NATURENGS
E-ISSN 2717-8013
https://dergipark.org.tr/en/pub/naturengs

Classification of Distributed Denial of Service Attacks Using Machine Learning Methods

Uğur İnce [a,*], Gülşah Karaduman [a]

[a] Elazig Firat University, Department of Computer Engineering Elazığ Türkiye – 23100

*Corresponding author

ARTICLE INFO

Received 12.03.2024
Accepted 01.05.2024

Doi: 10.46572/naturengs.1450965

ABSTRACT

With the digitalized world, the uninterrupted provision of services over the internet, especially in hospitals, banking, energy, etc. systems is of great importance. There are many attack methods to disrupt or disable these services. Denial of service attacks, which are one of these methods, are more complex and difficult to detect; Organizing such attacks becomes very easy and cost-effective thanks to many tools. Attackers can perform DDoS attacks on target systems with very little knowledge and skills, and they can render target systems inoperable, sometimes for a short time or for days. In this work, K Nearest Neighbor, decision trees, and support vector machines were used to classify the most recent and extensive assault dataset available online, CIC-DDoS2019. With recall and accuracy values of 0.95 and 0.94, respectively, the decision trees approach performed best in the classification; the support machine vectors method performed best with an f-1 score and precision value of 0.95. Given the parameters used and the machine learning techniques employed, this work will be a helpful resource for academics who are thinking about utilizing these techniques to categorize DDoS attacks.

Keywords: Denial of service attacks, distributed denial of service attacks, machine learning, classification

1. Introduction

With the advancement of broadband network technology, the internet is becoming an indispensable part of our lives. In addition, the concept of "smart" technology appears in every aspect of our daily lives. In early 2022, there will be 4.95 billion internet users worldwide, accounting for 62.5 percent of the world's population, according to a forecast released by We Are Social and Hootsuite [1]. TurkStat data shows that in Turkey, the percentage of people in the 16–74 age range who used the internet in 2021 and 2022 was 82.6% and 85.0%, respectively. In 2022, it was found that 80.9% of women and 89.1% of males used the Internet [2].

In addition to the opportunities they offer to make life easier, information technologies have also led to the development of new security concerns. In the new world, criminal acts such as theft and fraud have become possible without the need for physical contact or being in the same place as the victim [3]. Borders between countries are disappearing in the virtual world. The use of technology and the dependence of countries on technology in social, economic and military fields are increasing. The widespread use of technology in all fields brings both risks and benefits. With the development of communication technology, the concept of attack is also changing. Today, it is observed that attacks on information and communication infrastructures are increasing in order to damage the sectors where technology is widely used and critical infrastructures [4].

Attacks known as denial of service (DoS) aim to interfere with the target system's ability to function.

Conversely, distributed denial of service (DDoS) attacks seek to render the target system unusable as soon as possible for a large number of compromised network devices [5].

In DDoS attacks, the attacker aims to prevent the target system from responding to this traffic by creating continuous traffic in the target system using tools that collectively control many devices, called zombies/bots, that the attacker has captured in various ways. Since the target system does not know which of the incoming traffic is its real user and which is traffic generated by the attacker, it tries to respond to all of them and its resources are exhausted in a short time and it becomes unable to provide service.

Machine learning is used in many fields such as banking and finance, transportation, retail, healthcare, agriculture, customer service, etc. for purposes such as

* Corresponding author. e-mail address: ugurince@munzur.edu.tr
ORCID: 0000-0001-5265-4661

finding cause and effect relationships between variables, identifying unusual occurrences, and classification. All businesses, no matter how big or little, need to be able to recognize DDoS attacks from regular traffic and respond quickly to them. Using machine learning methods to detect DDoS attacks will bring great benefits to these organizations in terms of both manpower and cost.

2. Background

In this section, literature review, machine learning models, DDoS attacks, and data collection of the study were discussed.

2.1. Literature Review

Doshi et al. carried out a study on the internet of things (IoT) in 2018. Using a data set they generated on the smart home system, the researchers used K-nearest neighbors (KNN), Support vector machines with linear kernels (LSVM), Decision trees (DT), Random Forest (RF), and neural networks (NN) to classify DDoS attacks [6].

In 2020, Shanmuga et al. generated traffic in the virtual machines they built and used the data set from this traffic to identify DDoS assaults using Naive Bayesian, K-Mean, and Random Forest techniques [7].

Özçam employed Decision Tree (DT), Random Forest (RF), eXtreme Gradient Boosting (XGBoost), K-Means Clustering, and Isolation Forest techniques to identify TCP-SYN Flood and UDP Flood attacks on the BOUND DDoS data set in 2021 [8].

In 2022, Maniula and colleagues used the Apache Park Streaming tool to pre-process the data on the dataset they had created. They then classified DDoS assaults using techniques such as Random Forest (RF), K-Nearest Neighbor (KNN), and Naive Bayesian (NB) [9].

Maheswari et al. employed hybrid metaheuristic algorithms on the CAIDA-2007 and CIC-DDoS2019 dataset in 2022 [10].

The CIC-DDoS 2019 dataset was used by Akgun et al. in 2022 to evaluate a variety of deep learning models, including CNN, LSTM, and DNN, for varying units per layer [11].

2.2. Distributive Denial Of Service Attacks

Concurrent with the swift advancement of intelligent technology, there is a growing quantity of interconnected gadgets. For this reason, the disruption or interruptionof network-based services causes serious victimization to the person or organization during the realization of business and transactions. DDoS attacks are also seen as a serious threat to these systems and services. The competencies of DDoS attackers are increasing and attackers specify new targets for themselves every day. In this direction; analyzing DDoS attacks from past to present provides a general perspective on attacks.

2.2.1. Causes of DDoS Attacks

Looking at the reasons for DDoS attacks, it is observed that the desire to attack develops in five basic stages. These are: for demonstration or research purposes, for amateur hacking, for economic gain, as a means of social action, and for cyber warfare [12].

Today, DDoS attacks are mostly carried out for hobby purposes, personal ambitions, financial gains, and ideological approaches. Furthermore, the majority of these attacks nowadays revolve around three main goals: political objectives, financial gain based on winning or losing money, and the need to conceal the attack in order to mask the primary target: information theft. Figure 1 shows the reasons for attacks today in detail.

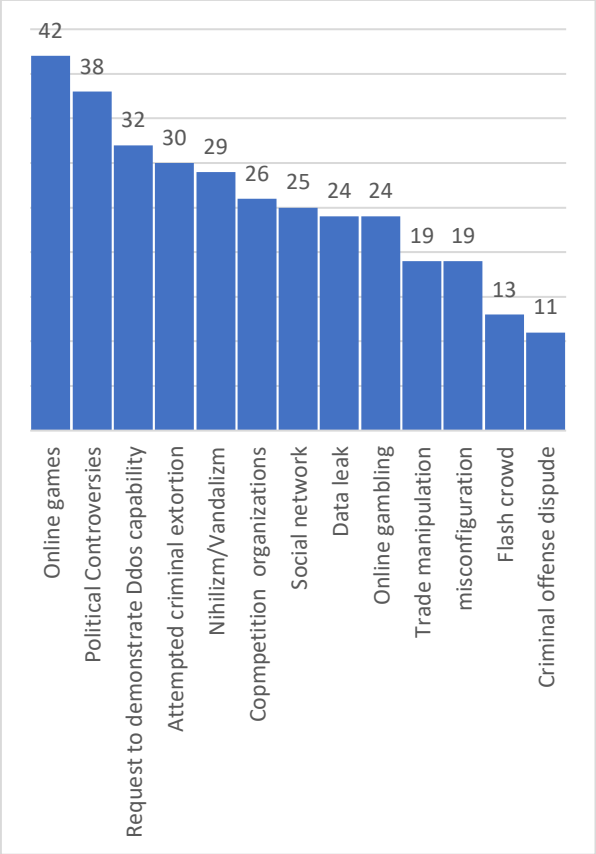

Figure 1. Causes of DDoS Attacks Today

2.2.2. Dimensions of DDoS Attacks

According to Figure 2, 78% of DDoS attacks in 2022 focused on the OSI model's application layer, 17% on the network and transport layers, and 3% on the DNS. Eighty percent of assaults in the third quarter of 2021 targeted the application and transport layers with packet flooding. However, while the cost of botnets has decreased this year, their impact has increased, leading to a shift towards application layer attacks.

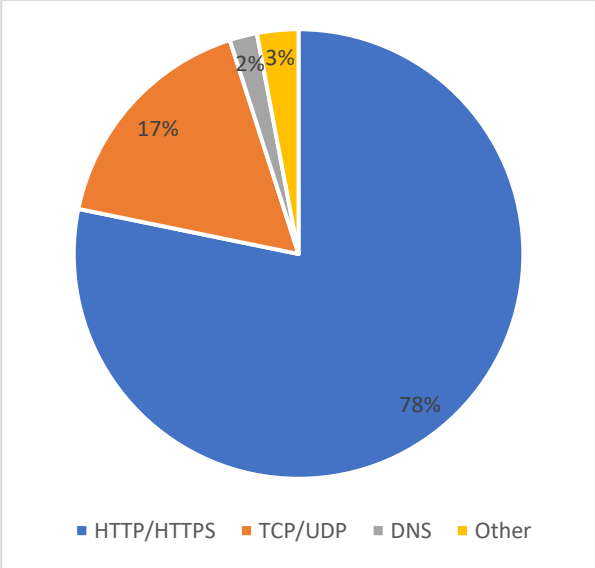

Figure 2. Protocol Based Attack Rates

When we look at DDoS attacks in sector-based services compared to the previous year, in 2022, it has reached 12 times the rate in financial services, 4 times in telecommunications, 1.5 times in the retail sector, 3 times in the entertainment sector, and 5 times in the insurance sector [13]. Once more, when it comes to sector-based assaults, the financial and telecommunications sectors led with 34% and 26%, respectively, while the education sector, which had not previously been the subject of many attacks, saw 2% of them, particularly with the introduction of the remote learning system following the Covid pandemic.

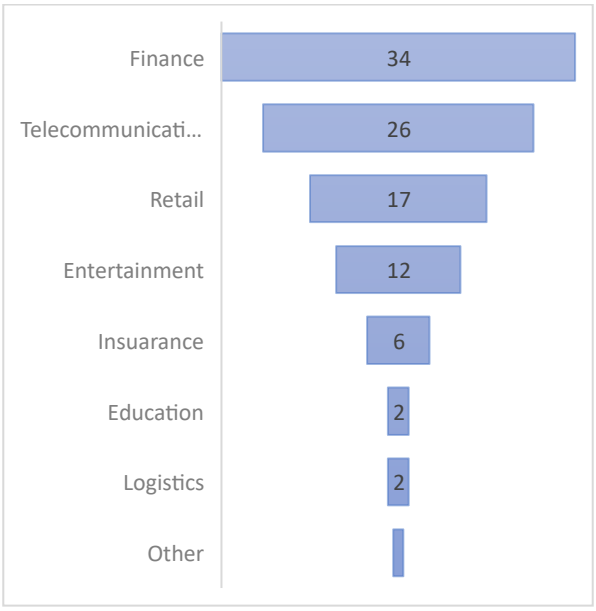

Figure 3. Sector Based DDoS Attacks

2.2.3. DDoS Attack Types

Due to the distributed nature of DDoS attacks, it is very difficult to distinguish between attack traffic and real traffic [14]. Nowadays, DDoS attacks are considered as the most powerful weapon preferred by the attacker to prevent the availability of Internet services. Attackers generally use three types of DDoS attack methods (Network, Protocol and Application).

Volmetric attacks, sometimes known as network attacks, are the most prevalent kind of attacks that aim to overload the target system's bandwidth. TCP/UDP flood, DNS/NTP/Mamcached amplification attacks are examples of these attacks.

Protocol attacks are attacks against targets that serve a large number of people by sending fake port-based requests. These attacks target network and transport layer protocols of the OSI model. SYN/SYN-ACK/ACK flood attacks are examples of these attacks.

At the application layer of the OSI model, Web applications are the target of application assaults. Attacks against HTTP, HTTPS, and SMTP services are examples of application attacks.

The attacks mentioned above can be performed by attackers one by one, or more than one attack type (multi-vector) can be performed simultaneously. In this way, attackers aim to make their attacks more complex and make it difficult for both incident response teams and security devices to prevent them [15].

2.3. MACHINE LEARNING

Through the use of machine learning, a computer can learn from given data without needing to be explicitly and thoroughly programmed for every problem [16]. Within artificial intelligence (AI), it is regarded as a subset. To find patterns in data, machine learning employs algorithms. Utilizing these patterns, a predictive data model is constructed. Machine learning outcomes improve with more data and experience, much like human performance does with further practice. Machine learning is an excellent alternative when coding a solution is impractical or when data, demands, or tasks are continually changing because of its versatility.

2.3.1. Support Vector Machines

Support Vector Machines (SVM), often used for linear data, are also used for nonlinear data with the help of a kernel function [17]. Encouragement One supervised learning technique that is frequently applied to classification issues is the vector machine. It creates a line to divide up points on a plane. It seeks to place this line at the farthest position for each class's points. Because SVMs reduce structural risk rather than the square or absolute size of the mistake, they are generally resistant to overfitting [18].

2.3.2. Decision Trees

Decision tree is a machine learning method for approximating objective functions where the learning function is represented by a decision tree [19]. Decision trees are structures that can perform classification and regression using a hierarchical structure of root, nodes and leaves. It is used in complex data sets.

2.3.3. K Nearest Neighbor (KNN)

Using the data stored in the training set, K-nearest neighbor (KNN), a supervised machine learning technique, classifies objects based on two fundamental values: neighborhood and distance [20]. In the KNN algorithm, it aims to determine the class to be formed based on which class the nearest neighbor of the variable to be predicted is from intensively.

2.4. Data Set

We made use of The Canadian Institute for Cybersecurity's (CIC) CIC-DDoS2019 database, which is updated with safe, typical DDoS attacks that mimic real-world data.

The safest and most recent DDoS assaults that mimic real-world data are included in CICDDoS2019 (PCAPs). It also contains the findings of a CICFlowMeter-V3 network traffic study, with flows labeled according to protocols, attacks, source and destination IP addresses, and timestamps (CSV files). Based on email, file transfer, remote access, and internet protocols, we created the abstract behavior of 25 individuals for this dataset.

Every day, the dataset is arranged. Raw data was recorded for each day, including the amount of network traffic for each system (pcaps) and the event logs from Ubuntu and Windows. CICFlowMeter-V3 was used to extract features from the raw data; on each system, more than 80 traffic features were retrieved and saved as a CSV file.

The dataset is organized on a daily basis. For each day, raw data including network traffic per machine (pcaps) and event logs (Windows and Ubuntu event logs) were recorded. For feature extraction from the raw data, CICFlowMeter-V3 was used and more than 80 traffic features per machine were extracted and saved as a CSV file [21].

3. Methodology

The study was conducted on a Dell desktop computer with Intel Core(TM) i5-6500 CPU @ 3.20GHz, 16GB RAM and Windows 10 Pro operating system. The Pyton programming language and out-of-the-box machine learning libraries were used to build the models. Pyton was executed in the Spyder editor.

The dataset was numbered with ID tags according to the data. In the data set, 30% of the data was randomly selected for training. Support vector machine, decision tree and nearest neighbor machine learning models were applied to the data and accuracy rates, recall, f1-score and precision values were compared.

3.1. Data Preprocessing

In the study, 5 classes consisting of DDoS attack types (LDAP, NETBIOS, PORTMAP, SYN, UDP) and 1 class consisting of normal traffic (BENIGN) were used together with 6 classes. In order to ensure a balanced distribution of the classes, the total number of records was reduced to 12000 based on the class with the least data. Additionally, the features were whittled down from 87 to 35. Table 1 lists the chosen characteristics along with their categories.

Before the data were given to the models for training purposes, the data were normalized without distorting the difference in value ranges since there were data with different interval values in the data set. It can be said that this process will make it easier for the models to learn the weights. For this, the Min Max Scaler method in the sklearn library was used and the data was scaled in the range [0,1].

Table 1. *Selected features and their types*

Feature Name	Feature Type
ID	int64
Source IP	object
Source Port	int64
Destination IP	object
Destination Port	int64
Protocol	int64
Flow Duration	int64
Total Forward Packets	int64
Total Backward Packets	int64
Total Length of Forward Packets	float64
Total Length of Backward Packets	float64
Flow Bytes/s	float64
Max Packet Length	float64
SYN Flag Count	int64
ACK Flag Count	int64
Down/Up Ratio	float64
Average Packet Size	float64
Average Forward Segment Size	float64
Average Backward Segment Size	float64
Forward Header Length.1	int64
Subflow Forward Packets	int64
Subflow Forward Bytes	int64
Subflow Backward Packets	int64
Subflow Backward Bytes	int64
Flow Packets/s	float64
Forward Header Length	int64
Backward Header Length	int64
Forward Packets/s	float64
Backward Packets/s	float64
Min Packet Length	float64
Init_Win_bytes_forward	int64
Init_Win_bytes_backward	int64
act_data_pkt_fwd	int64
min_seg_size_forward	int64
Label	object

4. Results and Discussion

In DDoS attacks, the attacker uses tools that collectively control numerous devices, known as zombies or bots, that the attacker has seized in various ways in order to create continuous traffic in the target system and prevent the target system from reacting to this traffic. The target system tries to reply to all incoming traffic since it is unsure which traffic is from its actual users and which is from the attacker. As a result, its resources are quickly depleted, making it impossible for it to continue providing service.

All companies, regardless of size, must be able to distinguish DDoS attacks from normal traffic and act swiftly in response. The benefits of using machine learning techniques to identify DDoS attacks are substantial.

One of the most recent and extensive datasets on the Internet, CICDDoS2019, was used for classification in this work utilizing DT, SVM, and KNN methods. The precision, f-1 score, recall, and accuracy values obtained using the KNN approach were 0.88, 0.90, 0.92, and 0.92, respectively. The Decision Tree approach yielded precision, f-1 score, recall, and accuracy scores of 0.91, 0.93, 0.95, and 0.94, respectively. Ultimately, precision, f-1 score, recall, and accuracy values using the Support Vectors approach were 0.95, 0.95, 0.91, and 0.91, respectively. Considering the results obtained, the decision tree classification method achieved a better accuracy rate than the other methods. Support vector machines produced higher outcomes in terms of f1-score and accuracy values, despite the decision tree being the most effective approach in terms of recall value. The results obtained are shown in Figure 4.

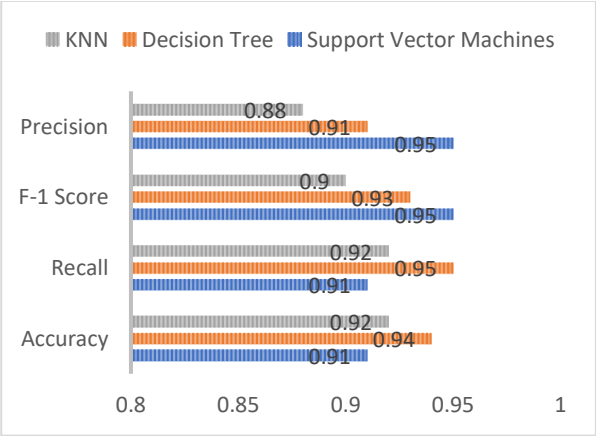

Figure 4. *Results of classification*

5. Conclusion

The destructive impact of DDoS attacks has increased since their inception. While in the early years the attacks were carried out for demonstration/research purposes, the reasons for the attacks have changed. In parallel with the reasons for DDoS attacks, there has been a significant increase in the size of the attacks. While in the early days the size of attacks was in the megabyte range, today they are in the terabyte range.

To minimize the impact of DDoS attacks, it is important to develop mechanisms that can distinguish malicious traffic from normal traffic. In this study, machine learning models are tested to distinguish DDoS attack traffic from normal traffic. One of the most current and extensive datasets available online, CICDDoS2019, underwent preprocessing to provide relevant results, and machine learning models were trained using this data. Support vector machines, decision trees and nearest neighbor models were applied to the dataset and successful classification was achieved.

In light of the information obtained, it is intended to take the study one step further by ensuring that both the dataset and the detection system are made using the artificial intelligence kit in future studies.

References

[1] https://recrodigital.com/dunyada-ve-turkiyede-internet-sosyal-medya-kullanimi-2022, Access May 2023

[2] https://data.tuik.gov.tr/Bulten/Index?p=Hanehalki-Bilisim-Teknolojileri-(BT)-Kullanim-Arastirmasi-2022-45587, Access April 2023

[3] **Hekim, H.,** BAŞIBÜYÜK, O. (2013). Siber Suçlar ve Türkiye'nin Siber Güvenlik Politikalari. Uluslararası Güvenlik ve Terörizm Dergisi, 4(2), 135-158.

[4] **Atasever, S.,** Özçelik, İ., Sağıroğlu, Ş. (2019). Siber Terör ve DDoS, Süleyman Demirel Üniversitesi Fen Bilimleri Dergisi, Cilt 23, Sayı 1, 238-244.

[5] **Masum, E.,** Samet, R. (2018). Mobil Botnet ile Ddos Saldırısı, Bilişim Teknolojileri Dergisi, Cilt:11, Sayı:2, Nisan

[6] **Doshi, R.,** Apthorpe, N. ve Feamster, N. (2018). Machine learning ddos detection for consumer internet of things devices, 2018 IEEE Symposium on Security and Privacy Workshops, 29-35

[7] **Shanmuga, S.,** Sivaram, M. ve Jayanthiladevi, A. (2020). Machine learning based DDOS detection, 2020 International Conference on Emerging Smart Computing and Informatics (ESCI), AISSMS Institute of Information Technology, Pune, India. Mar 12-14, 2020, 29-35.

[8] **Özçam, B.** (2021). DDoS Atak Tespiti İçin Makine Öğrenmesi Algoritmaları ile Anomaly Tespiti, Yüksek Lisans Tezi, İstanbul Ticaret Üniversitesi, Fen Bilimleri Enstitüsü.

[9] **Manjula, H.T.,** Mangla, N. (2022). An approach to on-stream DDoS blitz detection using machine Learning Algorithms, Materials Today: Proceedings, 1-8

[10] **Aastha Maheshwari,** Burhan Mehraj, Mohd Shaad Khan, Mohd Shaheem Idrisi, An optimized weighted voting based ensemble model for DDoS attack detection and mitigation in SDN environment,Volume 89, 2022,104412, ISSN 0141-9331, https://doi.org/10.1016/j.micpro.2021.104412.

[11] **Akgun, D.,** Hizal, S., Cavusoglu, U.(2022). A new DDoS attacks intrusion detection model based on deep learning for cybersecurity,Computers & Security,Volume 118,ISSN 0167-4048,2022, https://doi.org/10.1016/j.cose.2022.102748.

[12] **Atasever, S.,** Özçelik, İ., Sağıroğlu, Ş. (2019). Siber Terör ve DDoS, Süleyman Demirel Üniversitesi Fen Bilimleri Dergisi, Cilt 23, Sayı 1, 238-244.

[13] https://www.infosecurity-magazine.com/blogs/2022-ddos-yearinreview/, Acces March 2023

[14] **Asarkaya, S.,** Kaynar, O.,Yelmen, İ., Yıldırım, F.,Zontul, M. (2021). Tasarım Mimarlık ve Mühendislik Dergisi, Cilt 1, Sayı 3, 2021, 221 – 232.

[15] https://www.barikat.com.tr/images/blog/loddos-ddos-saldirilari-degerlendirme-raporu.pdf, Access May 2023

[16] **Meng, T.,** Jing, X., Yan, Z., Pedrycz, W. (2020). A survey on machine learning for data fusion, İnformation Fusion, Cilt: 57, Mayıs 2020, 115-129

[17] **Kaynar,** O., Arslan, H., Görmez, Y., Işık, Y.E. (2018). Makine Öğrenmesi ve Öznitelik Seçim Yöntemleriyle Saldırı Tespiti, Bilişim Teknolojileri Dergisi, Cilt:11, Sayı:2, Nisan2018, c doi: 10.17671/gazibtd.368583

[18] **Aastha Maheshwari,** Burhan Mehraj, Mohd Shaad Khan, Mohd Shaheem Idrisi, An optimized weighted voting based ensemble model for DDoS attack detection and mitigation in SDN environment, Volume 89, 2022, 104412,ISSN 0141-9331, https://doi.org/10.1016/j.micpro.2021.104412.

[19] **Gökay Emel, G.,** Taşkın, Ç. (2005). Veri Madenciliğinde Karar Ağaçları ve Bir Satış Analiz Uygulaması, Eskişehir Osmangazi Üniversitesi Sosyal Bilimler Dergisi, Cilt:6, Sayı:2, Aralık 2005, 222-239

[20] **Masum, E.,** Samet, R. (2018). Mobil Botnet ile Ddos Saldırısı, Bilişim Teknolojileri Dergisi, Cilt:11, Sayı:2, Nisan

[21] **Iman Sharafaldin,** Arash Habibi Lashkari, Saqib Hakak, and Ali A. Ghorbani, "Developing Realistic Distributed Denial of Service (DDoS) Attack Dataset and Taxonomy", IEEE 53rd International Carnahan Conference on Security Technology, Chennai, India, 2019.