Web Attacks Analytics

Luis O. Sanchez Pardella
Master in Computer Engineering
Alfredo Cruz, PhD
Director for Computer Engineering
Polytechnic University of Puerto Rico

Abstract — *Cyber-attacks are on the rise. Information systems are constantly being attacked, generating economic loss. Computer networks and web technologies are the preferred places for hackers to commit cybercrimes. Most computers and devices are connected to the Internet through a communication network. On many occasions, network security administrators do not monitor adequately the cyber-attacks that are occurring through the computer network, making it difficult to protect the organizations' networks. It is important for information security experts to learn how to use data mining tools to analyze cyber-attacks in the network and to make better decisions regarding security. Based on this need, this project has been developed using RStudio and R programming Language. Four datasets were selected, and web attack information was extracted and analyzed from these datasets. With this web attack information available, network security administrators can analyze the data and make decisions that contribute to enhance the protection of networks.*

Key Terms — *Analytics, NB15, RStudio, Web Attacks*

INTRODUCTION

Cyber-attacks are on the rise. Information systems are constantly being attacked, generating great economic loss in all sectors of the economy. Most computers and devices are connected to the Internet through a communication network. Computer networks and web technologies are the preferred places for hackers to commit cybercrimes.

The proportion of cyber-attacks is increasing daily and new attacks are emerging exponentially, making it difficult for security experts to maintain a safe and secure environment. Many of these experts do not have the tools and knowledge to be able to extract important information from network traffic and recognize the trends of cyber-attacks on the network. This makes it difficult for security experts to protect information systems using calculated methods of analysis and decision making [1].

Using programming tools and coding, these experts can provide better monitoring, extraction of important data, and the protection of their computer networks from new threats. Based on this need, this Master Project has been developed, so that network administrators can use various programming algorithms to extract important traffic data from the network. This data is analyzed, and decisions can be made that contribute to the enhancement and protection of the network [2].

In this project, Security Administrators are provided with step by step data analytics applied to public datasets of network traffic and web data attacks [3]. The software tool and programming language used is RStudio, and R, respectively [2]. RStudio is an Integrated Development Environment for the R programming language.

With the use of RStudio and R language, IT security administrators can learn to extract important data from web attacks and prepare the data for worst case scenarios. These tools help security administrators develop effective and efficient countermeasures against attacks on the web and networks, to mitigate the risk of cyber-attacks on the organization.

UNSW-NB15 DATASET

These datasets were created by the IXIA PerfectStorm tool (see Figure 1) in the Cyber Range Lab of the Australian Centre for Cyber Security (ACCS) to generate a hybrid of real modern normal activities and synthetic contemporary attack behaviors [4].

Tcpdump tool is utilized to capture 100 Gb of the raw traffic (e.g. Pcap files). This dataset has nine types of attacks namely: Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode, and Worms [3]. The Argus, Bro-IDS tools are used, and twelve algorithms are developed to generate a total of 49 features with the class label [4].

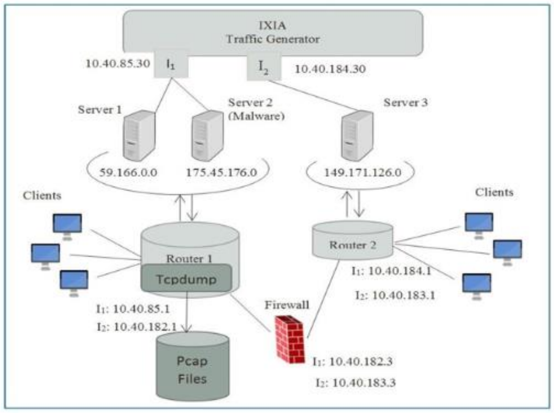

Figure 1
IXIA Traffic Generator

Description of the UNSW-NB15 Dataset

There are nine web attack types identified in the UNSW-NB15 Dataset [5]:

1. Fuzzers: an attack in which the hacker tries to find security loopholes in the operating system, program or network and make these resources suspended for some time and can even crash them.
2. Analysis: a type intrusion that get unauthorized access to web applications through port scanning, malicious web scripting, and dispatching spam, emails etc.
3. Backdoor: a technique in which an intruder can bypass the usual authentication and can get unauthorized remote access to a system.
4. DoS: an intrusion in which the hacker tries to disrupt the computing resources, by making them extremely busy to prevent the authorized access to the resources.
5. Exploit: the intrusions which utilize the software vulnerabilities, error, or glitch within the operating systems (OS) or software.
6. Generic: This attack acts against a cryptographic system and it tries to break the key of the security system.
7. Reconnaissance: Attacks begin with a scan of the network from the infected endpoint to locate the asset and services a hacker wants to target. Diversities of reconnaissance include active, random IP as well as stealth scanning.
8. Shellcode: a malicious software attack in which the hacker penetrates a slight piece of code starting from a shell to control the compromised machine.
9. Worm: a malicious software that replicate themselves and spread to other computers by using the network to spread the attack, depending on the security failures on the target computer which it wants to access.

Methodology & Design

This project will demonstrate how an adequate data analytics process can be applied to web attacks. Four datasets with network traffic information will be extracted and data analytics will be applied to the identified web attack data. Once the RStudio tool and the R programming language is used to upload the datasets to RStudio, the content of these datasets will be observed to choose which data set(s) to work with, for a subsequent data analytics process. Each data set has total of 49 columns with different data types [4].

Analysis of the Structures of Datasets

The analysis of the structure of the dataset must be done. Number of columns and rows is observed, and the different types of data it stores. The relationship of data with other data within the dataset is also observed.

Table 1 Head () function shows the first six rows of the data1 dataset. V48 column attacks are in blank because no attacks were detected in the first 6 rows of the dataset. NB15_1 dataset is stored in the data1 variable. The same procedure was applied to datasets data2, data3 and data4.

Consequently, the names are assigned again to the datasets that have been moved and stored in new data structures during the data transformation process (NB15_1 -> NB15_4).

Table 1
head(data1)

> head(data1)

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15
1	ï»¿59.166.0.0	1390	149.171.126.6	53	udp	CON	0.001055	132	164	31	29	0	0	dns	500473.94
2	59.166.0.0	33661	149.171.126.9	1024	udp	CON	0.036133	528	304	31	29	0	0	-	87676.09
3	59.166.0.6	1464	149.171.126.7	53	udp	CON	0.001119	146	178	31	29	0	0	dns	521894.53
4	59.166.0.5	3593	149.171.126.5	53	udp	CON	0.001209	132	164	31	29	0	0	dns	436724.56
5	59.166.0.3	49664	149.171.126.0	53	udp	CON	0.001169	146	178	31	29	0	0	dns	499572.25
6	59.166.0.0	32119	149.171.126.9	111	udp	CON	0.078339	568	312	31	29	0	0	-	43503.23

	V16	V17	V18	V19	V20	V21	V22	V23	V24	V25	V26	V27	V28	V29	V30	V31
1	621800.94	2	2	0	0	0	0	66	82	0	0	0.00000	0.00000	1421927414	1421927414	0.017
2	50480.17	4	4	0	0	0	0	132	76	0	0	9.89101	10.68273	1421927414	1421927414	7.005
3	636282.38	2	2	0	0	0	0	73	89	0	0	0.00000	0.00000	1421927414	1421927414	0.017
4	542597.19	2	2	0	0	0	0	66	82	0	0	0.00000	0.00000	1421927414	1421927414	0.043
5	609067.56	2	2	0	0	0	0	73	89	0	0	0.00000	0.00000	1421927414	1421927414	0.005
6	23896.14	4	4	0	0	0	0	142	78	0	0	29.68222	34.37034	1421927414	1421927414	21.003

	V32	V33	V34	V35	V36	V37	V38	V39	V40	V41	V42	V43	V44	V45	V46	V47	V48	V49
1	0.013000	0	0	0	0	0	0	0	0	3	7	1	3	1	1	1		0
2	7.564333	0	0	0	0	0	0	0	0	2	4	2	3	1	1	2		0
3	0.013000	0	0	0	0	0	0	0	0	12	8	1	2	2	1	1		0
4	0.014000	0	0	0	0	0	0	0	0	6	9	1	1	1	1	1		0
5	0.003000	0	0	0	0	0	0	0	0	7	9	1	1	1	1	1		0
6	24.315000	0	0	0	0	0	0	0	0	2	4	2	3	1	1	2		0

The dataset NB15_1 has 700,001 rows and 49 columns; NB15_2 has 700,001 rows and 49 columns; NB15_3 has 700,001 rows and 49 columns; and NB15_4 has 400,044 rows and 49 columns.

The data types of the columns of the 4 datasets are made up of integers, numbers, and factors. The domain of the cyber-attack data column is made up of 10 factors these are " ", "Fuzzers", "Analysis", "Backdoors", "DoS", "Exploits" "Generic", "Reconnaissance", "Shellcode", "Worms" [6][7].

Results And Discussion

The results of the data analytics process are discussed step by step. It shows how the data is cleaned and prepared to convert it into useful information [3]. It also shows how data is moved and stored through different data structures for the application of different data analytics algorithms. Once the results are obtained, they can be analyzed, and data analytics algorithms can be applied for in-depth statistical analysis [8].

Before extracting data from datasets, data cleaning algorithms are applied to make the data usable for extraction. Data cleaning is the process of transforming raw data into usable data. Cleaning data, checking quality, and standardizing data types, are part of the steps taken to analyze the data [3].

This project uses nine factors that represent the different types of web attacks. These datasets are moved and stored in different data structures to develop an efficient data analytics process [6].

Different programming algorithms are applied to manipulate and manage the data [6]. The data structures used in the project are vectors and data frames. After moving and storing the data in different data structures, different graphs and tables are generated to show useful information on the web attacks [6]. The graphic representation of this data helps security experts analyze web attacks to improve the protection of web systems [9].

Table 2 shows column V48 from the data1 dataset (just a part of the dataset rows). This column contains the different attacks detected and stored in the data1 dataset.

Table 2
Data1 column V48

> data1$V48

[1]						
[7]						
[13]						
[19]			Exploits	Exploits	Reconnaissance	
[25]						
[31]						
[37]				Exploits	Exploits	
[43]						
[49]						
[55]				DoS	Generic	
[61]						
[67]						
[73]						Exploits
[79]	DoS	Exploits	Exploits			
[85]						
[91]						Exploits
[97]						
[103]						
[109]		Exploits	Reconnaissance	Exploits		
[115]						
[121]						
[127]					Exploits	Exploits
[133]						
[139]						
[145]	Reconnaissance	Exploits	Exploits			
[151]						
[157]						
[163]				Exploits		
[169]						
[175]						
[181]	Exploits					
[187]					Reconnaissance	
[193]						Reconnaissance
[199]	DoS					Exploits
[205]						

Table 2 stores the data from the rows of dataset data1 consecutively. For example, the table consists of 6 columns and if an attack is detected in the first column it is stored in column 1; if in row 2 an attack is detected, it is stored in column 2, and so on.

If it does not detect an attack on any row, the column assigned to the row is left blank. The web attacks that are shown on the table are: DoS, Exploits, Generic, Reconnaissance. Blank spaces mean not attacks were detected; row 1 to row 205. This same process was applied to the other datasets.

The data of the different types of attacks that is stored in column V48 of the data1 dataset will move to a new data frame (data1Mod), registering only the types of registered attacks, thus eliminating the blank rows where no web attacks were registered (see Table 3). This same process was applied to the other datasets.

Table 3
data1Mod column V48

```
> data1Mod$V48
  [1] Exploits       Exploits       Reconnaissance Exploits       Exploits       DoS
  [7] Generic        Exploits       DoS            Exploits       Exploits       Exploits
 [13] Exploits       Reconnaissance Exploits       Exploits       Exploits       Reconnaissance
 [19] Exploits       Exploits       Exploits       Exploits       Reconnaissance Reconnaissance
 [25] DoS            Exploits       Exploits       Exploits       Exploits       Generic
 [31] Reconnaissance Exploits       Reconnaissance Exploits       Reconnaissance Reconnaissance
 [37] Reconnaissance Exploits       Exploits       Exploits       Exploits       Exploits
 [43] Exploits       Reconnaissance Exploits       DoS            Generic        Reconnaissance
 [49] Exploits       Reconnaissance Exploits       Exploits       DoS            Exploits
 [55] Shellcode      Shellcode      Exploits       Exploits       Reconnaissance Reconnaissance
 [61] Exploits       Exploits       Exploits       Exploits       Exploits       Exploits
 [67] Exploits       Shellcode      Exploits       Exploits       Reconnaissance Exploits
 [73] Exploits       Exploits       Reconnaissance Reconnaissance Generic        Reconnaissance
 [79] Generic        Shellcode      Generic        Exploits       Exploits       Exploits
 [85] DoS            Exploits       Exploits       Exploits       Exploits       Reconnaissance
 [91] DoS            Reconnaissance Exploits       DoS            Exploits       Reconnaissance
 [97] Exploits       Reconnaissance Exploits       Exploits       Exploits       Exploits
[103] Reconnaissance Exploits       Exploits       Exploits       Exploits       Exploits
[109] Reconnaissance Generic        Exploits       Reconnaissance Shellcode      Reconnaissance
[115] Exploits       Exploits       Exploits       Generic        Exploits       Shellcode
[121] Generic        Exploits       Exploits       Reconnaissance DoS            DoS
[127] Exploits       DoS            Reconnaissance Generic        Exploits       Exploits
[133] Exploits       DoS            Exploits       Reconnaissance Exploits       Exploits
[139] Exploits       Exploits       Exploits       Exploits       Reconnaissance Exploits
[145] Exploits       Shellcode      Exploits       Exploits       Exploits       DoS
```

From a total of 2,540,047 rows of network packet traffic, a total of 321,283 (12.648%) web attacks were detected. The type of Web attack with the highest number of attacks was Generic with a total of 215,481 (67.06%), and Worms represented the fewest attacks with a total of 174 (0.0541%) [4].

The cyber-attacks with the most occurrences and the least occurrences were selected, and statistical concepts of percentiles were applied. The percentiles of Generic attacks are 7,522.00 (0%), 22,792.75 (25%), 44,880.50 (50%), 75,958.00 (75%) and 118,198.00 (100%). The percentiles of Worms attacks are 24 (0%), 36 (25%), 41.50 (50%), 49 (75%) and 67 (100%).

Web Attacks Summary Results

Table 4 shows the total of each type of web attack that was extracted from each dataset and stored in new data frames named NB15_1 -> NB15_4. The results are displayed on the screen.

Table 4
Type of web attacks per dataset (NB15_1, NB15_2, NB15_3, NB15_4)

```
> NB15_1
                      Fuzzers    Analysis   Backdoors        DoS   Exploits
             0           5051         526         534       1167       5409
       Generic Reconnaissance   Shellcode       Worms
          7522           1759         223          24
> NB15_2
                      Fuzzers Reconnaissance   Shellcode     Analysis
             0           4668           3116         324          608
      Backdoor            DoS       Exploits     Generic        Worms
           370           4637          11103       27883           40
> NB15_3
                      Fuzzers Reconnaissance   Shellcode     Analysis
             0           9137           5582         593          873
      Backdoor            DoS       Exploits     Generic        Worms
           759           5642          16574      118198           67
> NB15_4
                      Fuzzers Reconnaissance   Shellcode     Analysis
             0           5390           3530         371          670
      Backdoor            DoS       Exploits     Generic        Worms
           666           4907          11439       61878           43
```

Web Attacks Pie Chart

Figure 2 shows Total Tuples (2,540,047), Total Attacks (321,283), Min (174) & Max (215,481).

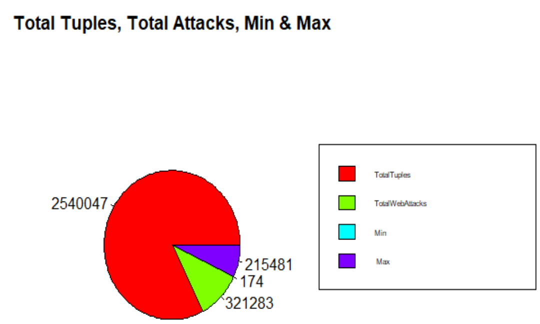

Figure 2
Total Tuples, Total Attacks, Min & Max

Quantiles of Web Attacks Results

Figure 3 Box plots show the five-number summary of a set of data: including the minimum score, first (lower), median, third (upper) quartile, and maximum score. Also explains the interquartile range (IQR) [10].

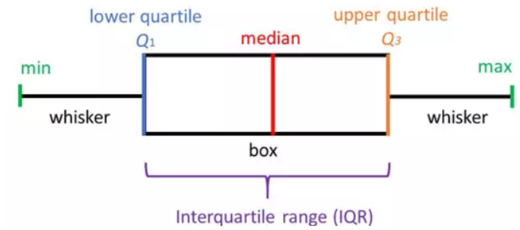

Figure 3
Explanation of Quantile Structure

Figure 4 shows Boxplot of Web Attacks Fuzzers, DoS, Exploits, & Reconnaissance. Min, 1st Quantile, Median, 3rd Quantile & Max. Fuzzers 4,668, 4,955, 5,220, 6320 & 9,137. DoS 1,167, 3,770, 4,772, 5,091 & 5,642. Exploits 5,409, 9,680, 11,271, 12,723 & 16,574. Reconnaissance 1,759, 2,777, 3,323, 4,043 & 5,582.

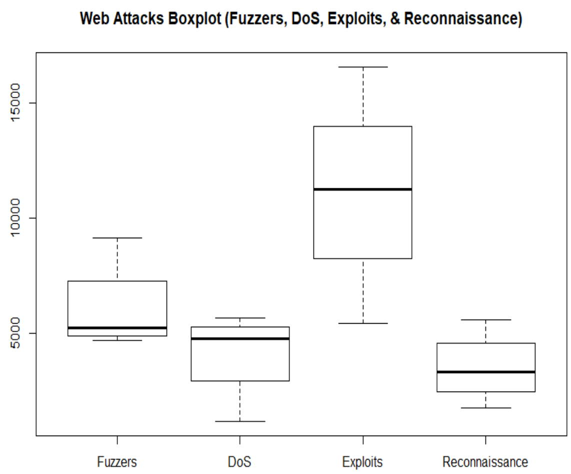

Figure 4
Web Attacks Boxplot (Fuzzers, DoS, Exploits, & Reconnaissance)

Figure 5 shows the Boxplot of different types of web attacks. The type of attacks are Analysis, Backdoors, Shellcode and Worms. Analysis with a Min of 526, 1st Quantile 587.5, Median 639, 3rd Quantile 720.80 and Max 873. Backdoors with a Min of 370, 1st Quantile 493, Median 600, 3rd Quantile 689.20 and Max 759. Shellcode with a Min of 223, 1st Quantile 298.80, Median 347.50, 3rd Quantile 426.50 and Max 593. Worms with a Min of 24, 1st Quantile 36, Median 41.50, 3rd Quantile 49 and Max 67.

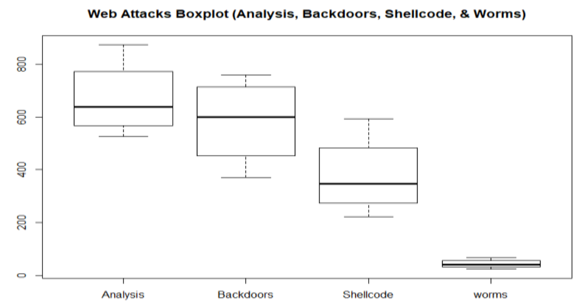

Figure 5
Boxplot of Web Attacks (Analysis, Backdoors, Shellcode, & Worms)

Generic Quantile Percentages Pie Chart

Figure 6 Pie chart of Generic Web Attacks Quantile and Percentile. 0% 7,522, 25% 22,792.75, 50% 44,880.50, 75% 75,958.00, 100% 118,198.00.

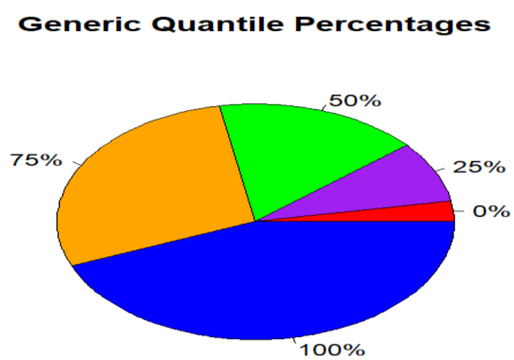

Figure 6
Generic Quantile Percentages

CONCLUSION

By following the steps and selecting the most appropriate data analytics algorithms, the result was the creation of pieces of useful information for analysis and decision making. Information could be extracted, moved, and stored in different data structures [11]. The different results obtained from the data can help the security administrator to see the attack behavior within the network [12].

The main goal of carrying out this web attack analysis project is to be able to contribute to the field of information security. It facilitates knowledge and skills for future information security professionals with examples and techniques of gathering data. It was possible to extract information about web attacks from four datasets using the RStudio and R Language for data analysis.

These network datasets were generated by the Cyber Range Lab of the Australian Center for Cyber Security (ACCS). The results obtained from the different types of attacks and the extraction data obtained can help experts to analyze web cyber-attacks. The objective of extracting data that can be used for analysis and decision making was achieved. The field of cybersecurity is complex, but with education and guidance security experts can learn the knowledge and skills to mine data in the information security field [13][14][15].

Acknowledgement

This material is based upon work supported by, or in part by, the Nuclear Regulatory Commission (NRC) Grant Fellowship Award under contract/ award # NRC-HQ-7P-15-G-0006.

REFERENCES

[1] D. Kaur, P. Kaur, "Empirical Analysis of Web Attacks," *International Conference on Information Security & Privacy*, 2015.Available: https://www.sciencedirect.com/science/article/pii/S1877050916000594 [Accessed on June 8, 2020].

[2] T. M. Davies, *The Book of R: A First Course in Programming and Statistics*. California, USA: No Starch Press, 2016.

[3] J. Grendon, *Introduction to R for Business Intelligence*. Birmingham, UK: Packt Publishing Ltd., 2016.

[4] R. Moustafa, "A Comprehensive Data set for Network Intrusion Detection systems," School of Engineering and Information Technology University of New South Wales at the Australian Defense Force Academy Canberra, Australia, UNSW-NB15, 2015. [Online]. Available: https://ieeexplore.ieee.org/document/7348942 [Accessed on May 20, 2020].

[5] N. Sonule, M. Kalla, A. Jain, & D. S. Chouhan, "UNSW-NB15 Dataset and Machine Learning Based Intrusion Detection Systems," *International Journal of Engineering and Advanced Technology (IJEAT).* ISSN: 2249 – 8958, vol. 9, no. 3, 2020.

[6] D. Canali, D. Balzarotti, "Behind the Scenes of Online Attacks: An Analysis of Exploitation Behaviors on the Web," *20th Annual Network & Distributed System Security Symposium (NDSS 2013)*. San Diego, USA. Hal 00799082, 2013.

[7] E. Levi, Y. Arce, "Worm Propagation and Generic Attacks," *The IEEE Computer Society (IEEE Security & Privacy)*, pp. 63-65, 2005. [Online]. Available: https://ieeexplore.ieee.org/document/1423964 [Accessed on May 20, 2020].

[8] V. Kumar, A. K. Das & D. Sinha, "Statistical Analysis of the UNSW-NB15 Dataset for Intrusion Detection," *Computational Intelligence in Pattern Recognition. Advances in Intelligent Systems and Computing*, vol. 999, 2020.

[9] W. Chang, *R Graphics Cookbook: Practical Recipes for Visualizing Data*. Sebastopol, CA: No Starch Press, 2019.

[10] S. Mcleod, "Box Plots (Also Known as Box and Whister Plots)," *Simply Psychology*, 19 July 2019. [Online] Available: www.simplypsychology.org/boxplots.html [Accessed on July 06, 2020].

[11] J. Yao, Y. Yao, "Web-based information retrieval support systems: building research tools for scientists in the new information age," *2012 Proceedings IEEE/WIC International Conference on Web Intelligence (WI 2003).* [Online]. DOI: 10.1109/WI.2003.1241270.

[12] R. Berthier, D. Korman, M. Cukier, & M. Hiltunen, "The Comparison of Network Attack Datasets: An Empirical Analysis," *2008 11th IEEE High Assurance Systems Engineering Symposium*. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/4708862 [Accessed on June 10, 2020].

[13] O. Tripp, M. Pistoia, S. Fink, S. Sridharan, & O. Weisman, *TAJ: Effective Taint Analysis of Web Applications*. IBM T. J. Watson Research Center, 2009. [Online]. Available: https://www.cs.tufts.edu/comp/150BUGS/taj-2009.pdf [Accessed on Month June 12, 2020].

[14] R. G. Mohammed, "Data Mining Based Network Intrusion Detection System: A Survey," College of Computer Science and Information Technology Sudan University of Science and Technology. *International Journal of Engineering and Advanced Technology (IJEAT)*, 2010. [Online]. DOI 10.1007/978-90-481-3662-9_86.

[15] N. Moustafa, J. Slay, "A hybrid feature selection for network intrusion detection systems: Central points," *Proceedings of the 16th Australian Information Warfare Conference*, pp. 5–13, 2015. [Online]. DOI: 10.4225/75/57a84d4fbefbb. [Accessed on June 11, 2020].