Uluslararası Mühendislik Araştırma ve Geliştirme Dergisi
International Journal of Engineering Research and Development

UMAGD, (2022) 14(2), 794-798.

10.29137/umagd.1117634

Cilt/Volume:14 Sayı/Issue:2 Temmuz/July 2022

Araştırma Makalesi / Research Article

Darknet Web Traffic Classification via Gradient Boosting Algorithm

Fahrettin Horasan [1], Ahmet Haşim Yurttakal[2]

[1]Kırıkkale University, Engineering Faculty, Department of Compuer Engineering, Kırıkkale, Turkey
[2]Afyon Kocatepe Üniversity, Engineering Faculty, Department of Computer Engineering, Afyon, Turkey

***Başvuru/Received**: 16/05/2022 **Kabul / Accepted**: 29/07/2022 **Çevrimiçi Basım / Published Online:** 31/07/2022*
***Son Versiyon/Final Version**: 31/07/2022*

Abstract
Classification of network traffic not only contributes to improving the quality of network services of institutions, but also helps to protect important data. Machine learning algorithms are frequently used in the classification of network traffic, since port-based and load-based classification processes are insufficient in encrypted networks. In this study, VPN and Tor network traffic combined in the Darknet category was classified with the Gradient Boosting Algorithm. 70% of the dataset is reserved for training and 30% for testing. 10 fold cross validation was applied in the training set. Network Flows in 8 different categories: Audio-Streaming, Browsing, Chat, E-mail, P2P, File Transfer, Video-Streaming and VOIP were classified with 99.8% accuracy. The proposed method automated the process of network analysis from the Darknet. It enabled organizations to protect the important data with high accuracy in a short time.

KeyWords
"Network flows, web traffic, darknet, gradient boosting, classification"

Sorumlu Yazar: fhorasan@kku.edu.tr

1. Introduction

The Internet is one of the greatest inventions in human history. Its development still continues today. According to a study on digital life, as of January 2021, the number of people using the internet in the world has increased by 7.3 percent compared to the previous year and has been determined as 4.66 billion(Anonymous, 2021).There are millions of web pages, databases and servers running constantly on the internet network. Websites that can be found on this network using known search engines are called the surface network. The network of hidden websites that can be accessed with a special web browser such as Tor is called the Darknet. Used to keep internet activity anonymous and private (Sui et al. 2015).The Darknet can contain malware such as keyloggers, botnets, ransomware. Analyzing Darknet traffic helps with early monitoring of malware before an attack and detecting malicious activity after an outbreak (Kaur and Randhawa, 2020).

Classification of web traffic flows is used effectively in areas such as network monitoring, service quality, intrusion detection and network security. Network flows created by packets that have the same source IP, port, destination IP, port, TCP, and UDP information are determined by which application they belong (Cao et al., 2014) . One of the methods used to classify web traffic is the classification for the use of port numbers. It can be easily classified according to the port information determined by IANA. For example, HTTP uses port 80, SSH uses port 22, SMTP uses port 25. However, for some applications, there may not be a port number registered with IANA (Degermark et al. 1999). In addition, the use of dynamic port numbers or masking techniques has led to the inadequacy of port-based classification systems. Load-based classification systems, which is another classification method, gave more successful results than port-based systems after examining the content of the load. However, its insufficiency in encrypted networks has left its place to machine learning-based approaches due to costs (Karagiannis et al.,2004).

Machine learning is the general name for automated data analysis methods for statistical pattern recognition and modelling. It allows learning new information from data, finding hidden information or patterns (Petit et al. 2018). Successful studies have been made in classifying network traffic with machine learning techniques. Jun et al. (2007) proposed a method using Genetic Algorithm-based feature selection and machine learning algorithms such as C4.5, Random Forest, K Nearest Neighborhood. Liu et al. (2007), use K-Means Clustering to classify network traffic and achieve 90% accuracy.

The studies on the CICDarkNet2020 dataset have been examined in detail. Li and Lu (2021) performed a two-stage classification in their proposed deep learning model. In the first stage, VPN-nonVPN-Tor-nonTor classification was made, while in the second stage, an application-based classification was made. As a result, an accuracy of 97.65% was obtained. Li et al. (2021), CNN and K-means Clustering combined methods obtained 97.4% accuracy. Iliadis and Kaifas (2021), Benign and Darknet binary classification and Tor-nonTor-VPN-nonVPN multiclass classification. They obtained an average of 98% accuracy in their studies using the Random Forest Algorithm. Jadav et al. (2021) benign Darknet classification in their work, Synthetic Minority Oversampling Technique and Principal Component Analysis, after making classification with machine learning algorithms.

Recently, Rust-Nguyen and Stamp (2022) made classification with Support Vector Machines, Random Forest, Convolutional Neural Networks and Generative Adverserial Networks. Gupta, Jindal, and Pedi (2021) used the Extreme Gradient Boosting algorithm to divide the dataset into 3 classes: normal traffic, TOR traffic and VPN traffic. Júnior and Bianchi (2021) on the other hand, tested the performance of the Deep Reinforcement Learning Algorithm. Chang and Branco (2021) recommended Graph-based methods. In another study, Sarwar et al. (2021) achieved 96% accuracy in detecting darknet traffic via Convolutional Neural Network-Long Short Term Memory. Aswad and Sonuç (2020) reached 96.76% accuracy in their classification using Apache Spark and artificial neural networks. Singh et al. (2021) After transforming the time-based features into a three-dimensional image with the DeepInsight method, it classified with 10 different pre-trained models such as AlexNet and ResNet50. As a result, the VGG19 achieved 96% accuracy. Demertzis et al. (2021) proposed the weight agnostic neural networks method to classify real-time network traffic. As a result, it achieved 94% accuracy.

In this study, encrypted network traffic in VPN and DarkNet networks is classified by Gradient Boosting Algorithm. As a result, it achieved a high accuracy rate of 99.8%. In classification performance, the F1 Score value of each class has exceeded 99%. The proposed method is fast, highly accurate, time-saving, user-independent and inexpensive. In the second part of the article, the material method was presented, in the third part the experimental findings, and in the fourth part the discussion and conclusion were presented.

2. Material And Methods

2.1. Dataset

In this study, CICDarkNet2020 dataset was used. The dataset is a combination of ISCXTor 2016 and ISCXVPN2016 datasets covering Tor and VPN traffic. There are 8 different types of encrypted benign and dark network traffic data: Browsing, P2P, Audio-Streaming, Chat, File-Transfer, Video-Stream, Email, VOIP. It consists of 116711 rows and 69 columns (Lashkari et al. 2020). In Table 1, traffic types of the dataset, label numbers assigned to each traffic type and application examples are given.

Table 1. Traffic Types

Label ID	Traffic Category	Applications used
1	Browsing	Firefox and Chrome
2	P2P	uTorrent and Transmission (BitTorrent)
3	Audio-Streaming	Vimeo and Youtube
4	Chat	ICQ, AIM, Skype, Facebook and Hangouts
5	File-Transfer	Skype, SFTP, FTPS using Filezilla and an external service
6	Video-Stream	Vimeo and Youtube
7	Email	SMTPS, POP3S and IMAPS
8	VOIP	Facebook, Skype and Hangouts voice calls

2.2. Gradient Boosting Algorithm

Gradient boosting is one of the ensemble learning techniques used in classification and regression tasks. A collection of weak predictive models typically produces a prediction model in the form of decision trees. It builds the model incrementally, as other augmentation methods do, and generalizes it, allowing optimization of the loss function. They are algorithms that optimize a cost function over the function space by iteratively selecting a function pointing to the negative gradient direction. (Friedman, 2001).

2.3. Evaluation Metrics

In order to evaluate the performance of the gradient boosting model proposed in the study, the error matrix, in which the predictions of the target feature and the actual values are compared, was used. Figure 1 shows confusion matrix.

		Predicted labels	
		Negative (N)	Positive (P)
True labels	Negative (N)	TN	FP
	Positive (P)	FN	TP

Figure 1. Confusion matrix

In the confusion matrix, the terms given can be used to calculate various metrics.The formulas of the metrics used in the evaluation are given in the following equations (1-4).

$$\text{Recall} = TP / (TP + FN) \quad (1)$$

$$\text{Precision} = TP / (TP + FP) \quad (2)$$

$$\text{Accuracy} = (TP + TN) / (P + N) \quad (3)$$

$$\text{F1 Score} = 2TP / (2TP + FP + FN) \quad (4)$$

3. Experimental Results

The application was developed in an open source Python environment. 70% of the dataset is reserved for training and 30% for testing. The features in the dataset of network traffic were used in the classificationThe training and test data numbers for each traffic type are given in Table 2.

Table 2. Train and Test Sets

Label ID	Traffic Type	Train Set	Test Set	Total
1	Browsing	22947	9736	**32683**
2	P2P	16776	7388	**24164**
3	Audio-Streaming	12535	5359	**17894**
4	Chat	8012	3450	**11462**
5	File-Transfer	7817	3276	**11093**
6	Video-Stream	6843	2874	**9717**
7	Email	4253	1884	**6137**
8	VOIP	2514	1047	**3561**
Total		**81697**	**35014**	**116711**

In the preprocessing phase, the timestap property is converted to seconds. All features used in classification have been digitized.In the classification phase, Gradient Boosting Algorithm hyper parameters have been determined. The learning rate is 0.1, the loss function

is log_loss, the boosting number is 100, and the maximum depth is 3.The confusion matrix resulting from the classification is given in Figure 2.

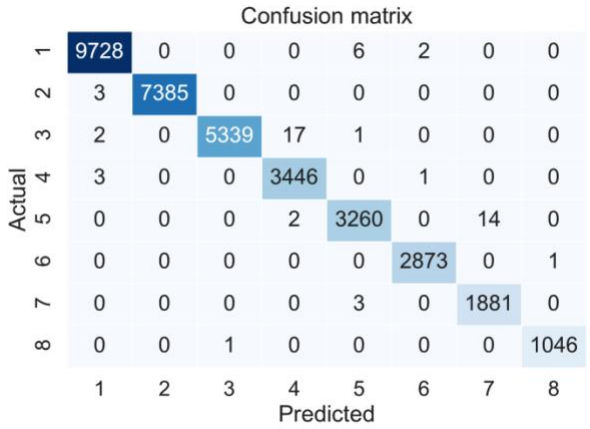

Figure 2. Confusion matrix

The dataset is trained with other machine learning algorithms.KNeighbors Algorithm reached 87%, Logistic Regression 78%, Random Forest 98%, Support Vector Machine 82%, Linear Discriminant Analysis 70%, Extra Tree Classifier 96% accuracy values. Gradient Boosting Algorithm achieved the most successful result with 99.8% accuracy.Other performance metrics obtained are given in Table 3. The proposed algorithm has achieved over 99% success in every metric for all classes.

Table 3. Performance metrics

Label Id	Class	Recall	Precision	F1-Score
1	Browsing	99.918	99.918	99.918
2	P2P	99.959	100.00	99.979
3	Audio-Streaming	99.627	99.981	99.803
4	Chat	99.884	99.452	99.667
5	File-Transfer	99.512	99.694	99.602
6	Video-Stream	99.965	99.896	99.930
7	Email	99.841	99.261	99.550
8	VOIP	99.904	99.904	99.904

Kim and Lee (2022) used XGBoost and LightGBM techniques for darknet traffic detection and classification. LightGBM algorithm showed faster and higher performance than XGBoost, reducing hyper parameter tuning time by more than 10 times. They achieved an F1 score of 94% for Browsing. Sridhar and Sanagaravapu (2021) performed feature selection on the dataset with the Chi-Square method, and then used Generative Contradiction Networks to eliminate the imbalance of the classes. In the classification phase, they obtained an F1 score of 97.87% with the Random Forest classifier. Lan et al. (2022) obtained 92.22% accuracy and 92.10% macro F1 score as a result of the classification they made with Convolutional Neural Network and bidirectional Long Short Term Memory network. In this study, over 99% F1 score was obtained for all classes with the proposed methods and parameters.

4. Conclusion

Classification of network traffic is used to increase service quality, to analyze network traffic data properly, and to detect attacks. In this study, encrypted network traffic in VPN and DarkNet networks is classified. Network traffic is classified only on the basis of application. After all network features have been digitized, they have been classified with 7 different machine learning algorithms: KNeighbors Algorithm Logistic Regression, Random Forest, Support Vector Machine, Linear Discriminant Analysis, Extra Tree Classifier and Gradient Boosting Algorithm. As a result, the Gradient Boosting Algorithm achieved the highest success with an accuracy rate of 99.8%.The proposed method saves time and is user-independent and has a high accuracy rate. The obtained results are encouraging for future studies.

Acknowledgements

There is no conflict of interest. There is no funding. The authors approved the final version of the article.

References

Aswad, S. A., & Sonuç, E. (2020, October). Classification of VPN network traffic flow using time related features on Apache Spark. In 2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT) (pp. 1-8). IEEE.

Cao, Z., Xiong, G., Zhao, Y., Li, Z., &Guo, L. (2014, November). A survey on encrypted traffic classification. In International Conference on Applications and Techniques in Information Security (pp. 73-81). Springer, Berlin, Heidelberg.

Chang, L., & Branco, P. (2021). Graph-based Solutions with Residuals for Intrusion Detection: the Modified E-GraphSAGE and E-ResGAT Algorithms. arXiv preprint arXiv:2111.13597.

de Araújo Júnior, S. R., & Bianchi, R. A. (2021, November). A Model for Traffic Forwarding through Service Function Chaining using Deep Reinforcement Learning Techniques. In Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional (pp. 619-630). SBC.

Degermark, M., Nordgren, B., & Pink, S. (1999). RFC2507: IP header compression. RFC Editor.

Demertzis, K., Tsiknas, K., Takezis, D., Skianis, C., & Iliadis, L. (2021). Darknet traffic big-data analysis and network management for real-time automating of the malicious intent detection process by a weight agnostic neural networks framework. Electronics, 10(7), 781.

Digital 2021: the latest insights into the state of digital - We Are Social UK, We Are Social UK, Jan. 27, 2021. [Online]. Available: https://wearesocial.com/uk/blog/2021/01/digital-2021-the-latest-insights-into-the-state-of-digital/. [Accessed: May 16, 2022]

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

Gupta, N., Jindal, V., & Bedi, P. (2022). Encrypted Traffic Classification Using eXtreme Gradient Boosting Algorithm. In International Conference on Innovative Computing and Communications (pp. 225-232). Springer, Singapore.

HabibiLashkari, A., Kaur, G., &Rahali, A. (2020, November). DIDarknet: A Contemporary Approach to Detect and Characterize the Darknet Traffic using Deep Image Learning. In 2020 the 10th International Conference on Communication and Network Security (pp. 1-13).

Iliadis, L. A., &Kaifas, T. (2021, July). Darknet Traffic Classification using Machine Learning Techniques. In 2021 10th International Conference on Modern Circuits and Systems Technologies (MOCAST) (pp. 1-4). IEEE.

Jun, L., Shunyi, Z., Yanqing, L., &Zailong, Z. (2007, August). Internet traffic classification using machine learning. In 2007 Second International Conference on Communications and Networking in China (pp. 239-243). IEEE.

Karagiannis, T., Broido, A., Faloutsos, M., &Claffy, K. C. (2004, October). Transport layer identification of P2P traffic. In Proceedings of the 4th ACM SIGCOMM conference on Internet measurement (pp. 121-134).

Kaur, S., & Randhawa, S. (2020). Dark web: A web of crimes. Wireless Personal Communications, 112(4), 2131-2158.

Kim, J., & Lee, S. J. (2022). Darknet Traffic Detection and Classification Using Gradient Boosting Techniques. Journal of the Korea Institute of Information Security & Cryptology, 32(2), 371-379.

Lan, J., Liu, X., Li, B., Li, Y., & Geng, T. (2022). DarknetSec: A novel self-attentive deep learning method for darknet traffic classification and application identification. Computers & Security, 116, 102663.

Li, Y., & Lu, Y. (2021). ETCC: Encrypted two-label classification using CNN. Security and Communication Networks, 2021, 1-11.

Li, Y., Lu, Y., & Li, S. (2021, May). EZAC: Encrypted Zero-day Applications Classification using CNN and K-Means. In 2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD) (pp. 378-383). IEEE.

Petit, C., Bezemer, R., &Atallah, L. (2018). A review of recent advances in data analytics for post-operative patient deterioration detection. Journal of clinical monitoring and computing, 32(3), 391-402.

Rust-Nguyen, N., & Stamp, M. (2022). Darknet Traffic Classification and Adversarial Attacks. arXiv preprint arXiv:2206.06371.

Sarwar, M. B., Hanif, M. K., Talib, R., Younas, M., & Sarwar, M. U. (2021). DarkDetect: darknet traffic detection and categorization using modified convolution-long short-term memory. IEEE Access, 9, 113705-113713.

Singh, D., Shukla, A., & Sajwan, M. (2021). Deep transfer learning framework for the identification of malicious activities to combat cyberattack. Future Generation Computer Systems, 125, 687-697.

Sridhar, S., & Sanagavarapu, S. (2021, November). DarkNet Traffic Classification Pipeline with Feature Selection and Conditional GAN-based Class Balancing. In 2021 IEEE 20th International Symposium on Network Computing and Applications (NCA) (pp. 1-4). IEEE.

Sui, D., Caverlee, J., &Rudesill, D. (2015). The deep web and the darknet. Washington DC: Publication of the Wilson Center.