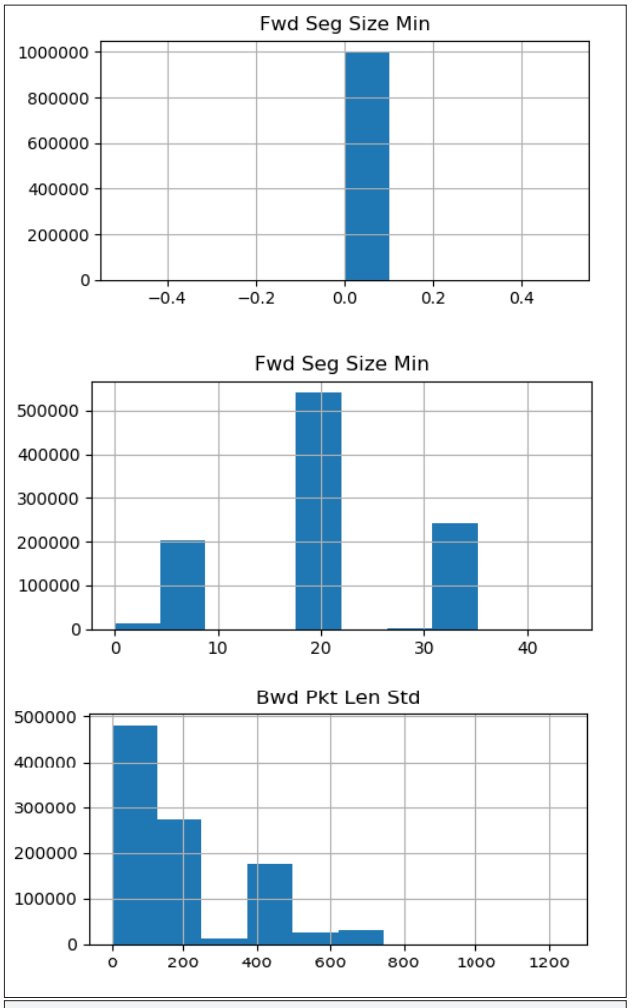

Figure 5. Graph of the best features in the third Data Set

When we check the properties and the success rate, the DDOS attack's determination at the soonest possible date with the highest success rate had carried out with the values that have four property quantities.

With creating 500 thousand data sets have each of 4,6,8,10 and duodenary property had been trained by Gaussian Naïve Bayes, Multinomial Naive Bayes, Bernoulli Naive Bayes, Logistic Regression, K-nearest neighbor (KNN), Decision Tree (entropy-gini), Random Forest and SVM algorithms. Properties that determined the DDOS attack at the highest rate and in the shortest time and the result of the train were shown in Table 2.

In the graphic in Figure 2, the algorithms and datasets with the highest success rate were included at the end of the training

It is seen that the most determinant feature in the detection of a DDOS attack is the Fwd_Pkt_Len_Std feature in Figure 1 when the data is analyzed in the data set consisting of four properties and label values, which is the first data.

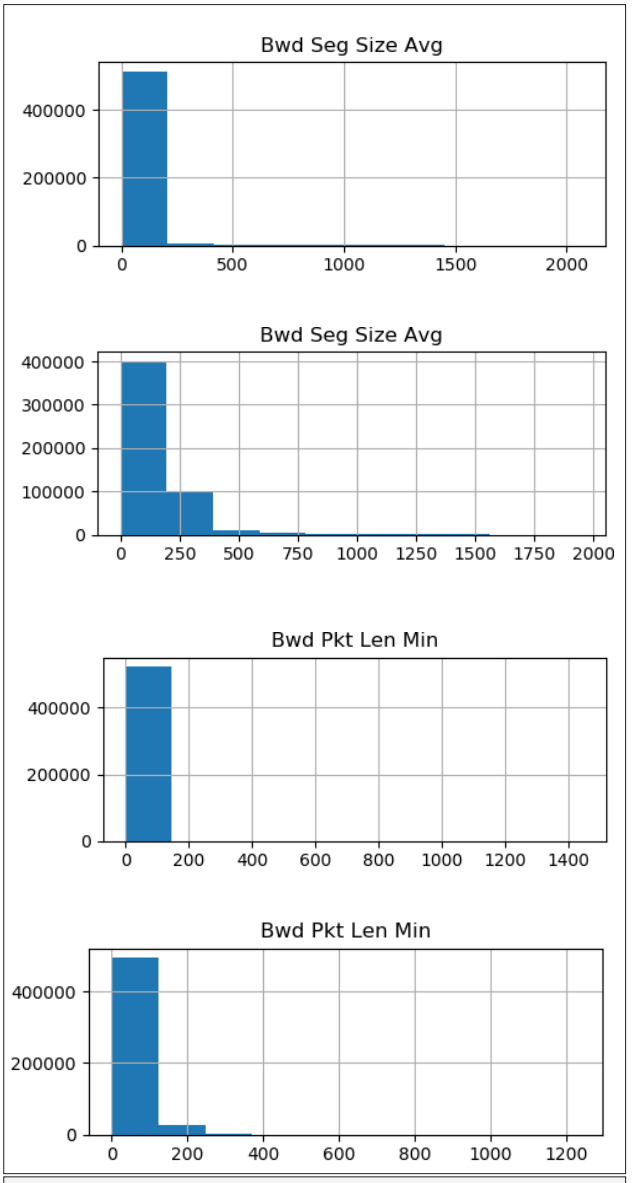

Figure 6. Chart of the best features in the fourth Data Set

It is seen that the most determinant features in the detection of DDOS attack are the Fwd Seg Size Min and Bwd Pkt Len Std features in Figure 2 when the data is analyzed.

The most determinant features in the detection of DDOS attack are Bwd Pkt Len Min and Bwd Seg Size Avg in Figure 4 when the data is analyzed.

The most determinant features in the detection of DDOS attacks are Bwd Pkt Len Min and Bwd Seg Size Min feature in Figure 5 when the data is analyzed.

Conclusion

Data sets have different 4,6,8,10 and 12 properties and were taken fifty each quantity from every category in total by using the

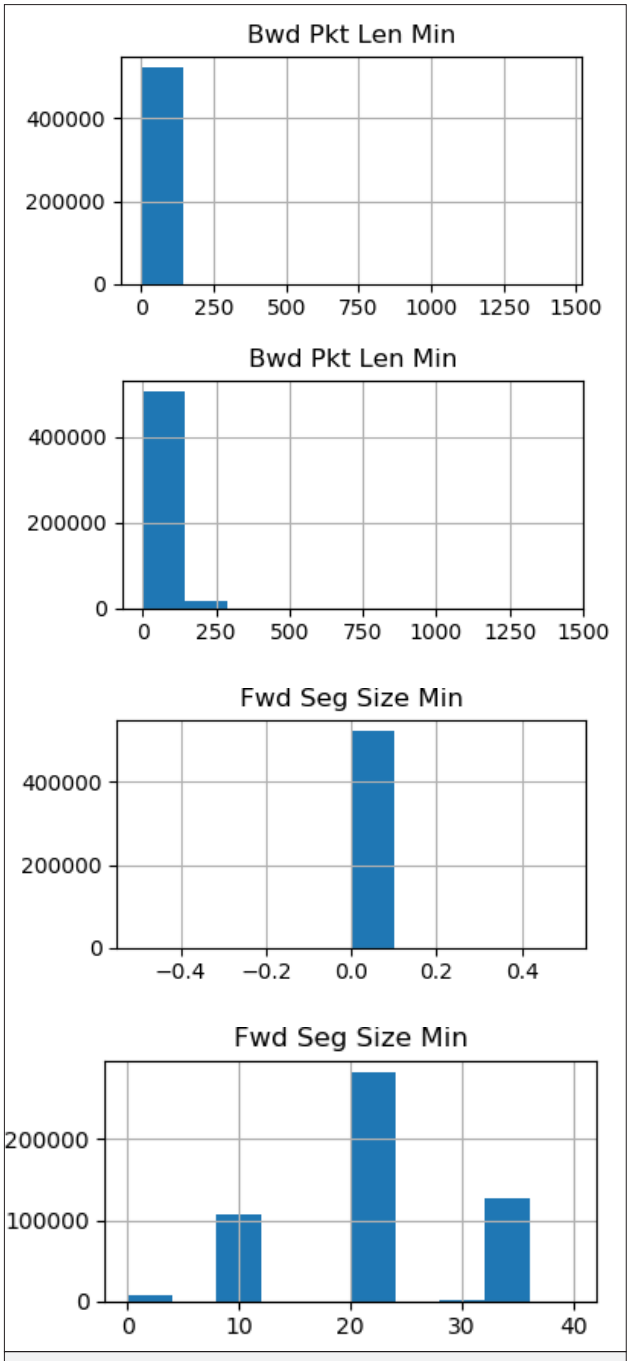

Figure 7. Graph of the best features in the fifth Data Set

Brute Force method to find the best properties in determining attack types by using the best current data set CICDDoS2019.

Every data set was trained by using YSA, Gaussian Naive Bayes, Multinomial Naive Bayes, Bernoulli Naive Bayes, Logistic Regression, K-nearest neighbor (KNN), Decision Tree (entropy-gini), Random Forest and SVM algorithms. The highest success rate of data sets was used for training and test obtained by using K-nearest neighbor, Logistic Regression, Naive Bayes, (Multinomial - Bernoulli) algorithms.

When we consider the training and test period, attack determination was done by a high rate as a 99.7% attack with the SVM algorithm, and it was identified that its performance is the best.

By analyzing five data sets from which have the highest accuracy rate in data that have been trained with different algorithms, it was seen that Fwd Pkt Len Std, Fwd Seg Size Min, Bwd Pkt Len Std, Bwd Pkt Len Min, CWE Flag Count, Bwd Seg Size Avg, Bwd Seg Size Min properties were the most distinct values for determination of DDOS attacks.

Peer-review: Externally peer-reviewed.

Conflict of Interest: The authors has no conflicts of interest to declare.

Financial Disclosure: The authors declared that the study has received no financial support.

References

1. Evans, D. (2011, 4). The Internet of Things How the Next Evolution of the Internet Is Changing Everything. CISCO.
2. Sağıroğlu, Ş., Yolaçan, E. N., & Yavanoğlu, U. (2011). Zeki SaldırıTespit Sistemi Tasarımı ve Gerçekleştirilmesi. Journal of the Faculty of Engineering & Architecture of Gazi University , 26 (2), 325-340.
3. Zulkernine, M. M. (2004). A Neural Network Based System for Intrusion Detection and Classification of Attacks. Natural Sciences and Eng Research Council of Canada (NSERC) Reports , 148-04.
4. Kraur, R., & Singh, M. (2014). Efficient Hybrid Technique for Detecting Zero-Day Polymorphic. In 2014 IEEE International Advance Computing Conference [Crossref]
5. İTUBİDB. (2013, 7). Saldırı Tespit Sistemleri. Retrieved 4 5, 2020, Available From: URL: bidb.itu.edu.tr: https://bidb.itu.edu.tr/seyir-defteri/blog/2013/09/07/sald%C4%B1r%C4%B1-tespit-sistemleri
6. Öztemel, E. (2006). Yapay Sinir Ağları. İstanbul: Papatya Yayıncılık.
7. Tanrıkulu, H., & Sazlı, M. H. (2007). Saldırı Tespit Sistemlerinde Yapay Sinir Ağlarının Kullanılması.
8. Hatipoğlu, E. (2018, 7 13). Machine Learning — Classification — Naive Bayes — Part 11. Retrieved 4 5, 2020, from medium.com: https://medium.com/@ekrem.hatipoglu/machine-learning-classification-naive-bayes-part-11-4a10cd3452b4
9. Panda, M., & Ranjan, P. M. (2007). Network Insrusipn Detection Using Naive Bayes. International Journal of Computer Science and Network Security , 7 (12), 258-263.
10. Ajagekar, S. K., & Jadhav, V. (2016, 12). Study on Web DDOS Attacks Detection Using Multinomial Classifier. International Conference on Computational Intelligence and Computing Research (ICCIC) , 1-5. [Crossref]
11. Çavuşoğlu, Ü., & Kaçar, S. (2019). Anormal Trafik Tespiti için Veri Madenciliği Algoritmalarının Performans Analizi. Akademik Platform Mühendislik ve Fen Bilimleri Dergisi , 7 (2), 205-216.
12. Özekes, S., & Karakoç, E. N. (2019). Makine Öğrenmesi Yöntemleriyle Anormal Ağ Trafiğinin Tespit Edilmesi. Düzce Üniversitesi Bilim ve Teknoloji Dergisi , 7 (1), 566-576. [Crossref]
13. Belgiu, M., & Draguţ, L. (2016). Random forest in remote sensing: A review of applications and future direction. ISPRS Journal of Photogrammetry and Remote Sensing , pp. 24-31. [Crossref]
14. Atay, R., Odabaş, D. E., & Pehlivanoğlu, M. K. (2019). İki Seviyeli Hibrit Makine Öğrenmesi Yöntemi ile Saldırı Tespiti. Gazi Mühendislik Bilimleri Dergisi , 5 (3), 258-272. [Crossref]
15. Wani, A. R., Rana, Q. P., Saxena, U., & Pandey, N. (2019). Analysis and Detection of DDoS Attacks on Cloud Computing Environment using

Machine Learning Techniques. 2019 Amity International Conference on Artificial Intelligence (AICAI) pp. 870-875. Dubai: IEEE. [Crossref]

16. CyberSecurity, C. I. (2019). unb.ca/cic/datasets/ddos-2019. Available From: URL: unb.ca: https://www.unb.ca/cic/datasets/ddos-2019.html

17. Sharafaldin, I., Lashkari, A. H., Hakak, S., & Ghorbani, A. A. (2019). Developing Realistic Distributed Denial of Service (DDoS) Attack Dataset and Taxonomy. 2019 International Carnahan Conference on Security Technology (ICCST) pp. 1-8. Chennai: IEEE. [Crossref]

Tuğba Aytaç received her Bachelor's Degree in 2016 from Süleyman Demirel University Department of Electronic and Communication Engineering.She studies for master degree in cyber security from İstanbul Commerce University Institute of Science.She works in the private sector as IT auditor

Muhammed Ali Aydın received his B.S degree in 2001, M.S degree in Computer Engineering in 2005 from Istanbul Technical University and he holds a doctorate in the same discipline from Istanbul University, received in 2009. Dr. Aydın is currently working as an Assistant Professor in Computer Engineering Department of Istanbul University - Cerrahpasa. His main research interests involve computer networks, cryptography and cyber security.

Abdül Halim Zaim Computer Engineering at the Faculty of Engineering at Istanbul University and the Director of the Center for Information Technology Application and Research at Istanbul Commerce University. Abdul Halim Zaim served as Vice Rector and Vice President of Academic Evaluation Commission at Istanbul Commerce University. He received his MS degree in Computer Engineering from Bogazici University in 1996 and his PhD in Electrical and Computer Engineering from North Carolina State University (NCSU) in 2001.