Classification of Malicious Websites Using Machine Learning Based on URL Characteristics

Muon Ha[1,2(✉)], Yulia Shichkina[1], Nhan Nguyen[2], and Thanh-Son Phan[2]

[1] St. Petersburg State Electrotechnical University, St. Petersburg, Russia
muon.ha@mail.ru, strange.y@mail.ru
[2] Telecommunications University, Nha Trang, Vietnam
ptson@tcu.edu.vn

Abstract. This article focuses on evaluating the efficiency of machine learning classification algorithms in detecting malicious websites based on their URL addresses. A highly reliable dataset of URL addresses is used to train the machine learning classification model. The results show that the Random Forest algorithm achieves a high accuracy of 95.68% in detecting malicious websites and provides an effective solution to this problem. At the same time, software for detecting and warning of malicious website in two forms is also implemented: web applications and browser extensions.

Keywords: Malicious website detection · URL · machine learning · cybersecurity

1 Introduction

These days, malicious websites are tools that cybercrimes often use to attack end-users. The tricks they often use include:

- Phishing websites: created by criminals with the goal of obtaining confidential information of the users such as usernames, passwords, credit card numbers, one-time authentication codes, and related banking account information, etc.
- Defacement website – a kind of cyberattack that destroys a website and changes the display interface of the website. In other words, when a user visits the address of this website, the interface of another website will be displayed. Typically, the displayed content is a message that the hacker wants to convey.
- A website containing malicious code (malware) is a trick that uses the website as a tool for spreading and installing malicious code on the victim's computer in order to steal information or gain control of the victim's computer for malicious purposes.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
O. Gervasi et al. (Eds.): ICCSA 2023 Workshops, LNCS 14112, pp. 317–327, 2023.
https://doi.org/10.1007/978-3-031-37129-5_26

According to statistics from the APWG [1] (Anti-Phishing Working Group), in the third quarter of 2022, there was a 47% increase in the number of detected phishing websites compared to the previous quarter.

Therefore, research on detecting and quickly preventing malicious websites has always been a particular interest to internet organizations and user communities. Nowadays, many researchers have also proposed numerous solutions to minimize cyberattacks via websites. However, cybercrimes always try to change their methods to evade detection. Thus, proposing an evolutionary solution is considered an effective method for detecting and blocking malicious websites (Fig. 1).

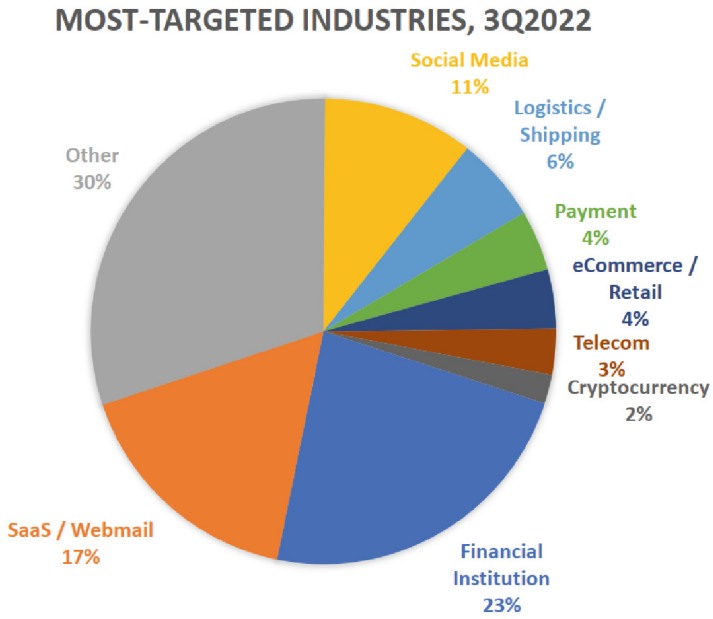


Fig. 1. The most targeted industries in the third quarter of 2022 [1]

The proposed approach is using machine learning methods to detect malicious websites based on URL characteristics. Different types of malicious websites, such as phishing, defacement, and web-spam, were taken into consideration in our work. To increase the accuracy of classification methods, we used a dataset updated from recent sources of URL data. In practice, we implemented a software tool for detecting and warning about malicious websites in the form of web applications and browser extensions.

The article is structured as follows. In the second section, a review of related works on the topic in the past few years is provided. In the third section, we developed a machine learning-based model for detecting malicious websites. In the fourth section, a description of the implemented software tool for detecting and warning about malicious websites is presented.

2 Related Works

Many studies have been conducted proposing different methods for detecting malicious URL addresses. One of them is maintaining a list of domain names or IP addresses of previously detected malicious websites [2–4]. This approach was used in [5], where the authors achieved 90% accuracy in classifying phishing websites after a three-week training using Google's phishing website blacklist. Or the Phishnet system presented in [6] uses a white list verification against a black list method. In this method, all URL addresses are checked against both the black and white lists respectively. Users are safe when the URL address is found in an official document; at the same time, users will be warned if the URL they visit is found in the black list. Another approach is presented as the use of shortened URL verification method, presented by the authors in [7–9], where if the attacker uses a URL shortening service, the user will receive a warning message about the risk of phishing attacks. DNS verification [10] of malicious URL addresses is also a widely used method.

Machine learning methods have been proven to quickly provide high accuracy in classifying objects based on available features. In studies [12–18], the use of machine learning methods in solving the problem of detecting malicious websites is presented. The rule-matching method, proposed in [19] for detecting malicious and safe URL addresses, uses 14 different features extracted from the URL address. The authors used the TF-IDF algorithm to search for high-frequency words in malicious URL addresses. Approximately 93.00% of malicious URL addresses were correctly identified by the Apriori algorithm in a dataset of 1400 URL addresses. In [20], the author presented the Phishing Alerting System (PHAS), capable of detecting and warning of all types of phishing emails, to help users make decisions. A dataset of email is used in this study and based on the extracted features, an accuracy of approximately 93.11% is proposed to be achieved using such machine learning methods as decision trees and random forests.

Although the above-mentioned works have achieved the main results, there are still some limitations, such as: firstly, the lack of flexibility in responding to new tricks of cybercriminals and the incomplete use of URL features for detecting malicious websites, and secondly, most of the existing works focus on detecting phishing websites and other types of malicious websites, such as defacement, web spam, are still limited. Thus, in this work, we propose the approach to the application of machine learning methods for the classification of multi-types of malicious websites based on their specific features.

3 Detection of Malicious Websites Using Machine Learning Methods

The malicious website detection model based on the proposed machine learning methods is divided into two stages, as shown in Fig. 2, as follows:

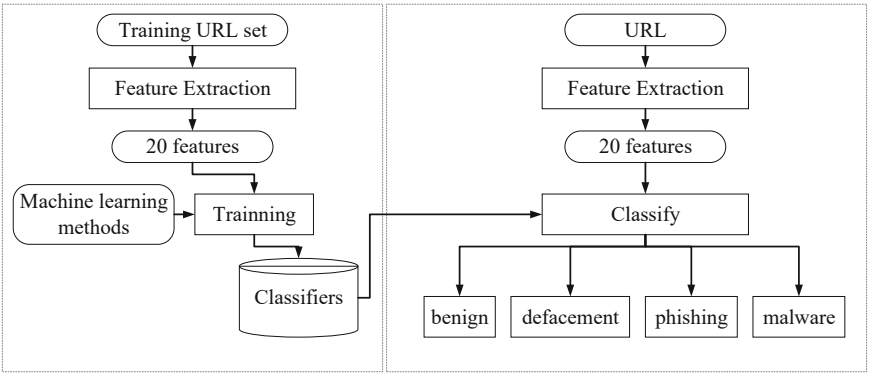


Fig. 2. Proposed model for classifying malicious websites

(a) Training stage: The training data file includes URLs of websites classified into 5 categories (labels): clean site, phishing, deception, spam web and malicious code website. The extracted URL features include 20 features. Using machine learning methods to train and provide classifiers.
(b) The detection stage: URLs are tracked and features are extracted using the trained classifier to identify malicious sites with URLs.

3.1 Feature Extraction URL

The basic structure of a URL consists of 3 main parts: protocol, host name, and path. Each URL has 24 features, but we only use 20 main features to judge whether a website is malicious or not. Some selected important features are shown in Table 1.

3.2 Machine Learning Classification Algorithms

Classification of objects based on feature extraction can use machine learning methods. There are several well-known machine learning methods such as:

- Decision Tree - A decision-making tool that uses a tree-like diagram or decision model and their possible outcomes, including the results of random events, resource expenditures, and benefits.
- Random Forest - A supervised learning algorithm that can solve both regression problems and classification tasks. Random forests generate decision trees on randomly selected data samples, predict for each tree and select the best decision by voting.
- K-Nearest Neighbors (K-NN) - one of the simplest supervised learning algorithms in machine learning. During training, this algorithm doesn't learn anything from the training data (which is also the reason why this algorithm is classified as lazy learning), all computations are done when it needs to predict the result of new data.

Table 1. Important features of URL

Features	Description
Having ip address	If the URL contains an IP address, it can be determined that the website is malicious. For example: http://192.70.12.33/phising.html
URL length	The average length of a URL address is 54 characters. If the URL address is longer than 54 characters, the website may contain malicious content
Shortening Service	The use of URL shortening service may be aimed at hiding the real address of the website and is also an indication that the website is not trustworthy
httpSecure	Cybercriminals can deceive users by attaching the "https" string to the URL address. For example: http://https-www-facebook.soft-hair.com/
Digit count	If a URL address contains too many characters and numbers, it is also a sign of a malicious website
Abnormal URL	This information can be extracted from the WHOIS database. For a legitimate website, the identifier is typically part of the URL address

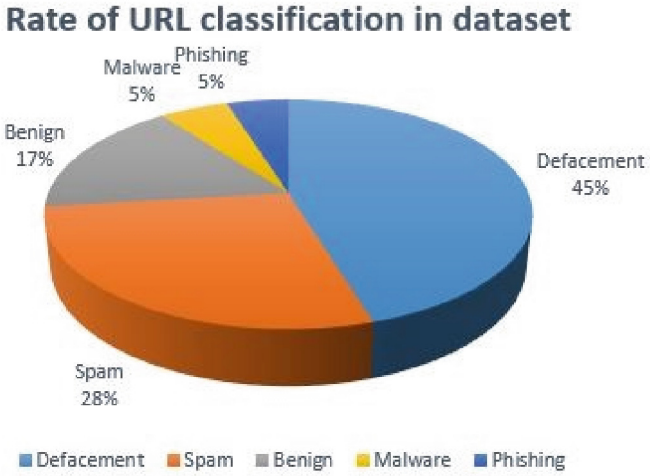
- Adaptive Boost (AdaBoost) - a powerful learning algorithm that accelerates the creation of a strong classifier by selecting good features from a family of weak classifiers and linearly combining them using weights.

In this work, we will apply the aforementioned algorithms in succession for website classification and malicious site detection, thereby evaluating the performance of each algorithm to select the best algorithm for the implementation process of the malicious web site detection application for various popular web browsers.

3.3 Dataset

In this study, we have collected a sample of data from public and regularly updated data sources. Our sample dataset includes 213,345 URLs. They are divided into 5 groups and labeled as: Benign - 0; Defacement - 1; Phishing - 2; Malware - 3 and Spam - 4. The number of URLs in the dataset is distributed as shown in Fig. 3.

In this study, malicious URL addresses were collected from sources such as OpenPhish [22], Phishtank [23], Zone-H [24], and WEBSPAM-UK2007 [25]. For each data source, we collect the URL as a character string and its corresponding label, then merge them into a large dataset for training. The data preprocessing process includes feature extraction, removing duplicated URL addresses, and filtering out URL addresses that do not have the necessary features.



Label	Defacement	Spam	Benign	Malware	Phishing
Count	96657	59279	35378	11566	10465

Fig. 3. Count and ratio of groups of websites in the dataset

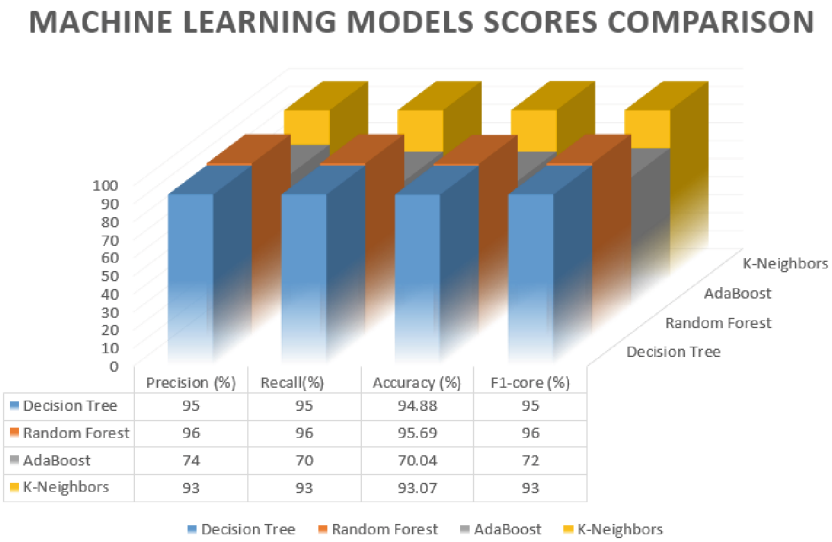
4 Results and Discussion

With the training data set, we used the most popular machine learning algorithms for 10-fold cross-validation to determine the model's performance. The performance metrics used to evaluate the models are Precision, Recall, accuracy, and F1-Score. Compared to some similar studies [12, 26], our model shows a significant effectiveness as demonstrated by the values mentioned above.

Based on the results in Fig. 4, with accuracy and F1-score values of 94.8388% and 95% respectively, the Random Forest method gives the best efficiency. Figure 5 shows its normalized confusion matrix, from which it can be seen that the model's recognition accuracy with data labels ranges from 4.3% to 56.37%. Meanwhile, the false recognition rate is relatively low, with the highest being 1.59%. This demonstrates the high diagnostic efficiency of the model. Therefore, we choose the Random Forest method in the implementation step of the application for detecting and warning against malicious websites based on URL characteristics.

The final step of the research involves implementing software to detect and prevent malicious website using the Random Forest method. To increase the program's efficiency, we combine the Random Forest method and the blacklist check method. The input URL address will first be checked in the database of malicious URL addresses, and then it will be verified using the machine learning model. Figure 6 shows the process of the malicious website detection and prevention program.

Our software has been implemented in two forms: a web application and a browser extension. The web application is built using programming languages JavaScript and Python, and its interface shows in Fig. 7.



	Precision (%)	Recall(%)	Accuracy (%)	F1-core (%)
Decision Tree	95	95	94.88	95
Random Forest	96	96	95.69	96
AdaBoost	74	70	70.04	72
K-Neighbors	93	93	93.07	93

Fig. 4. Performance metrics for a malicious dataset URL with different classifiers

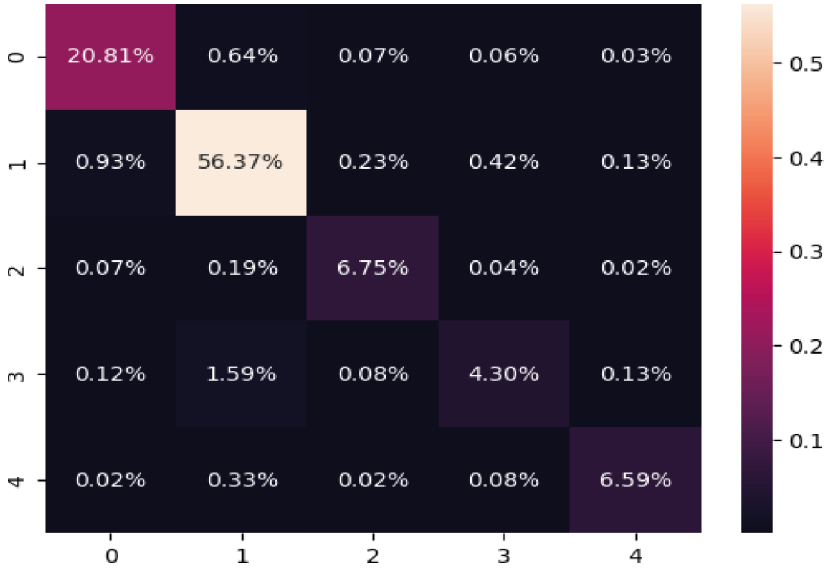


Fig. 5. Confusion matrix of the random forest method

The browser extension is a small part of the software used for configuring the web browser. The browser extension for detecting malicious websites is built using HTML, JavaScript, and CSS, while the Random Forest model was implemented using the Python programming language. Figures 8 shows the results of the browser extension check for a malicious website.

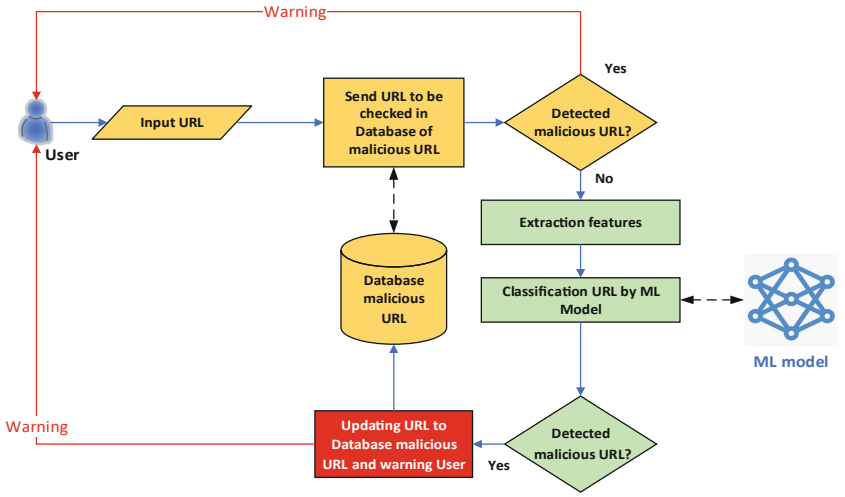


Fig. 6. The process of the developed program for detecting and preventing a malicious website

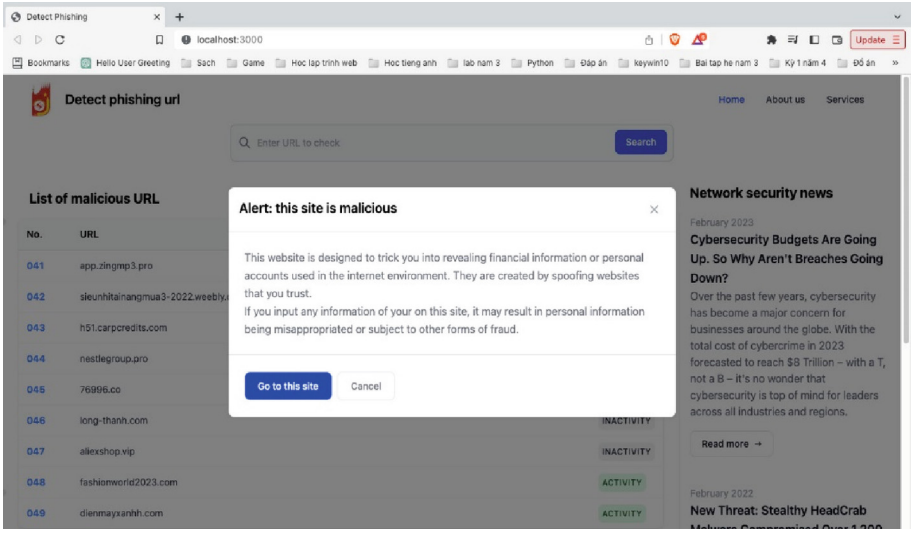


Fig. 7. Web application interface for detecting and warning a malicious website.

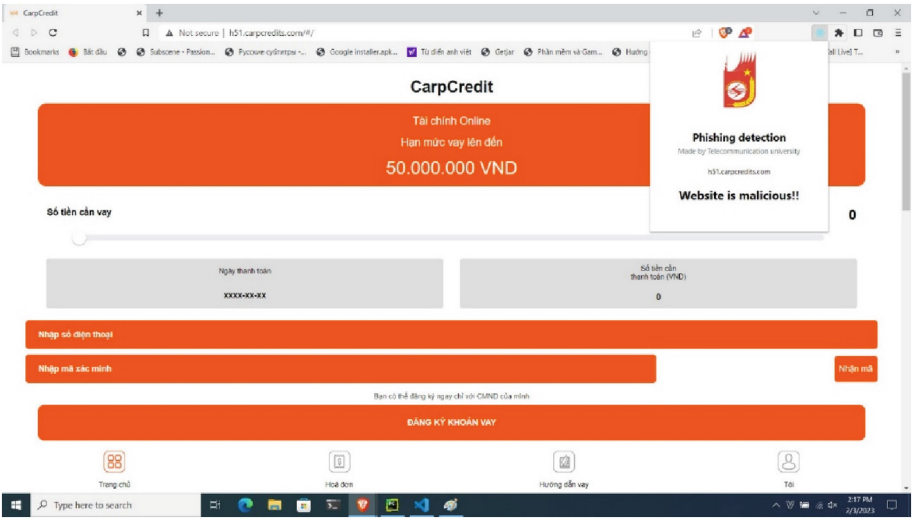


Fig. 8. Malicious Website Detection Results of Developed Browser Extension Application

5 Conclusion

Thus, this article presents an approach to using machine learning methods for classifying malicious websites based on URL characteristics. The results of the tests showed that the Random Forest method provides the best efficiency. This work also created software in the form of a web application and a browser extension for detecting and warning against malicious websites using random forest methods and blacklists. In the future, we will continue to explore different feature sets and use larger data sets to more accurately and efficiently detect malicious websites.

Acknowledgment. This work was supported by the Ministry of Science and Higher Education of the Russian Federation by the Agreement № 075-15-2022-291 on the provision of a grant in the form of subsidies from the federal budget for the implementation of state support for the establishment and development of the world-class scientific center «Pavlov center Integrative physiology for medicine, high-tech healthcare, and stress-resilience technologies».

References

1. Phishing activity trends reports. APWG. https://apwg.org/trendsreports/. Accessed 12 Feb 2023
2. Akiyama, M., Yagi, T., Hariu, T.: Improved blacklisting: Inspecting the structural neighborhood of malicious urls. IT Prof. **15**, 50–56 (2013). https://doi.org https://doi.org/10.1109/MITP.2012.118
3. Hong, J., Kim, T., Liu, J., Park, N., Kim, S.-W.: Phishing URL detection with lexical features and blacklisted domains. In: Jajodia, S., Cybenko, G., Subrahmanian, V.S., Swarup, V., Wang, C., Wellman, M. (eds.) Adaptive Autonomous Secure Cyber Systems, pp. 253–267. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-33432-1_12

4. Sun, B., Akiyama, M., Yagi, T., Hatada, M., Mori, T.: Automating URL blacklist generation with similarity search approach. IEICE Tran. Inf. Syst. **E99**(D), 873–882 (2016). https://doi.org/10.1587/transinf.2015ICP0027
5. Whittaker, C., Ryner, B., Nazif, M.: Large-scale automatic classification of phishing pages. In: Network and Distributed System Security Symposium (2010)
6. Prakash, P., Kumar, K., Kompella, R., Gupta, M.: PhishNet: predictive blacklisting to detect phishing attacks. In: 2010 Proceedings IEEE INFOCOM, San Diego, CA, USA, pp. 1–5 (2010). https://doi.org/10.1109/INFCOM.2010.5462216
7. Joshi, Y., Saklikar, S., Das., D, Saha, S.: PhishGuard: a browser plug-in for protection from phishing. In: 2008 2nd International Conference on Internet Multimedia Services Architecture and Applications, Bangalore, India, pp. 1–6 (2008). https://doi.org/10.1109/IMSAA.2008.4753929
8. Nikiforakis, N., et al.: Stranger danger: exploring the ecosystem of ad-based URL shortening services. In: WWW 2014 - Proceedings of the 23rd International Conference on World Wide Web, pp. 51–62 (2014). https://doi.org/10.1145/2566486.2567983
9. Gupta, N., Aggarwal, A., Kumaraguru, P.: bit.ly/malicious: deep dive into short url based e-crime detection. In: 2014 APWG Symposium on Electronic Crime Research (eCrime), Birmingham, AL, USA, 2014, pp. 14–24 (2014). https://doi.org/10.1109/ECRIME.2014.6963161
10. Singh, C., Meenu: Phishing website detection based on machine learning: a survey. In: 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2020, pp. 398–404 (2020). https://doi.org/10.1109/ICACCS48705.2020.9074400
11. Chou, N., Ledesma, R., Teraguchi, Y., Mitchell, J.: Client-side defense against web-based identity theft. In: Proceedings of the Network and Distributed System Security Symposium, San Diego, California, USA (2004)
12. Sharma, A., Thakral, A.: Malicious URL classification using machine learning algorithms and comparative analysis. In: Raju, K.S., Govardhan, A., Rani, B.P., Sridevi, R., Murty, M.R. (eds.) Proceedings of the Third International Conference on Computational Intelligence and Informatics. AISC, vol. 1090, pp. 791–799. Springer, Singapore (2020). https://doi.org/10.1007/978-981-15-1480-7_73
13. Wu, T., Xi, Y., Wang, M., Zhao, Z.: Classification of malicious urls by CNN model based on genetic algorithm. Appl. Sci. **12**(23), 12030 (2022). https://doi.org/10.3390/app122312030
14. Dutta, A.K.: Detecting phishing websites using machine learning technique. PLoS ONE **16**(10), e0258361 (2021). https://doi.org/10.1371/journal.pone.0258361
15. Monther, A., Rami, A.: MALURLS: a lightweight malicious website classification based on URL features. J. Emerg. Technol. Web Intell. **4**(2), 128–133 (2012). https://doi.org/10.4304/jetwi.4.2.128-133
16. Patgiri, R., Katari, H., Kumar, R., Sharma, D.: Empirical study on malicious URL detection using machine learning. In: Fahrnberger, G., Gopinathan, S., Parida, L. (eds.) ICDCIT 2019. LNCS, vol. 11319, pp. 380–388. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-05366-6_31
17. Alsaedi, M., Ghaleb, F.A., Saeed, F., Ahmad, J., Alasli, M.: Cyber threat intelligence-based malicious URL detection model using ensemble learning. Sensors. **22**(9), 3373 (2022). https://doi.org/10.3390/s22093373
18. Alsmadi, M., Alsmadi, I., Wahsheh, H.A.M.: URL links malicious classification towards autonomous threat detection systems. In: Al-Emran, M., Al-Sharafi, M.A., Al-Kabi, M.N., Shaalan, K. (eds.) ICETIS 2021. LNNS, vol. 322, pp. 497–506. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-85990-9_40
19. Jeeva, S.C., Rajsingh, E.B.: Intelligent phishing URL detection using association rule mining. HCIS **6**(1), 1–19 (2016). https://doi.org/10.1186/s13673-016-0064-3

20. Fon, K.H., Lashkari, A., Ghorbani, A.: A phishing email detection approach using machine learning techniques. Int. J. Comput. Inf. Eng. **44**, 2340 (2017)
21. Firake, S.M., Soni, P., Meshram, B.B.: Tool for prevention and detection of phishing e-mail attacks. In: Wyld, D.C., Wozniak, M., Chaki, N., Meghanathan, N., Nagamalai, D. (eds.) CNSA 2011. CCIS, vol. 196, pp. 78–88. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-22540-6_8
22. Phishing intelligence. OpenPhish. https://openphish.com/. Accessed 12 Feb 2023
23. Join the fight against phishing. PhishTank. https://phishtank.org/. Accessed 12 Feb 2023
24. Zone-H. Zone. http://www.zone-h.org/archive. Accessed 12 Feb 2023
25. Webspam-UK2007 (current dataset). In: WEBSPAM-UK2007. https://chato.cl/webspam/datasets/uk2007/. Accessed 12 Feb 2023
26. Das, A., Das, A., Datta, A., Si, S. and Barman, S.: Deep approaches on malicious URL classification, In: 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT*)*, Kharagpur, India, 2020, pp. 1–6 (2020). https://doi.org/10.1109/ICCCNT49239.2020.9225338