Detect Darknet URL Based on Artificial Neural Network

Jie Xu*
Jiangsu Police Institute, Nanjing, Jiangsu, China
xujieip@163.com

Ao Ju
Jiangsu Police Institute, Nanjing, Jiangsu, China
2296381432@qq.com

ABSTRACT

Darknet is a network that transmits data on the Internet through anonymous network technology and protects the relationship between the two sides of communication from being leaked. Because the IP addresses of both sides of the communication cannot be traced on the darknet, the identity of the user cannot be determined. The darknet is used by criminals to engage in criminal activities. This paper studies the URL address of the darknet, proposes an algorithm for darknet URL recognition using artificial neural network. The algorithm transforms URL into a fixed length vector, and then uses it as a part of the input data of artificial neural network for learning and classification. Experiments show that the proposed algorithm has high accuracy, can accurately identify the darknet URL through multiple iterations under different attribute accuracy. Experimental results show that the proposed algorithm can achieve 99.3% detection accuracy.

CCS CONCEPTS

• **Networks**; • **Network properties**; • **Network manageability**;

KEYWORDS

Darknet, URL classification, Artificial neural network, Network security

ACM Reference Format:
Jie Xu* and Ao Ju. 2021. Detect Darknet URL Based on Artificial Neural Network. In *The 5th International Conference on Computer Science and Application Engineering (CSAE 2021), October 19–21, 2021, Sanya, China.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3487075.3487132

1 INTRODUCTION

With the popularity of the Internet, people spend a lot of time on the Internet. The increasing network activities include not only legal but also illegal activities. For example: network fraud, hacker attacks, phishing and so on. Network crime has become an important hidden danger threatening people's safety, and it is also an important part of the national crackdown on supervision. One of the key steps to crack down on network crime is to find out the criminals in the network, that is, to find out the IP addresses of illegal users, and then to find the criminals.

TCP / IP protocol is mainly used in Internet communication. In the pure TCP/IP protocol, each user's IP address is trackable. To avoid tracking, criminals do not communicate directly through IP address, but use anonymous communication network to hide their identity, such as Tor [1], I2P [2], Freenet [3] and so on. Anonymous communication network has a popular name - Darknet. Lots of content on Darknet are dangerous [4]. Many researches had been done on darknet, such as Tor based application classification [5], encrypted traffic classification [6-7], bridge discover [8], running environment [9] and so on. This paper will study the identification of darknet URL.

On the Internet, people can visit a website through its IP address or URL. Similar to web sites on the Internet, users need to know the address of a darknet site when they want to visit it. Because the darknet server will not disclose its IP address, the users have to access the darknet site through its URL. Identifying the address of the darknet and reducing its spread can effectively curb the development of the darknet. This paper studies how to use the method of artificial neural network to identify the address of darknet and Internet website.

The main contributions of this paper are as follows:

- An artificial neural network (ANN)-based method is proposed to detect the darknet URL;
- Collect the darknet URL from Internet
- Develop a prototype system to evaluate the performance of the darknet URL detection algorithm.

The arrangement of this paper is as follows. In the second section, this paper introduces the related work of darknet URL classification. The third section introduces the address recognition method based on artificial neural network; In section 4, the algorithm proposed in this paper is analyzed by experiments. Finally, the conclusion of this paper is summarized.

2 RELATED WORK

URL recognition is a hot topic in the research. Many algorithms for URL recognition on the Internet are proposed.

The simplest way to classify URLs is to use regular expressions for filtering. This method is applicable to URLs with characteristic strings. For example, the URL contains some sensitive strings, the characteristic domain name ".onion" in the dark network address, etc. However, in order to avoid detection, the feature string in the URL can be easily hidden by character replacement. At this time, we can use the method of machine learning to classify URLs.

The first step is to find suitable attributes of the URL. Then, classification algorithms, such as decision tree [10-11], support vector machine [12-13], Naive Bayesian [14], random forest [15] and so on, are used to learning relationship from these attributes.

Mohammed et al. [16] used clustering method to assign a ID to each URL based on lexical features and classified phishing site URL based on the clustering result.

Artificial neural network has a better classification performance and is widely used in URL classification. K. Shima et al. [17] uses a

CSAE 2021, October 19–21, 2021, Sanya, China

ACM ISBN 978-1-4503-8985-3/21/10…$15.00
https://doi.org/10.1145/3487075.3487132

512 dimensions vector to represent the original URL and normalizes the vector. Then, the deep neural network is used to detect the URL classification. The method proves the advantage of artificial neural network in URL classification and recognition. But the vector dimension adopted in this method is large, which increases the calculation time.

Afzal et al. [18] used both LSTM (Long Short-Term Memory) and k-means clustering algorithm to classify malicious URL based on semantic and lexical features.

In some situations, the URL's number of different classes is different. The imbalance data will affect the classification result. Seok et al. [19] acquired the features based on deep learning to solve the problem of URL class imbalance. They also learned the similarity of URL based on a triplet network structure.

In order to improve the computational efficiency of darknet URL, this paper will introduce a new method to extract vectors from the characters of URL, and use the artificial neural network method to identify the darknet URL.

3 DARKNET URL DETECTION

Artificial neural network (ANN) is an algorithm that can automatically learn the rules of data. The application of artificial neural network in URL recognition of darknet can obtain higher accuracy.

The first step to detection darknet URL is to acquire the apposite features of URL.

We can think of a URL as a string, and then extract string related properties from the URL. For example, the length of a string, the number of a special character, whether a substring appears, etc. In URL detection, we pay more attention to some special characters or substrings according to a priori knowledge. For example, the number of occurrences of HTTP or HTTPS can reflect whether there is redirection in the URL, the symbol "/" can reflect the number of virtual directories in the URL, and ". JSP" ". PHP "can reflect whether the website uses dynamic web pages, etc. These string attributes do not fully represent the characteristics of the URL, or the original URL cannot be restored from these attributes. The string rule contained in the darknet URL is not as obvious as the Internet URL. In order to detect the darknet URL, more URL information needs to be retained. Therefore, we use a vectorization method to represent URL features. For shorter URLs, this vector attribute has a one-to-one relationship with the URL.

A URL is made up of characters. If each character is represented by a number, the URL can become a vector composed of a set of numbers. The length of different URL is different, so the vector is different. In order to make the length of the URL vector the same, we truncate or fill in the length of the URL vector, and only keep the previous fixed length. Let L indicate the reserved character length. When the length of URL n is greater than L, only the front L characters are reserved. When n is less than L, the last (L − n) positions of the vector are filled with 0. We call such a vector a character numeric vector.

In order to make full use of the URL vector, when L > n, the last (L-n) positions can be used to represent the occurrence of a character combination. Each character combination is represented by a unique value different from the single character. When such a character combination appears in the URL, fill in the value at a certain position in the URL vector. Which character combination can be recorded and where they are placed can be obtained statistically. For example, for the character combination "HTTPS", assign a unique value v1 to it. When L > n, fill v1 in the nth position of the URL vector. Because the value of character combination is different from that of a single character, the original URL can still be restored according to the URL vector.

There are different ways to map each character in a URL to a numeric value. Users can randomly assign different values to each character, or sort the number of appearing times of each character in the URL according to the past URL statistical results, and set the value of each character according to the sorted order value. No matter which method is adopted, it is necessary to ensure that each character corresponds to only one numerical value, and each numerical value corresponds to only one character. In the following experiments, the second method is used to determine the corresponding value of each character, and the character combination information is not saved in the URL vector to simplify the calculation process.

When n is less than L, the length of the URL can be seen through the character value vector, but when the length of the URL is greater than n, the length of the URL cannot be seen from the character value vector. In order to keep the URL length information, we add a numerical value before the character array vector - the length of the URL. After adding the URL length, the character array vector is called the URL numeric vector. The length of the URL numeric vector is L + 1 ◦

It can be seen from the construction process of URL numeric vector that when n is less than or equal to L, the URL numeric vector contains all the information of the URL. When n is greater than L, the URL numeric vector can only partially recover the URL, and part of the URL information will be lost. In theory, the larger the URL numeric vector is, the more information the URL contains and the more accurate the classification is. But the longer it takes to compute. In the following experimental part, we will do experiments on different URL numeric vector lengths and artificial neural network parameters, compare and analyze their effects on URL recognition in darknet.

4 EXPERIMENTAL ANALYSIS

In order to detect the darknet address recognition algorithm based on artificial neural network proposed in this paper, 2500 Internet website URLs are collected from Alexa [20], and 2500 darknet addresses are collected from the darknet search engine for experiments.

When using artificial neural network for darknet URL recognition, each URL will be vectorized first. Then the URL vector is input into the artificial neural network for learning or classification.

The neural network used in this experiment consists of an input layer, a density layer composed of N nodes, a dropout layer and an output layer. The number of nodes N in the density layer after the input layer will affect the detection accuracy. This experiment compares and analyzes the accuracy of different N and L values for darknet URL detection.

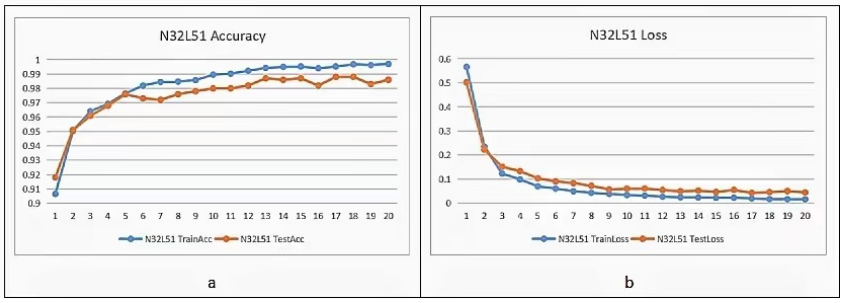

Figure 1: Comparison of Test Data Set and Training Data Set.

4.1 Comparison between Training Data Set and Test Data Set

In the experiment, we use 500 Internet URLs and 500 darknet URLs as the test data set, and the remaining 2000 Internet URLs and 2000 darknet URLs as the training data set. First, we compare the accuracy and loss values of training data set and test data set when N is 32 and L is 51, as shown in Figure 1

In the following figure, N represents the number of nodes in the second layer of neural network; L represents the number of attributes contained in each data; "Train" represents training data set; "Test" stands for test data set. The full name of ACC is accuracy, which means accuracy; "Loss" means loss rate. The abscissa represents the number of iterations, and each data set has 20 iterations; The ordinate represents the accuracy or loss rate corresponding to the icon title.

It can be seen from sub figure a in figure 1 that when the parameter is N32L51, the accuracy difference between the training data set and the test data set is small, and the accuracy of the training data set is slightly higher than that of the test data set. In the sub figure b of figure 1, the loss rate difference between the training data set and the test data set is small, and the loss rate of the training data set is slightly lower than that of the test data set. This is because the training results are based on the training data set, so the accuracy will be higher than the test data set.

4.2 Influence of Parameter L on Test Results

Then, in order to analyze the influence of parameter L on the detection results, we adopt the method of N unchanged and L changed to obtain the accuracy and loss values of the training data set and the test data set in turn, as shown in figure 2, figure 3, figure 4 and figure 5 respectively.

It can be seen from the experimental results that when the number of iterations increases, the accuracy of different L improves and tends to be stable. When N is constant, the greater L is, the higher the accuracy is. This is because the larger the L, the more URLs can completely save their information in the URL numeric vector.

4.3 Influence of Parameter N on Experimental Results

Then, in order to analyze the influence of parameter N on the experimental results, we adopt the method of L unchanged and N changed, and compare the experimental data when L is 11, 31, 51 and 71, as shown in figure 6, figure 7, figure 8 and figure 9 respectively.

Combining with the above analysis, we can get that the size of N has a great influence on the initial value of experimental results.

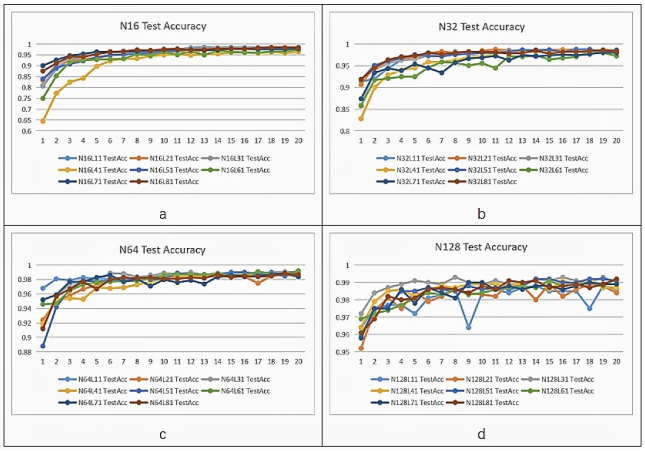

Figure 2: Comparison of Influence Results of Parameter L on Accuracy of Test Data Set.

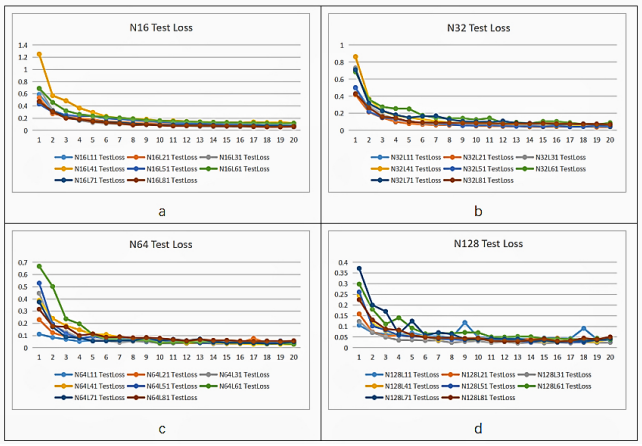

Figure 3: Comparison of Influence Results of Parameter L on Loss Value of Test Data Set.

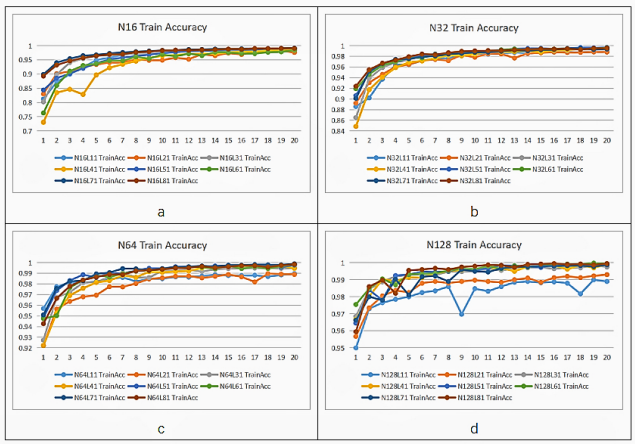

Figure 4: Comparison of the Influence of Parameter L on the Accuracy of Training Data Set.

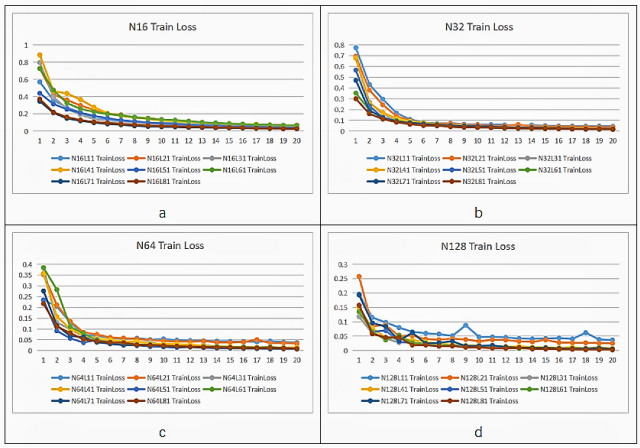

Figure 5: Comparison of the Influence of Parameter L on the loss Value of Training Data Set.

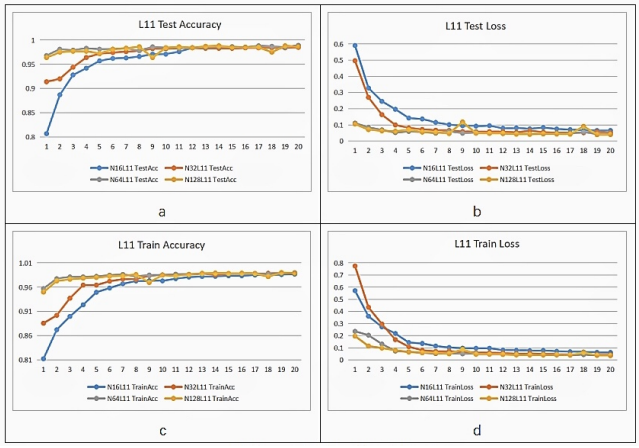

Figure 6: Comparison of the Influence of Parameter N on the Experimental Results when L is 11.

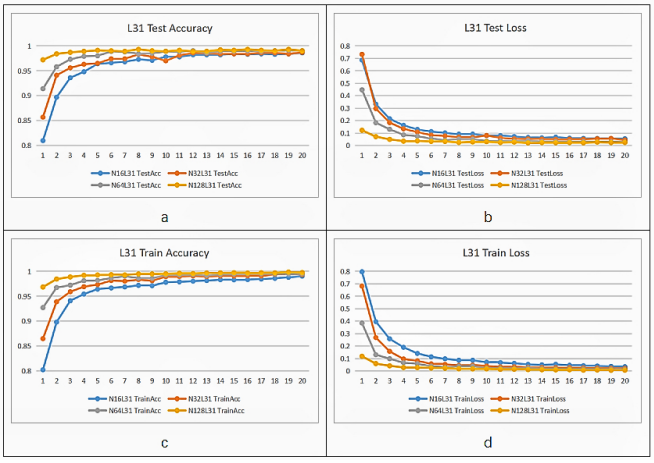

Figure 7: L Is the Comparison of the Influence of Parameter N on the Experimental Results when L is 31.

Larger N can obtain higher accuracy and lower loss value. This is because the more neural network nodes contain more network parameters, which can better learn the relationship of URL.

5 CONCLUSION

The darknet is used to carry on the network crime. Detecting darknet URL is one of the key steps to curb darknet crime. This paper extracts a numeric vector from URL and uses artificial neural network to detect darknet URL. The length of URL numeric vector can be adjusted flexibly, and the URL content can be partially recovered. At the same time, different length of URL numeric vector can get different classification effect. In this paper, the URL numeric vector is input into the artificial neural network for training and detection, and the relationship between the URL numeric vectors is automatically learned by using the artificial neural network. Experiments show that the detection accuracy is up to 99.3%. This paper only uses an artificial neural network model, does not analyze the impact of other network models, such as convolutional neural network, long short term memory, recurrent neural network and transform model, on the detection results. These works will be carried out in future research. In the future work, we will also study how to use neural network to detect phishing sites, darknets and malicious URLs in darknet.

ACKNOWLEDGMENTS

The research of this paper is supported by the science and technology research project of Jiangsu Provincial Public Security Department (2020KX007Z), the project of Jiangsu Provincial Department of Education (20KJB413002) and the "Jiangsu Police Institute high level talent introduction research start-up fund" (JSPIGKZ).

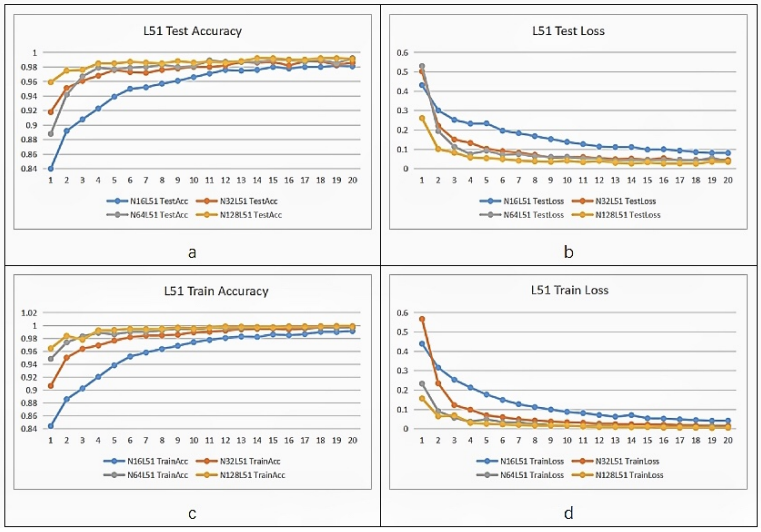

Figure 8: Comparison of the Influence of Parameter N on the Experimental Results when L is 51.

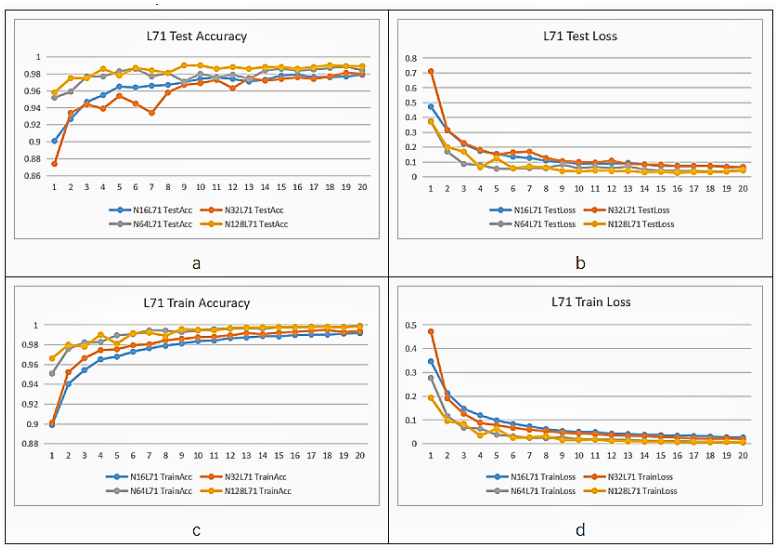

Figure 9: Comparison of the Influence of Parameter N on the Experimental Results when L is 71.

REFERENCES

[1] M. A. Irsyad Mohd Aminuddin, Z. F. Zaaba, A. Samsudin, N. B. Anuar Juma'at and S. Sukardi (2020). Analysis of the Paradigm on Tor Attack Studies. 2020 8th International Conference on Information Technology and Multimedia (ICIMU), pp. 126-131, doi: 10.1109/ICIMU49871.2020.9243607.

[2] H. Yin and Y. He (2019). I2P Anonymous Traffic Detection and Identification. 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), pp. 157-162, doi: 10.1109/ICACCS.2019.8728517.

[3] G. Tian, Z. Duan, T. Baumeister and Y. Dong (2017). A Traceback Attack on Freenet, in IEEE Transactions on Dependable and Secure Computing, vol. 14, no. 3, pp. 294-307, 1 May-June 2017, doi: 10.1109/TDSC.2015.2453983.

[4] H. Zhang and F. Zou (2020). A Survey of the Darknet and Dark Market Research. 2020 IEEE 6th International Conference on Computer and Communications (ICCC), pp. 1694-1705, doi: 10.1109/ICCC51575.2020.9345271.

[5] He, G., Yang, M., Luo, J., and Gu, X. (2015). A novel application classification attack against Tor. Concurrency Computat.: Pract. Exper., 27: 5640– 5661. doi: 10.1002/cpe.3593.

[6] F. Zhang, T. Shang and J. Liu (2020). Imbalanced Encrypted Traffic Classification Scheme Using Random Forest. 2020 International Conferences on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics), pp. 837-842, doi: 10.1109/iThings-GreenCom-CPSCom-SmartData-Cybermatics50389.2020.00142.

[7] Z. Cai, B. Jiang, Z. Lu, J. Liu and P. Ma (2019). isAnon: Flow-Based Anonymity Network Traffic Identification Using Extreme Gradient Boosting. 2019 International Joint Conference on Neural Networks (IJCNN), pp. 1-8, doi: 10.1109/IJCNN.2019.8851964.

[8] Z. Ling, J. Luo, W. Yu, M. Yang and X. Fu (2012). Extensive analysis and large-scale empirical evaluation of tor bridge discovery. 2012 Proceedings IEEE INFOCOM, pp. 2381-2389, doi: 10.1109/INFCOM.2012.6195627.

[9] S. Kim, J. Han, J. Ha, T. Kim and D. Han (2018). SGX-Tor: A Secure and Practical Tor Anonymity Network With SGX Enclaves, in IEEE/ACM Transactions on Networking, vol. 26, no. 5, pp. 2174-2187, Oct. 2018, doi: 10.1109/TNET.2018.2868054.

[10] B. Zhang (2021). Tactical Decision System of Table Tennis Match based on C4.5 Decision Tree. 2021 13th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), pp. 632-635, doi: 10.1109/ICMTMA52658.2021.00146.

[11] T. V. Olegario, R. G. Baldovino and N. T. Bugtai (2020). A Decision Tree-based Classification of Diseased Pine and Oak Trees Using Satellite Imagery. 2020 IEEE 12th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM), pp. 1-4, doi: 10.1109/HNICEM51456.2020.9400002.

[12] D. Yang, N. Chhatre, F. Campi and C. Menon (2016). Feasibility of Support Vector Machine gesture classification on a wearable embedded device. 2016 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), pp. 1-4, doi: 10.1109/CCECE.2016.7726800.

[13] L. Mohan, J. Pant, P. Suyal and A. Kumar (2020). Support Vector Machine Accuracy Improvement with Classification. 2020 12th International Conference on Computational Intelligence and Communication Networks (CICN), pp. 477-481, doi: 10.1109/CICN49253.2020.9242572.

[14] X. Wang and X. Sun (2016). An Improved Weighted Naive Bayesian Classification Algorithm Based on Multivariable Linear Regression Model. 2016 9th International Symposium on Computational Intelligence and Design (ISCID), pp. 219-222, doi: 10.1109/ISCID.2016.2059.

[15] K. RAMESH, M. A. BENNET, J. VEERAPPAN and P. RENJITH (2021). Performance Metric System for Malicious URL Data using Revised Random Forest Algorithm. 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), pp. 1188-1191, doi: 10.1109/ICCMC51019.2021.9418480.

[16] M. N. Feroz and S. Mengel (2015). Phishing URL Detection Using URL Ranking. 2015 IEEE International Congress on Big Data, pp. 635-638, doi: 10.1109/BigDataCongress.2015.97.

[17] K. Shima *et al.* (2018). Classification of URL bitstreams using bag of bytes. 2018 21st Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN), pp. 1-5, doi: 10.1109/ICIN.2018.8401597.

[18] Afzal, S., Asim, M., Javed, A.R. *et al.* (2021). URLdeepDetect: A Deep Learning Approach for Detecting Malicious URLs Using Semantic Vector Models. J Netw Syst Manage 29, 21 (2021). https://doi.org/10.1007/s10922-021-09587-8.

[19] S. -J. Bu and H. -J. Kim (2021). Learning Disentangled Representation of Web Address via Convolutional-Recurrent Triplet Network for Classifying Phishing URLs. 2021 International Conference on Electronics, Information, and Communication (ICEIC), pp. 1-4, doi: 10.1109/ICEIC51217.2021.9369758.

[20] Alexa, http://www.alexa.cn/