International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

An Identification and Analysis of Harmful URLs through the Application of Machine Learning Techniques

Dr. Swagat M. Karve[1], Dr. Shital Kakad [2], Dr. Swapnaja Amol Ubale [3], Dr. Ashwini B. Gavali[4], Prof. Sonali B. Gavali [5], Dr. Shrinivas T. Shirkande[6]

Submitted: 22/12/2023 **Revised**: 28/01/2024 **Accepted**: 08/02/2024

Abstract- Malicious URLs pose a significant cyber security threat, posing risks to user security and causing substantial financial losses. Traditional detection methods relying on blacklists are limited in addressing rapidly evolving threats. As a response, machine learning approaches have gained popularity for enhancing the efficiency of malicious URL detection. This paper presents a detailed analysis, offering a structured insight into various aspects and formally defining the machine learning task of identifying malicious URLs. It delves into feature representation, algorithm design. The objective of survey is to provide a detailed analysis of harmful URLS not only to researchers but to cyber security experts.

Keywords- *Malicious URLs, Cyber security, Malware, Phishing, Machine learning, Deep learning.*

1. Introduction- The contemporary digital era witnesses millions of individuals engaging globally primarily through social networking platforms, raising significant concerns regarding privacy and security [1]. The prevalence of Internet applications has led to a rise in network attacks, employing tactics like malware distribution, spam, and phishing to generate profits. Unfortunately, as technology advances, so do the methods for exploiting users, encompassing activities such as creating counterfeit websites, financial scams, and the installation of harmful software [2]. There might be misleading information in emails through for users through various ways like job links, gift winner site; social media friends etc. potentially leading to unwitting access of harmful content [3]. Malicious URLs are employed to deceive users into clicking on them, compromising security of system or cracking information privacy through granting unauthorized access [4].

A uniform resource locator (URL) is a location of website indicating where a data or information is kept on the internet and that could be entered into a browser to access a specific website. For instance, "https://www.Google.com" is an example of a URL.

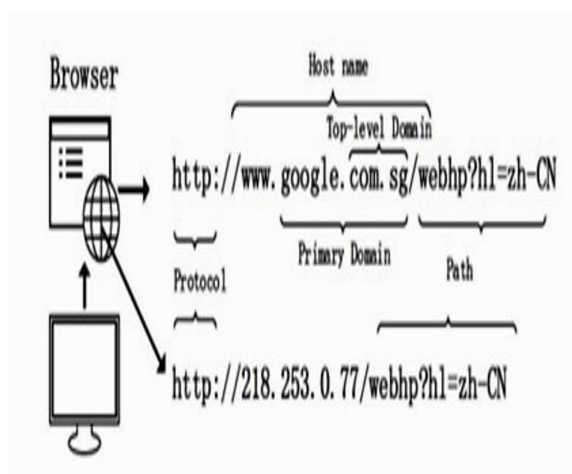

Fig 1. Example of URL

On the flip side, a malicious URL refers to a web address crafted with the intent to harm or exploit users. These URLs commonly lead to websites designed for

[1]Assistant Professor, S. B. Patil College of Engineering, Indapur (MH), India, swagatkarve@gmail.com
[2]Assistant Professor, Information Technology Department, Marathwada Mitra Mandal College of Engineering, Pune, shitalkakad2604@gmail.com
[3]Associate Professor, Information Technology Department, Marathwada Mitra Mandal College of Engineering, Pune, swapnaja.b.more@gmail.com
[4]Assistant Professor, S. B. Patil College of Engineering, Indapur (MH), India, dnyane.ash@gmail.com
[5]Asssistant Professor, Dr.D.Y.Patil, Institute of Technology, Pimpri,Pune, sonygavali1991@gmail.com
[6]Assistant Professor, S.B.Patil College of Engineering, Indapur (MH), India, shri.shirkande8@gmail.com

distributing malware, extracting sensitive information, or executing other harmful activities. Clicking on such URLs can lead to cyberattacks, data breaches, and security vulnerabilities. The deceptive nature of these URLs, often designed to mimic trustworthy sites, poses a significant threat to unsuspecting users. According to findings from Kaspersky [5], web security software detected 173 million dangerous URLs in the year of 2020. And as per the report 66.07% of these suspicious URLs were linked to the 20 most recently identified harmful applications.

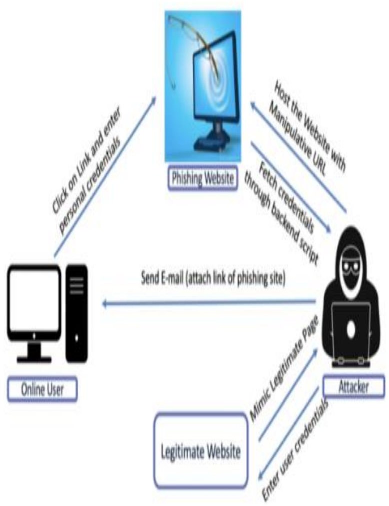

Fig 2. Data Stealing Procedure

Malignant URLs frequently serve as conduits for ransomware, phishing, malware dissemination, and various intrusions. The identification and blocking of such URLs play a crucial role in safeguarding users and systems from these forms of threats. Cyber attackers leverage malicious URLs to execute diverse attacks, spanning spam, phishing, malware distribution, and defacement. Cyberattacks typically unfold when unsuspecting visitors click on deceptive URLs. The misuse of URLs, diverging from legitimate online resources, jeopardizes data integrity, confidentiality, and accessibility [4].

To detect malicious URLs, a diverse range of methods must be employed, including traditional approaches. URL Phishing utilizes 2 primary strategies: 1. Blacklisting 2. Whitelisting, supplemented by sophisticated techniques. Intelligent methods involve the manual or statistical selection of discriminative features, essential for improving accuracy in classification and overall effectiveness [6].

2. Related Work

Here in section 2 there is study of URL features and different types of possible URL attack .

2.1 Features of URL

2.1.1 Lexical Features:

These features encompass various characteristics such as length, frequency and the prevalence of high-frequency words [7]. In the context of URLs, lexical features extend to aspects like URL length, the count of special characters, the ratio of digits to letters, the proportion of uppercase to lowercase characters, and the presence of single characters. These static lexical features are derived directly from the URL string [8]. Visual and textual attributes of a URL, including factors like length, domain length, special characters, and digits, fall under lexical features. These features offer statistical insights into the structural aspects of the URL, contributing to the evaluation of potential threats [4].

A list of lexical features commonly used in the analysis of URLs includes:

- **URL Length:** URL characters count.
- **Special Characters count:** Count of non-alphanumeric characters such as hyphens, underscores, and other symbols.
- **Digit to Letter Ratio:** The ratio of numeric characters to alphabetical characters in the URL.
- **Uppercase and Lowercase Ratio:** The proportion of uppercase letters to lowercase letters in the URL.
- **Presence of Single Characters:** Indication of whether the URL contains single characters.
- **Domain Length:** Domain Part length in URL.
- **TLD:** The last part of the domain, indicating the website's type or purpose (e.g., .com, .org, .edu).
- **Use of Hyphens in Domain:** Presence of hyphens within the domain part of the URL.
- **Use of Subdomains:** The presence and count of subdomains in the URL.

- **Character Frequency Distribution:** Analysis of the frequency distribution of individual URL characters.

- **Word Length:** Words length within the URL.
- **Word Frequency:** Words frequency within the URL.

These lexical features are utilized to extract statistical information from the URL string, aiding in the identification and differentiation between malicious and benign URLs.

2.1.2 Content Features:

Web address (URL) functions as a unique identifier for locating resources on the Internet [9]. Elements within the URL string content are often referred as URL features. These features can provide insights into the nature of the URL and its potential threat level. They play a crucial role in identifying problematic elements or patterns within the URL. The HTML structure of a webpage is also analyzed to extract webpage content features (CONTs), including HTML tags, iframes, zero-size iframes, lines, and hyperlinks. The process said is designed to scrutinize the webpage structure and detect suspicious code [10].

A URL content features, which includes elements providing information about the URL's nature and potential threat level, consists of:

- **Keywords:** Specific words or terms within the URL string.
- **Patterns:** Recognizable sequences or arrangements of characters in the URL.
- **Encoded Material:** Content that has been encoded or encrypted within the URL.
- **HTML Tags:** Elements within the HTML structure of a webpage, denoting various types of content or formatting.
- **Iframes:** HTML elements used to embed another document or webpage within the current one.
- **Zero-Size Iframes:** Iframes with no visible dimensions on the webpage.
- **Lines:** Quantification of lines within the HTML structure of the webpage.
- **Hyperlinks:** Links within the URL that direct to other resources or pages.
- **Native JavaScript Functions:** Analysis of specific JavaScript functions within the webpage.

2.1.3 Network Features:

These features within a URL encompass details pertaining to the online infrastructure, incorporating factors like the domain's age, the reputation of the associated IP address, and the geographical location of the server. Examination of WHOIS records provides valuable insights into domain ownership, contributing to the assessment of a URL's trustworthiness and potential risk. These features play a crucial role in identifying potentially harmful online resources. The network features of a URL include characteristics related to DNS, network, and host aspects [4].

List of network features in a URL, providing insights into the online infrastructure and aiding in the assessment of potential threats, includes:

- **Domain Age:** The length of time since the creation of the domain associated with the URL.
- **IP Address Reputation**: The reputation or historical behavior of the IP address linked to the URL.
- **Server Geographical Location:** The physical location of the server hosting the website.
- **WHOIS Records:** Information extracted from WHOIS records, including details about domain ownership and registration.
- **Resolved IP Count:** The number of resolved IP addresses associated with the URL.
- **Latency:** The delayed time between a request and response, indicating the responsiveness of the server.
- **Redirection Count**: The number of times the URL redirects to another location.
- **Domain Lookup Time:** The time it takes to look up the domain associated with the URL.
- **DNS Queries**: The number of queries made to the Domain Name System (DNS) for the URL.
- **Connection Speed:** The speed at which a connection to the URL's server is established.
- **Open Ports:** Identification of open ports on the server associated with the URL.

2.2 URL Attack Types

Malicious URLs can compromise data integrity, confidentiality, and internet availability [4]. Various attacking techniques through URLs are detailed below:

2.2.1. Spam URL Attacks:

These attacks involve the use of email URLs, forums, or websites to disseminate unsolicited or undesirable content, often with false or commercial intentions. Hackers create webpages designed to deceive web browsers into perceiving them as legitimate, leading to three main objectives in transmitted emails:

- Imitating well-known websites to acquire user credentials.
- Infecting the user's PC.
- Distributing spam [11].

2.2.2. Malware Attacks:

The primary goal of malware attacks through URLs is to steal sensitive user information or gain unauthorized access to systems. Malicious URLs attack occurs when users unknowingly download malware after visiting deceptive websites, posing significant harm to their computers and cracks the privacy[12].

2.2.3. Phishing URL Attacks:

Here login credentials are stolen without knowing users through various URLS. These harmful URLs can be distributed in both public and private environments. Without measures to limit or eliminate these URLs, attackers can easily retrieve user credentials, which may scarify the fund or privacy loss.

2.2.4. Defacement URL Attacks:

Defacement URL attacks involve unauthorized alterations to a website's appearance or content. Motivations for these attacks can vary, including stump speech, showcasing treating, or expressing personal antagonism. Consequences may include damage to reputation of corporations, suspicion etc [4]. Hacktivists often use website defacement as a tool to promote socio-political and ideological goals, with instances targeting specific organizations, governments, or companies [15,16].

3. Techniques for malicious URL detection

Various methods exist for detecting fraudulent URLs, encompassing traditional techniques, machine learning approaches, and more. Few listed ways are as follows for identifying malicious URLs:

3.1 Blacklists:

Blacklists comprise a compilation of known harmful URLs, and access to these URLs is prohibited if they match any entries in the list. Blacklisting involves creating a list of suspicious websites and blocking them to prevent access [6]. However, this method has limitations as phishing URLs may undergo slight changes, making it challenging for traditional spam filters to identify them. Additionally, blacklists are less effective for newly added or altered URLs, and lexical comparisons can be resource-intensive and incompatible with real-time streaming [2,13,17].

3.2 Whitelists:

Whitelists consist of normal URL addresses, and to determine the legitimacy of a URL, one can check if it is included in the whitelist [18,19,20].

3.3 Heuristic Approach:

It identifies zero-hour phishing threats by recognizing features observed in actual phishing attacks. While this approach provides versatile protection against evolving threats, improvements are needed to reduce false positives [14,20]. Some researchers, such as C. Seifert et al. [21], employ a heuristic approach alongside blacklists, dynamically creating signatures for new URLs targeting unique elements of phishing sites. Paper [22], propose a heuristic-based method analyzing features specific to phishing sites, effectively identifying and mitigating potential attacks. As per authors of [23] there is use of heuristic method for classification of URLs into harmful and safe types.

3.4. Machine Learning Approach:

To overcome the limitations of blacklists and heuristics, researchers have turned to machine learning for more effective detection. Before applying any algorithm, feature extraction is crucial, involving characteristics of the URL. Two feature extraction methods include tokenization and vectorization, and lexical feature selection. Afterward, machine learning or hybrid approaches can be implemented using various classifiers such as SVM, RF, NB, LSTM, LR, GB, DT, and deep learning methods. There is detection of malicious URLs using four machine learning algorithms, with RF achieving the highest accuracy of 92.18% as per paper [26]. Other studies have explored character-aware language models like LSTM, CNN, and CharacterBERT, achieving success in URL-based detection models [27,28]. The paper [30] proposed a DDQN classifier and Deep reinforcement algorithm for web phishing classification, demonstrating superior accuracy.

4. Datasets Used

Researchers use various datasets for training detection models, ensuring real-world relevance [5]. The ISCX-URL-2016 dataset, among others, has been utilized for classifying URLs into categories.

5. Malicious URL Detection Using ML

Table-1 provides an overview of previous URL detection using machine learning methods, showcasing the evolving landscape of research in this field.

6. Challenges and Future Directions

Over the past decade, significant strides have been made in using machine learning for identifying harmful URLs. However, some critical challenges persist. One key issue in the reviewed literature is the size of the data, suggesting the need for ample samples with a balanced ratio of normal to malicious URLs. Balancing strategies can enhance detection accuracy while maintaining an adequate sample size. Detection challenges arise when machine learning models lack historical data, making it difficult to identify emerging threats or zero-day attacks. Adaptable models capable of swiftly adjusting to evolving trends are crucial. Malicious actors often employ techniques to regularly modify URL structures, necessitating machine learning models that can withstand such polymorphic attacks. As URLs may contain sensitive information, ensuring data confidentiality while using URL data for training is essential.

7. Conclusion

As a concluding remark this paper emphasizes the significant role of machine learning in cyber security for detecting malicious URLs. The section 1 gives the brief introduction about malicious URLs including the data stealing procedure. In section 2 there is discussion about features of URLs and different types of attacks. Based on that in section 3 there is coverage of various techniques for detecting malicious URLs. From the researches point of view which data sets need to be considered are discussed in brief in section 4. Section 5 contains about how Machine Learning techniques can be implemented with the tabular form. Challenges and Future Directions are discussed in section 6.

Table 1. Study of malicious URLs detection based on machine learning

Reference	Year	URL classification	Classifier/Method	Result
[1]	2021	Malicious, Phishing and benign URLs	XGBoost, CS-XGBoost, SMOTE+XGBoost FNN (Fuzzy Neural Networks)	99.8%
[3]	2021	Malicious website	LR, DT	97.5% 85%
[5]	2021	Malicious and benign URLs	combining the attention-based bidirectional independent recurrent network (Bi-IndRNN) and capsule network (CapsNet)	99.89%
[6]	2020	Malicious and safe URLs	RF, Single class SVM	86.24% 96-97%
[8]	2019	Malicious and benign URLs	Random forest, Gradient boost, AdaBoost, Logistic regression, Naïve Bayes	92%, 90%, 90%, 87%, 70%
[11]	2020	Malicious and benign URLs	RF, fast.ai, Keras-TensorFlow(deep learning framework)	96.99% 97.55% 93.81%
[17]	2022	Malicious or benign URLs	LR, MLP neural network	93.26% 96.35%
[18]	2017	Malicious or benign URLs	Multi-layer filtering model, Simple NB, Simple DT, Simple SVM	79.55% 77.30% 79.35% 76.80%
[25]	2022	Malicious or benign URLs	Logistic regression, SVM, RF,	92.80% 97.32%

			GB, Bagging	97.35% 96.27% 97.35%
[26]	2023	Malicious and safe URLs	SVM, RF, DT, KNNs	91.25% 92.18% 90.18% 86.64%
[28]	2022	Malicious and benign URLs	CNN LSTM NB RF	95.13% 95.14% 96.01% 95.15%
[29]	2021	Malicious and benign URLs	XGBoost, CS-XGBoost, SMOTE+XGBoost	97.83% 99.05% 98.43%
[30]	2023	Malicious URLs using unbalanced classification	a double deep Q-Network (DDQN)-based classifier, Deep Reinforcement Learning	93.4%
[31]	2023	Phishing, benign, defacement and malware	RF, LightGBM, XGBoost	96.6% 95.6% 93.2%
[32]	2020	Malicious and benign URLs	RF, SVM	99.77% 93.39%
[33]	2019	Good and bad URLs	RF SVM	92.38% 87.93%
[34]	2023	Malicious website	MM-ConvBERT-LMS	98.72%
[35]	2023	Phishing URLs through parallel processing	NB, CNN, RF, LSTM	96%
[36]	2022	Malicious and benign URLs	RF	96%
[37]	2019	Phishing and benign URLs	CNN	86.63%

[38]	2022	Malware	Logistic regression, SVM, ELM, ANN	89.99% 96.49% 98.17% 97.20%
[39]	2022	Malicious and benign URLs	MLP	99.62%
[40]	2022	Phishing website	BERT, NLP, Deep CNN	96.66%
[41]	2023	Phishing and benign URLs	RF, GB, XGB	97.44% 98.27% 98.21%
[42]	2021	Malicious URLs using data mining approach	CBA(classification based on association)	91.30%
[43]	2022	Phishing and legitimate URLs	LSTM, Bi-LSTM, GRU	97% 99% 97.5%
[44]	2021	Threats and alerts on network log by pfSense	1D-CNN, LSTM	~ 99%
[45]	2022	Phishing URLs using homoglyph attack detection	RF	99.8%
[46]	2017	Intrusion detection	eXpose neural network that uses deep learning method	97-99%
[47]	2020	Fraudulent URLs which work in the Splunk platform	RF SVM	Precision: 85%, Recall:87 % Precision: 90%, Recall:88 %
[48]	2012	Suspicious URLs detection for twitter	Logistic regression, support vector classification (SVC)	87.67% 86%

[49]	2022	Malicious and benign URLs	DT, RF	96.33% 97.49%
[50]	2016	Phishing and legitimate sites	Auto-updated whitelist	89.38%
[51]	2014	Phishing URLs	Heuristic based approach	error rate-0.3%, false positive rate-0.2%, false negative rate- 0.5%
[52]	2020	Phishing website	AdaBoost-Extra Tree (ABET), Bagging –Extra tree (BET), Rotation Forest – Extra Tree (RoFBET), LogitBoost-Extra Tree (LBET)	97.485%, 97.404%, 97.449%, 97.576%
[53]	2021	Malware and malicious codes	LSTM, DCNN, CNN-LSTM, DTCNN-LSTM	79.5%, 80.6%, 91.4%, 93.2%
[54]	2021	Anomaly and malicious traffic in IoT	Feature selection based on chi-square, Pearson correlation, and score correlation	99.93%
[55]	2018	Malicious browser extensions	SVM, MLP, BN, LR	96.52% 93.48% 88.99% 86.16%
[56]	2021	Malicious application	KNN, NBM, TextCNN	92.17%
[57]	2017	Malicious JavaScript code	NB, J48,	95.06% 99.22%

			SVM, KNN	94.55% 97.14%
[58]	2019	Malicious domain name detection	N-gram	94.04%
[59]	2023	Malicious TLS flow	Unsupervised method	Precision, recall and F1: 99%
[60]	2019	Malicious behavior	H-gram, RF, AdboostM1, Bagging	96.8%
[61]	2022	Phishing and benign URLs	Conditional Generative Adversarial Network	ACC-87.45% F1-score-85.6% AUC-87.45%
[62]	2020	Malicious URL related to COVID-19	KNN (without entropy)	99.2%
[63]	2020	Phishing website	LR2, SVM, CNN, DBN-SVM	95.13%, 95.34%, 96.87%, 99.96%

References

[1] P. Ashwini and N. Vadivelan, "Security from phishing attack on internet using evolving fuzzy neural network," CVR Journal of Science and Technology, vol. 20, no. 1, pp. 50-55, 2021.

[2] D. Sahoo, C. Liu, and S. C. Hoi, "Malicious URL detection using machine learning: A survey," arXiv preprint arXiv:1701.07179, 2017.

[3] H. V. S. Aalla, N. R. Dumpala, and M. Eliazer, "Malicious URL prediction using machine learning techniques," Annals of the Romanian Society for Cell Biology, pp. 2170-2176, 2021.

[4] M. Aljabri et al., "Detecting malicious URLs using machine learning techniques: review and research directions," IEEE Access, 2022.

[5] J. Yuan, Y. Liu, and L. Yu, "A novel approach for malicious URL detection based on the joint model," Security and Communication Networks, pp. 1-12, 2021.

[6] G. N. Anil, "Detection of phishing websites based on feature extraction using machine learning," International Research Journal of Engineering and Technology (IRJET), 2020.

[7] J. Liu, "Lexical Features of Economic Legal Policy and News in China Since the COVID-19 Outbreak," Frontiers in Public Health, vol. 10, p. 928965, 2022.

[8] A. Joshi, L. Lloyd, P. Westin, and S. Seethapathy, "Using lexical features for malicious URL detection--a machine learning approach," arXiv preprint arXiv:1910.06277, 2019.

[9] TechTarget. [Online].

Available: https://www.techtarget.com/

[10] H. Choi, B. B. Zhu, and H. Lee, "Detecting malicious web links and identifying their attack types," in 2nd USENIX Conference on Web Application Development (WebApps 11), 2011.

[11] C. Johnson, B. Khadka, R. B. Basnet, and T. Doleck, "Towards Detecting and Classifying Malicious URLs Using Deep Learning," J. Wirel. Mob. Networks Ubiquitous Comput. Dependable Appl., vol. 11, no. 4, pp. 31-48, 2020.

[12] M. Cova, C. Kruegel, and G. Vigna, "Detection and analysis of drive-by-download attacks and malicious Javascript code," in Proc. 19th Int. Conf. World Wide Web (WWW), 2010, pp. 281–290, doi:10.1145/1772690.1772720.

[13] M. Sánchez-Paniagua et al., "Phishing URL detection: A real-case scenario through login URLs," IEEE Access, vol. 10, pp. 42949-42960, 2022.

[14] A. Pandey and J. Chadawar, "Phishing URL Detection using Hybrid Ensemble Model," international journal of engineering research & technology (IJERT), vol. 11, no. 04, 2022.

[15] M. Romagna and N. J. van den Hout, "Hacktivism and website defacement: motivations, capabilities and potential threats," in 27th virus bulletin international conference, 2017, pp. 1-10.

[16] M. Romagna and N. J. van den Hout, "Hacktivism and website defacement: motivations, capabilities and potential threats," in 27th virus bulletin international conference, 2017, pp. 1-10.

[17] P. Chang, "Multi-Layer Perceptron Neural Network for Improving Detection Performance of Malicious Phishing URLs Without Affecting Other Attack Types Classification," arXiv preprint arXiv:2203.00774, 2022.

[18] H. A. Tariq, W. Yang, I. Hameed, B. Ahmed, and R. U. Khan, "USING black-list and white-list technique to detect malicious URLs," IJIRIS::International Journal of Innovative Research Journal in Information Security, vol. 4, pp. 01-07, 2017.

[19] R. Kumar, X. Zhang, H. A. Tariq, and R. U. Khan, "Malicious URL detection using multi-layer filtering model," in 2017 14th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), 2017, pp. 97-100, doi: 10.1109/ICCWAMTIP.2017.8301457.

[20] W. Chu et al., "Protect sensitive sites from phishing attacks using features extractable from inaccessible phishing URLs," in 2013 IEEE international conference on communications (ICC), 2013, pp. 1990-1994.

[21] C. Seifert, I. Welch, and P. Komisarczuk, "Identification of malicious web pages with static heuristics," in 2008 Australasian Telecommunication Networks and Applications Conference, 2008, pp. 91-96.

[22] L. A. T. Nguyen, B. L. To, H. K. Nguyen, and M. H. Nguyen, "A novel approach for phishing detection using URL-based heuristic," in 2014 international conference on computing, management and telecommunications (ComManTel), 2014, pp. 298-303.

[23] M. G. Schultz, E. Eskin, F. Zadok, and S. J. Stolfo, "Data mining methods for detection of new malicious executables," in Proceedings 2001 IEEE Symposium on Security and Privacy. S&P 2001, 2001, pp. 38-49.

[24] R. A. Lekshmi and S. Thomas, "Detecting malicious URLs using machine learning techniques: a comparative literature review," Int Res J Eng Technol (IRJET), vol. 6, no. 06, 2019.

[25] Y. Wang, "Malicious URL Detection An Evaluation of Feature Extraction and Machine Learning Algorithm," Highlights in Science, Engineering and Technology, vol. 23, pp. 117-123, 2022.

[26] S. Abad, H. Gholamy, and M. Aslani, "Classification of Malicious URLs Using Machine Learning," Sensors, vol. 23, no. 18, p. 7760, 2023.

[27] M. Almuhaideb and M. Anwar, "A URL-Based Social Semantic Attacks Detection With Character-

Aware Language Model," IEEE Access, vol. 11, pp. 10654-10663, 2023.

[28] M. Aljabri et al., "An assessment of lexical, network, and content-based features for detecting malicious urls using machine learning and deep learning models," Computational Intelligence and Neuroscience, vol. 2022, 2022.

[29] S. He et al., "An effective cost-sensitive XGBoost method for malicious URLs detection in imbalanced dataset," IEEE Access, vol. 9, pp. 93089-93096, 2021.

[30] A. Maci et al., "Unbalanced Web Phishing Classification through Deep Reinforcement Learning," Computers, vol. 12, no. 6, p. 118, 2023.

[31] U. S. DR, A. Patil, and M. Mohana, "Malicious URL Detection and Classification Analysis using Machine Learning Models," in 2023 International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), 2023, pp. 470-476.

[32] X. Cho, D. Hoa, and V. Tisenko, "Malicious url detection based on machine learning," International Journal of Advanced Computer Science and Applications, 2020.

[33] R. Patgiri et al., "Empirical study on malicious URL detection using machine learning," in Distributed Computing and Internet Technology: 15th International Conference, ICDCIT 2019, Bhubaneswar, India, January 10–13, 2019, Proceedings 15, 2019, pp. 380-388.

[34] X. Tong et al., "MM-ConvBERT-LMS: Detecting Malicious Web Pages via Multi-Modal Learning and Pre-Trained Model," Applied Sciences, vol. 13, no. 5, p. 3327, 2023.

[35] N. Nagy et al., "Phishing URLs Detection Using Sequential and Parallel ML Techniques: Comparative Analysis," Sensors, vol. 23, no. 7, p. 3467, 2023.

[36] M. Alsaedi et al., "Cyber threat intelligence-based malicious url detection model using ensemble learning," Sensors, vol. 22, no. 9, p. 3373, 2022.

[37] B. Wei et al., "A deep-learning-driven light-weight phishing detection sensor," Sensors, vol. 19, no. 19, p. 4258, 2019.

[38] C. Hajaj, N. Hason, and A. Dvir, "Less is more: Robust and novel features for malicious domain detection," Electronics, vol. 11, no. 6, p. 969, 2022.

[39] M. Umer et al., "Deep Learning-Based Intrusion Detection Methods in Cyber-Physical Systems: Challenges and Future Trends," Electronics, vol. 11, no. 20, p. 3326, 2022.

[40] M. Elsadig et al., "Intelligent Deep Machine Learning Cyber Phishing URL Detection Based on BERT Features Extraction," Electronics, vol. 11, no. 22, p. 3647, 2022.

[41] S. R. Abdul Samad et al., "Analysis of the Performance Impact of Fine-Tuned Machine Learning Model for Phishing URL Detection," Electronics, vol. 12, no. 7, p. 1642, 2023.

[42] K. Sandra, L. ChaeHo, and S. G. Lee, "Malicious URL Detection Based on Associative Classification," entropy, 2021.

[43] S. S. Roy et al., "Multimodel phishing url detection using lstm, bidirectional lstm, and gru models," Future Internet, vol. 14, no. 11, p. 340, 2022.

[44] K. Fotiadou et al., "Network traffic anomaly detection via deep learning," Information, vol. 12, no. 5, p. 215, 2021.

[45] A. M. Almuhaideb et al., "Homoglyph Attack Detection Model Using Machine Learning and Hash Function," Journal of Sensor and Actuator Networks, vol. 11, no. 3, p. 54, 2022.

[46] J. Saxe and K. Berlin, "eXpose: A character-level convolutional neural network with embeddings for detecting malicious URLs, file paths and registry keys," arXiv preprint arXiv:1702.08568, 2017.

[47] O. Christou et al., "Phishing URL Detection Through Top-level Domain Analysis: A Descriptive Approach," arXiv 2020, arXiv preprint arXiv:2005.06599.

[48] S. Lee and J. Kim, "Warningbird: Detecting suspicious urls in twitter stream," in Ndss, 2012, pp. 1-13.

[49] S. P. Tung et al., "Using a machine learning model for malicious url type detection," in International Conference on Next Generation Wired/Wireless Networking, 2021, pp. 493-505.

[50] A. Jain and B. B. Gupta, "A novel approach to protect against phishing attacks at client side using

auto-updated white-list," EURASIP Journal on Information Security, vol. 2016, 2016.

[51] R. B. Basnet, A. H. Sung, and Q. Liu, "Learning to detect phishing URLs," International Journal of Research in Engineering and Technology, vol. 3, no. 6, pp. 11-24, 2014.

[52] Y. A. Alsariera et al., "AI meta-learners and extra-trees algorithm for the detection of phishing websites," IEEE access, vol. 8, pp. 142532-142542, 2020.

[53] L. Liu et al., "Learning-Based Detection for Malicious Android Application Using Code Vectorization," Security and Communication Networks, vol. 2021, pp. 1-11, 2021.

[54] T. D. Diwan et al., "Feature entropy estimation (FEE) for malicious IoT traffic and detection using machine learning," Mobile Information Systems, vol. 2021, pp. 1-13, 2021.

[55] Y. Wang et al., "A combined static and dynamic analysis approach to detect malicious browser extensions," Security and Communication Networks, vol. 2018, 2018.

[56] Y. Song et al., "Permission Sensitivity-Based Malicious Application Detection for Android," Security and Communication Networks, vol. 2021, pp. 1-12, 2021.

[57] N. Khan, J. Abdullah, and A. S. Khan, "Defending malicious script attacks using machine learning classifiers," Wireless Communications and Mobile Computing, vol. 2017, 2017.

[58] H. Zhao, Z. Chang, G. Bao, and X. Zeng, "Malicious domain names detection algorithm based on N-gram," Journal of Computer Networks and Communications, vol. 2019, 2019.

[59] G. Gomez, P. Kotzias, M. Dell'Amico, L. Bilge, and J. Caballero, "Unsupervised detection and clustering of malicious TLS flows," Security and Communication Networks, 2023.

[60] Y. Zhao, B. Bo, Y. Feng, C. Xu, and B. Yu, "A feature extraction method of hybrid gram for malicious behavior based on machine learning," Security and Communication Networks, 2019.

[61] S. A. Kamran, S. Sengupta, and A. Tavakkoli, "Semi-supervised conditional GAN for simultaneous generation and detection of phishing URLs: A game theoretic perspective," arXiv preprint arXiv:2108.01852, 2021.

[62] J. Ispahany and R. Islam, "Detecting malicious URLs of COVID-19 pandemic using ML technologies," arXiv preprint arXiv:2009.09224, 2020.

[63] X. Yu, "Phishing websites detection based on hybrid model of deep belief network and support vector machine," in IOP Conference Series: Earth and Environmental Science, vol. 602, no. 1, pp. 012001, IOP Publishing, November 2020.

[64] F. A. Aboaoja, A. Zainal, F. A. Ghaleb, B. A. S. Al-rimy, T. A. E. Eisa, and A. A. H. Elnour, "Malware detection issues, challenges, and future directions: A survey," Applied Sciences, vol. 12, no. 17, pp. 8482, 2022.