www.dergipark.gov.tr
ISSN:2148-3736
El-Cezerî Fen ve Mühendislik Dergisi
Cilt: 9, No: 2, 2022 (644-655)

El-Cezerî Journal of Science and Engineering
Vol: 9, No: 2, 2022 (644-655)
DOI :10.31202/ecjse.990318

ECJSE

Research Paper / Makale

Classification of Malware in HTTPs Traffic Using Machine Learning Approach

Abhay Pratap SINGH [1a*], Mahendra SINGH [1b]

[1]Department of Computer Science, Gurukula Kangri (Deemed to be University) Haridwar, India
rs.abhaypratapsingh@gkv.ac.in msa@gkv.ac.in

Received/Geliş: 04.09.2021 **Accepted/Kabul:** 13.01.2022

Abstract: Cybersecurity and cyberwar have become crucial for a world backed by continuous development and expansion of digitalization. In the current digital era, malware has become a significant threat for internet users. Malware spreads faster and poses a big threat to cyber security. Hence, network security measures have an important role to play for neutralizing these cyber threats. In our research study, we collected some malicious and self-generated benign PCAP's and then applied a Random Forest (RF) machine learning algorithm to build a traffic classifier. The proposed classifier classifies the HTTPs traffic as benign or malicious one. Experimental results exhibit the average accuracy of 90% and a false-positive rate of 0.030 for RF classifier.

Keywords: Network traffic, Classification, HTTPs, Malware, Wireshark, Machine learning.

Makine Öğrenimi Yaklaşımını Kullanarak HTTPs Trafiğindeki Kötü Amaçlı Yazılımların Sınıflandırılması

Öz: Siber güvenlik ve siber savaş, dijitalleşmenin sürekli gelişimi ve genişlemesiyle desteklenen bir dünya için çok önemli hale geldi. Mevcut dijital çağda, kötü amaçlı yazılım internet kullanıcıları için önemli bir tehdit haline geldi. Kötü amaçlı yazılımlar daha hızlı yayılır ve siber güvenlik için büyük bir tehdit oluşturur. Bu nedenle, ağ güvenlik önlemleri bu siber tehditleri etkisiz hale getirmek için önemli bir role sahiptir. Araştırma çalışmamızda, bazı kötü niyetli ve kendi kendine oluşturulan iyi huylu PCAP'ler topladık ve ardından bir trafik sınıflandırıcısı oluşturmak için bir Rastgele Orman (RF) makine öğrenimi algoritması uyguladık. Önerilen sınıflandırıcı, HTTP trafiğini iyi huylu veya kötü niyetli olarak sınıflandırır. Deneysel sonuçlar, RF sınıflandırıcı için ortalama %90 doğruluk ve 0.030 yanlış pozitif oranı sergiler.

Anahtar Kelimeler: Ağ trafiği, Sınıflandırma, HTTP'ler, Kötü Amaçlı Yazılım, Makine öğrenimi.

1. Introduction

Along the proliferation of the internet usage, the menace of malware is increasing day by day, posing a threat to the integrity, confidentiality, authentication, control flow, and the functionality of a system or network [1]. Any malicious software, that performs unwanted or undesirable action in a computer or a network, can be considered a malware including viruses, logic bombs, worms and spyware etc. [2]. In recent years, due to the widespread attack of malware in the network, the number of incidents related to network security is increasing rapidly year by year. The relevant statistics demonstrates that the number of network security incidents triggered by malware have increased by more than 50% per year since the 90s [3]. As these network security incidents exhibit the vulnerability of system and networks, combating against these attacks is a challenging task for the cybersecurity research community. Nowadays, malware communicate through HTTPs

How to cite this article
Singh A. P., and Singh, M., "*Classification of Malware in HTTPs Traffic Using Machine Learning Approach*" El-Cezerî Journal of Science and Engineering, 2022, 9 (2); 644-655.

Bu makaleye atıf yapmak için
Singh A. P., and Singh, M., "*Makine Öğrenimi Yaklaşımını Kullanarak HTTPs Trafiğindeki Kötü Amaçlı Yazılımların Sınıflandırılması*" El-Cezerî Fen ve Mühendislik Dergisi 2022, 9 (2); 644-655.
ORCID ID: [a]0000-0002-1044-7557; [b]0000-0001-5182-127X

(hypertext transfer communication with SSL/TLS) web traffic to secure their malicious activities. In web address bar, a URL generelly begin with HTTPs:// and the information is transferred over port number 443 by default. The role of HTTPs protocol is to encipher the content between client and server, so that network communications remain secured between both entities. For a legitimate user, encryption is used for a useful purpose; on the other hand, malware developers use it for harmful or illegal intent. Traditional approaches of detecting malware in HTTPs (encrypted) traffic, such as DPI (Deep Packet Inspection) and signature-based antivirus are not liable to apply on encrypted traffic because encryption reduces the efficacy of pattern matching approaches. The use of machine learning techniques for analyzing and detecting malware is a new trend, which is being used by many researchers successfully. The proposed research study primarily focuses on the domain of network security by applying machine learning as well as statistical computing in the field of network security. For this purpose, we collected malicious PCAPs from packet total [4] and generated benign traffic through Wireshark tool [5]. We extracted and calculated more than 80 statistical network traffic features for all benign and malware Biflows by using CICFlowmeter software which is publicly available on Canadian Institute for Cybersecurity website [6]. Finally, we analyzed the self-created dataset for malware category classification.

The main contribution of our study can be stated as follows :

- A machine learning-based classification model is proposed to classify the HTTPs traffic.
- The model was built, trained, and evaluated using RF classifier, which produces a higher accuracy and low false positive rate.

The rest of the paper is organized as follows: section 2 introduced the family of malware. Section 3 elaborates the traditional network traffic classification techniques. Related work is presented in section 4. Methodology and details of the proposed machine learning model are described in section 5. Section 6 presents the experimental results and finally, section 7 concludes the paper.

2. Background

2.1. Malware Family

To have a clear understanding of the methods and rationality behind the malware, it is beneficial to categorize it. Malware can be categorized into several classes, depending on their various purposes. The following list shows the common types of malware.

2.2.1 Virus

It is a malicious code that replicates itself by injecting its code into different platforms, such as, operating system, and entails running within the victim host. The virus can spread very fast in a short period, and it can damage one computer to another with the help of human assistance. Some of the most prevalent virus types are Expiro, Sality, Virat etc.[7]

2.2.2 Worm

This type of malware is different from a virus in the context of the transfer medium. The worm can explore network vulnerability, and it can carry other malware in its payload. It can self-replicate among networks without human assistance. Some of the common worms are Allaple, Vobfus, etc.[7].

2.2.3 Trojan Horse

A Trojan horse works on the principle of client and server. Attacker cleverly designs a malicious software and then send it to the user via social engineering techniques, which appears as a legitimate software to the user. When its payload executed, it performs malicious activities controlled by the attacker.

2.2.4 Spyware

Spyware is a class of software which keeps track of user information, like information of regularly visited websites and credit card number, etc. without victim knowing it. Recently, Pegasus (spyware) came into existence that keeps spying the user's data. It has been developed by the Israeli cyber arms firm, NSO Group [8].

2.2.5 Backdoor

These kinds of malware are continuously exploring the security loopholes or vulnerabilities in the system. Using a backdoor, the attacker can perform malicious activities on a target object, for example, installing a keylogger, illegally accessing useful information and infecting networks as well. Some of the common backdoors are Rbot, Hupigon, Bifrose, etc. [9]

2.2.6 Rootkit

A rootkit is an ensemble of many malicious software such as password stealer, dropper, and bots that allows an unauthorized way to handle administration-level access to a system or network [10]. Rustock and Mebroot are famous examples of rootkit malware.

2.2.7 Adware

Adware is a malicious type of software that exhibits undesirable advertisements on the user's browser, sometimes it raises so many pops options in a browser window and automatically downloads the malicious script and executes it to the victim system [11]. The most popular way to exploit such type of malware is cross-site scripting (XSS) attack and traffic redirection.

2.2.8 Ransomware

Ransomware is an amalgamation of ransom and malware that has a prime objective of enciphering the content of the user's data and then asks for a ransom to allow the user system access [12]. Wanna cry, locky, and cerber are the widespread ransomware attacks.

2.2.9 Digital Currency Mining Malware

Cryptocurrency mining malware is a new term that tries to access system resources for digital mining currency without knowing the user's consent [12]. This type of malware typically reduces the processing speed of system.

3. Network Traffic Classification Techniques

Researchers have conducted various studies associated with malware analysis and detection based on network traffic features. Within the domain of network security, network traffic classification has gained much popularity in security-related applications in firewalls and intrusion detection systems. Further, network traffic classification is divided into four main categories viz. port-based classification, payload classification, behavioral classification, and statistical features based classification, as shown in Figure 1.

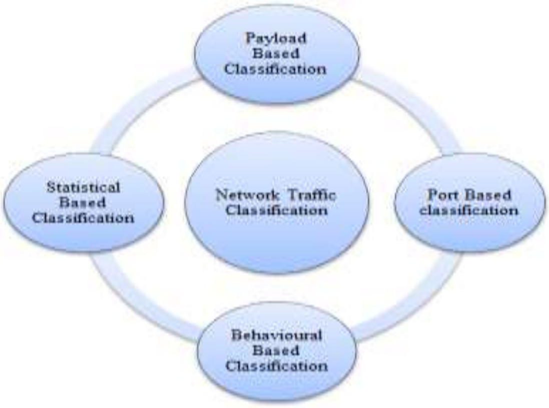
Payload Based Classification
Statistical Based Classification
Network Traffic Classification
Port Based classification
Behavioural Based Classification

Figure 1. Traffic Classification

3.1. Port Based Classification

Port-based classification is mainly concerned with endpoint's connection or TCP/UDP port numbers allotted by IANA. It is a useful technique and simple to implement for network traffic identification. However, in modern applications, which make use of dynamic port numbers unregistered with IANA, the accuracy of port-based classification is significantly decreasing [13]. Despite its inaccuracy, the port-based classification is still extensively utilized.

3.2.Payload based Classification

This type of method is also called as Deep Packet Inspection (DPI). Payload based classification can be classified into two subclasses. The first type of methods are based on DPI, which tries to match a given set of signatures against packet payload and the second ones are Stochastic Packet Inspection (SPI) methods, which check the statistical properties of packet content [14]. This is a viable alternative to port-based classification. DPI solution is good for unencrypted traffic, but when it comes to encrypted traffic, this solution does not provide accurate results.

3.3 Behavioral Based Classification

This type of classification method mainly deals with packets and a flow based analysis. It is concerned with traffic pattern between client and server such as the number of systems communicated, the protocol used, and the bidirectional flows being used on the host [15-18]. Behavior based approaches yield accurate classification to the near optimum level with decreased overhead when compared to the payload based method [19].

3.4 Statistical Based Classification

Classification based on statistical features is regarded trivial and highly extensible from the functioning aspect [20]. The statistical properties like packet length, flow duration, forward packets, and backward packets, are used to classify the network traffic based on many applications. This type of method is highly reliable even in encrypted traffic. In our study, we are primarily focused on web-based (HTTPs protocol) traffic for effectively classify it.

Table 1. Summary of Traffic Classification

Category	Features	Advantages	Disadvantages
Port-based method	Protocol port numbers	Easy to implement. Very efficient in large networks.	Some applications use ports other than its well-known ports.
Payload inspection method	Rely on packet header and payload	It’s good for early classification	Privacy policy breaching and unable to perform on encrypted traffic.
Behavioral techniques	Host level and end-mile connection	It can achieve good accuracy in less information	It requires huge flows to be analyzed before successful application identification
Statistical method	Flow and packet-based features	Effectively handle both encrypted and unencrypted traffic. Can detect real-time traffic identification.	Performance heavily relies on the human-engineered features.

4. Related Works

In recent years, network traffic classification techniques based on machine learning, have been extensively utilized to identify and detect malware in HTTPs traffic. The purpose of a machine learning algorithm is to provide data access to machines and let them learn by themselves [21]. It enables computer systems to learn from the data by the use of statistical techniques. Classification models usually require a substantial amount of labeled samples for training, which results in the enhanced ability of the classification model to identify the trained samples [22]. We are reviewing here a few appreciable contributions made towards classification of the encrypted traffic by using the machine learning techniques.

Lokoc et al. in [23] proposed a classifier to detect the secure HTTP connection related malware families. The testing was completed with the help of an ECM linear classifier. The metric index k-NN classification method improves the efficiency of detecting malware in HTTPs traffic on the small dataset of high-dimensional network traffic descriptors, which reduced the false positive rate. However, authors have not shown the average accuracy of classifier but focused on the problem of recognizing malicious servers instead of understanding malicious traffic of various types.

The authors in [24] developed and studied LSTM-based malware detection model that utilizes only the observable aspects of HTTPS traffic. LSTM networks [25] are extensively utilized for speech recognition, translation, and natural-language-processing tasks. The developed method makes it possible to gather huge size of malicious and benign network traffic for model evaluation. Malware is detected in the context of the host address, timestamps, and data volume information of the computer network traffic.

The authors in [26] proposed a behavior testing method to identify HTTP and HTTPs network packets in a more detailed manner and utilized machine learning techniques to detect malware characteristics. The experimental results exhibit that precision and recall are more than 96% on an average. In future, authors proposed to simulate in real-time environment to detect malware and to discover TLS metadata in a more precise way.

5. Proposed Machine Learning Model

The proposed classification model is illustrated in Figure 2. The prime objective of the proposed model is to identify/classify benign and malicious HTTPs network traffic with the help of a machine learning technique.

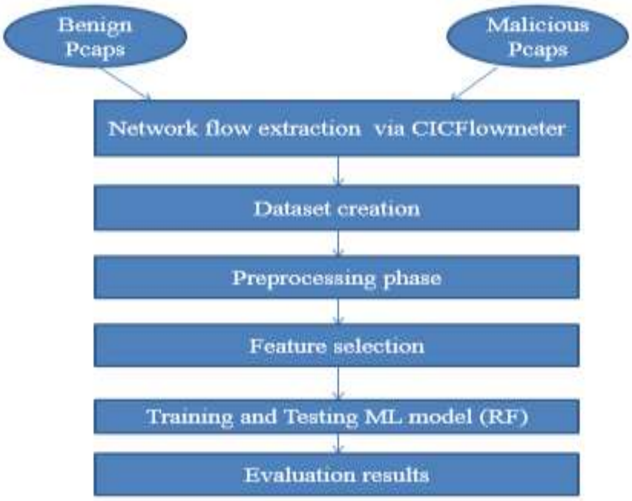

Figure 2. Methodologies of Proposed Model

5.1 Network Flow Extraction

First and foremost, we generated benign network traffic data and collected malicious samples [4]. In this step, statistical network traffic features are extracted via the CICFlowMeter tool [6] from both benign and malicious PCAP's. This tool can extract more than 80 network flow features. The period of flow is defined by five attributes, *Source IP, Destination IP, Source port, Destination port, and Protocol.*

5.2 Dataset

After extraction of statistical network flow features, we labeled them as which one is specific malware or benign network flow traffic. We divided our dataset into four classes Benign, Scareware, Ransomware, Adware (as shown in Figure 3).

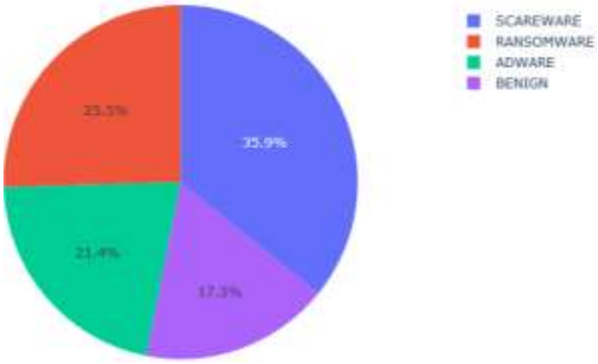

Figure 3. Dataset Class.

5.3 Preprocessing Phase

Preprocessing is a vital phase to maneuver real-world data into a logical layout. In the real-world, data is often shortened and noisy in a specific behavior. Here, we used a tool open refine [27] previously called as Google refine for preprocessing purpose in order to enhance the superiority of data.

5.4 Feature Selection

This section impersonates an essential part of our research. We extracted multiple statistical features of the network traffic, but we need to determine correct and appropriate features related to our research. A feature selection policy identifies only relevant features thus reducing the data dimension. We applied the backward elimination feature selection technique, which comes under the category of wrapper method and infogain with the ranker search in Weka tool [28] and finds the most common features, which will be used in the next phase. In backward elimination, we begin with initial features and eliminate the least essential features at every stage, which further enhances the performance of the model. We reiterate this process until no improvement is seen on the deletion of features. Table 2 shows the selected features out of 80 statistical features of the network traffic.

Table 2. List of Selected Statistical Features of the Network Traffic

Feature Name	Meaning
SIP, DIP, Sport, Dport	The period of the flow
Total fwd packets	Total packets in forward track
Total bwd packets	Total packets in backward track
Total length of fwd packets	Total length of forward packets
Total length of bwd packets	Total length of backward packets
Fwd packet length (min, max, mean, std)	Min, max, mean, std of the packet size in forward track
Bwd packet length (min, max, mean, std)	Min, max, mean, std of the packet size in backward track
Forward header length	The total number of bytes used for headers in forward track
Backward header length	The total number of bytes used for headers in backward track
Fwd packets/s	The number forward packets per second
Bwd packets/s	The number backward packets per second
Flow packet length (min, max, mean, std, variance)	Min, max, mean, std of the length of flow

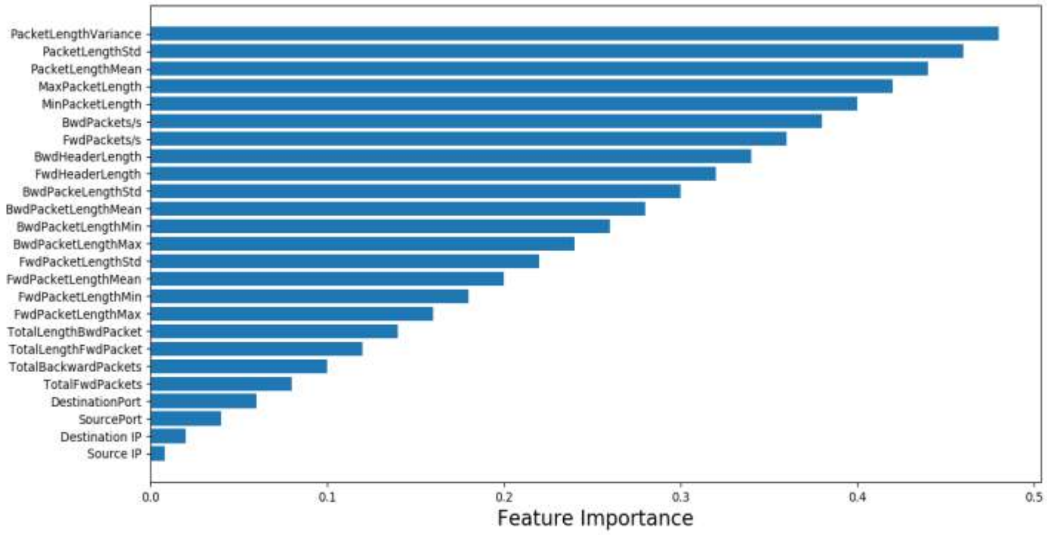

Figure 4. Feature Visualization

5.5 Training and Testing

We applied the Weka simulation tool for training and testing the model. We built and trained our model with 10-fold cross-validation on the 80% of training data, and evaluated our model on the 20% of testing data by applying Random Forest classifier. Random Forest is an ensemble classifier that performs better in contrast with other conventional classifiers for efficient classification of network traffic. Random forest algorithm is one of the most popular classifier among all other machine learning algorithms. It utilizes an ensemble method that consists of various decision trees to make predictions using the voting process. The main advantage of random forest is, that it can be utilized for both regression and classification problems, which form most current machine learning systems. The training and testing parameters used in the simulation tool for the classification purpose are as shown in Table 3.

Table 3. Summary of Parameters

Parameters	Value
Number of iterations	100
Batch size	100
Random seed	1
Number of execution slots	1

5. Experiment Results

For the implementation of the proposed approach, we setup the experiment by installing the software tools on the machine. The experimental environment used in our research is system with Windows 10 operating system, 2.00GHz Intel core i3 CPU, 8GB memory and the Python version 3.6. After the application of Random Forest classifier to our dataset, simulation tool provided a detailed output of the applied algorithm. Tabel 4 illustrates that random forest classifier produces accurate and reliable results with relevant features for the testing data set.

Table 4. Efficiency and Accuracy of Model in Terms of Classification Results

Result Parameters	Value
Correctly Classified Instances%	90.4762 (456)
Incorrectly Classified Instances %	9.5238 (48)
Kappa statistic	0.8694
Mean absolute error	0.097
Root mean squared error	0.1963
Relative absolute error%	26.5153
Root relative squared error%	45.9046
Total Number of Instances	504

Table 5 represents the result of classification model for testing data set in terms of several performance metrics. Random forest classifier takes 1.39 second to build basic model and 0.03 second to build test model on the test set. Confusion matrix is shown in Table 6 which is the basis for checking accuracy and credibility of the proposed model. Simulation results shows that the proposed approach is feasible and effective to classify malware attacks when compared with the approach proposed by Lokoc et al. [23]. To evaluate the performance of the proposed classification model, we need to know the significance of some of the essential evaluation metrics. Some of the common metrics are defined here under.

Table 5. Detailed Average Accuracy for all Classes of Malware

TPR	FPR	Precision	Recall	F-Measure	ROC Area	Class
0.793	0.031	0.833	0.793	0.813	0.977	BENIGN
0.934	0.024	0.934	0.934	0.934	0.987	RANSOMWARE
0.827	0.046	0.835	0.827	0.831	0.960	ADWARE
0.983	0.024	0.956	0.983	0.969	0.995	SCAREWARE
0.905	0.030	0.904	0.905	0.904	0.982	Weighted Avg.

Table 6. Confusion Matrix

a	b	c	d	<-- classified as
65	0	17	0	a = BENIGN
1	127	1	7	b = RANSOMWARE
11	7	91	1	c = ADWARE
1	2	0	173	d = SCAREWARE

6.1 Accuracy

It is the proportion of the correctly identified malicious traffic to the whole size of the test set.

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

where the parameters are TP = True Positive, FP = False Positive. , TN = True Negative and FN = False Negative

6.2 True Positive Rate (TPR)

It is the proportion of malware packets classified as malware among all packets which truly have malware class.

$$TPR = \frac{TP}{TP+FN} \tag{2}$$

6.3 False Positive Rate (FPR)

It is the proportion of benign packets that are flagged as malicious packets to the total number of benign packets.

$$FPR = \frac{FP}{FP+TN} \tag{3}$$

6.4 Precision

The proportion of the correctly identified malware packets to the number of all identified malicious packets.

$$precision = \frac{TP}{TP+FP} \tag{4}$$

6.5 Recall

It points to the total percentage of relevant results out of correctly classified through an algorithm.

$$Recall = \frac{TP}{TP+FN} \tag{5}$$

6.6 Roc (Receiver Operating Characteristic)

It is one of the most important evaluation metric for checking the performance of a classification model. The basic purpose of this curve is to introduce the feat relationship between TP and FP. Roc curve in Figure 6 shows that the Roc value is closer to 1, which is the perfect classification rate [29].

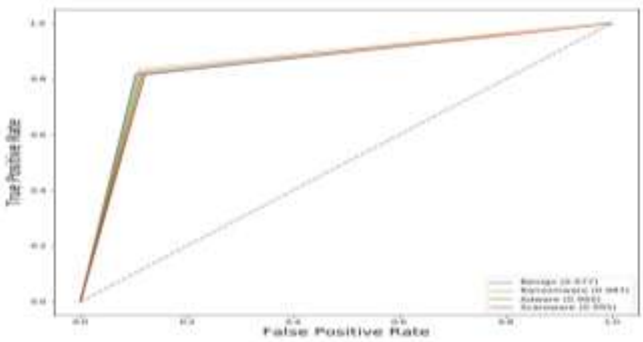

Figure 5. Roc Curve

6. Conclusion

In the current circumstances of cybersecurity, network traffic classification is not easy as traditional techniques have their limitations. The traditional approaches are not capable of combating modern threats like encrypted traffic, zero-day malware attack, and ransomware. To classify threats in HTTPs traffic is an cumbersome task for the researchers due to the in-built encryption. In modern era of artificial intelligence, machine learning based technologies have attained popularity in classifying these advanced security threats. In this paper, the proposed classification model classify both benign as well as malicious traffic without decrypting the network traffic by using the random forest ML algorithm. Experimental results indicate that the average accuracy of proposed classification model is 90% with a precision of 0.904. Another merit of proposed model is that it utilizes statistical-based features, which include flow-based, packed-based, and behavioral-based features to classify both known and unknown malware. In future, we propose to develop a malware detection mechanism based on deep learning technique.

Authors' Contributions

The authors read and approved the final manuscript.

Competing Interests

The authors declare that they have no competing interests.

Refereneces

[1]. Singh A.P., Singh M., "A comparative review of malware analysis and detection in Https traffic" International Journal of Computing and Digital Systems (IJCDS), 2021, 10(1): 111-123.

[2]. McCarthy C., Zincir-Heywood., "An investigation on identifying SSL traffic" IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA), 2011, 115–122.

[3]. Husák M., Čermák M., Jirsík T., Čeleda P., "HTTPS traffic analysis and client identification using passive SSL/TLS fingerprinting" EURASIP Journal on Information Security, 2016, (1): 1-14.

[4]. Becker, Jamin., https://packettotal.com, 12.03.2021.

[5]. Wireshark., https://wireshark.org, 28.03.2021.

[6]. CICFlowMeter., https://www.unb.ca/cic/reserach/applications.html, 28.03.2021.

[7]. What is a computer virus or a computer worm., https://usa.kaspersky.com/resource-center/threats/computer-viruses-vs-worms, 10.06.2021.

[8]. Bill Marczak., John Scott-Railton., Sarah Mckune., Bahr Abdulrazzak., Ron Deibert., "HIDE AND SEEK Tracking NSO Group's Pegasus Spyware to Operations in 45 Countries" University of Toranto, 2018, Citizen Lab Resrach Report No: 113.

[9]. What is a backdoor., https://www.wired.com/2014/12/hacker-lexicon-backdoor/, 10.06.2021.

[10]. Kim S., Park J., Lee K., You I., Yim K., "A brief survey on rootkit techniques in malicious Codes" Journal of Internet Services and Information Security, 2012, 3(4):134-147.

[11]. Malode S.K., and Adware R.H., "Regenerative braking system in electric vehicles" International Research Journal of Engineering and Technology (IRJET), 2016, 3(3):394-400.

[12]. Mohurle S., Patil M., "A brief study of wanna cry threat: Ransomware attack" International Journal of Advanced Research in Computer Science (IJARCS), 2017, 8(5):1938-1940.

[13]. Rezaei S., Liu X., "Deep learning for encrypted traffic classification: An overview" IEEE communications magazine, 2019, 57(5):76-81.

[14]. Valenti S., Rossi D., Dainotti A., Pescapè A., Finamore A., Mellia M., "Reviewing traffic classification" In Data Traffic Monitoring and Analysis, Lecture Notes in Computer Scince Springer, Berlin, Heidelberg, 2013, 123-147.

[15]. Zhao J., Jing X., Yan Z., Pedrycz W., "Network traffic classification for data fusion: A survey", Information Fusion, 2021, 72:22-47.

[16]. Karagiannis T., Papagiannaki K., Taft N., Faloutsos M., "Profiling the end host", In International Conference on Passive and Active Network Measurement, Lecture Notes in Computer Science, 2007, Springer, Berlin, Heidelberg, 186-196.

[17]. Xu K., Zhang Zhi., Bhattacharyya S., "Profiling internet backbone traffic: behavior models and applications", ACM SIGCOMM Computer Communication Review, 2005, 35(4): 169–180.

[18]. Iliofotou M., Pappu P., Faloutsos M., Mitzenmacher M., Singh S., Varghese G., "Network monitoring using traffic dispersion graphs (TDGs)", Proceedings of the 7th ACM SIGCOMM Internet Measurement Conference, 2007, San Diego, California, USA, 315-320.

[19]. Bermolen P., Mellia M., Meo M., Rossi D., Valenti S., "Abacus: accurate behavioral classification of P2P-TV traffic", Computer Networks, 2011, 55(6): 1394–1411.

[20]. Bakhshi T., Ghita B., "On internet traffic classification: A two-phased machine learning approach" Journal of Computer Networks and Communications, 2016, 1-21.

[21]. Zhang XD., "Machine learning", A Matrix Algebra Approach to Artificial Intelligence, 2020 Springer, Singapore.

[22]. Zheng R., Liu J., Niu W., Liu L., Li K., Liao S., "Preprocessing method for encrypted traffic based on semi supervised clustering", Security and Communication Networks, 2020, 1-13.

[23]. Lokoc J., Kohout J., Cech P., Skopal T., Pevny T., "k NN classification of malware in Https traffic using the metric space approach" In Pacific Asia Workshop on Intelligence and Security Informatics, 2016, Springer, Cham,131–145.

[24]. Prase P., Machlica L., Pevny T., Harvelka J., Scheffer T., "Malware detection by analyzing encrypted network traffic with neural networks" In Joint European Conference on Machine Learning and Knowledge Discovery in Database, 2017, Springer, Cham ,73-88.

[25]. Soutner D., Müller L., "Application of LSTM neural networks in language modelling", In International Conference on Text, Speech and Dialogue, 2013, Springer, Berlin, Heidelberg, 105-112.

[26]. Calderon P., Hasegawa H., Yamaguchi Y., Shimada H., “Malware detection based on Https characteristics via machine learning”, In Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP), 2018, 410-417.

[27]. Openrefine., https://openrefine.org/, 05.07.2021.

[28]. Waikato Environment for Knowledge Analysis (WEKA)., https://www.cs.waikto.ac.nz/ml, 05.07.2021.

[29]. Moustafa N., Hu J., Slay J., “A holistic review of network anomaly detection systems: A comprehensive survey”, Journal of Network and Computer Applications, 2019, 128:33-35.