International Journal of Scientific Research in Science, Engineering and Technology

Print ISSN - 2395-1990
Online ISSN : 2394-4099

Available Online at : www.ijsrset.com
doi : https://doi.org/10.32628/IJSRSET

Malicious URL Detection based on Machine Learning

[1]N Sudha Laxmaiah, [2]Kandoju Praveshika, [3]Pallerla Ruthvika

[1]Assistant Professor, Department of CSE, Bhoj Reddy Engineering College for Women, Hyderabad, Telangana, India

[2,3]Students, Department of CSE, Bhoj Reddy Engineering College for Women, Hyderabad, Telangana, India

ARTICLEINFO

Article History :

Accepted: 10 June 2023
Published: 04 July 2023

Publication Issue :
Volume 10, Issue 4
July-August-2023

Page Number :
42-48

ABSTRACT

Currently, the risk of network information in security is increasing rapidly in number and level of danger. The methods mostly used by hackers to day are to attack end to end technology and exploit human vulnerabilities. These techniques include social engineering, phishing, pharming, etc. One of the steps in conducting these attacks is to deceive users with malicious Uniform Resource Locators (URLs). As a results, malicious URL detection is of great interest now a days. There have been several scientific studies showing a number of methods to detect malicious URLs based on machine learning and deep learning techniques. In this paper, we propose a malicious URL detection method using machine learning techniques based on our proposed URL behavior sand attributes. Moreover, bigdata technology is also exploited to improve the capability of detection malicious URLs based on abnormal behaviors. In short, the proposed detection system consists of a new set of URLs features and behaviors, a machine learning algorithm, and a big data technology. The experimental results show that the proposed URL attributes and behavior can help improve the ability to detect malicious URL significantly. This is suggested the proposed system may be considered as an optimized and friendly used solution for malicious URL detection.

Keywords: URL, malicious URL detection; feature extraction; feature selection; Machine learning.

I. INTRODUCTION

Uniform Resource Locator (URL) is used to refer to resources on the Internet. In [1], Sahooetal. Presented about the characteristics and two basic component sof the URL as: protocol identifier, which indicates what protocol to use, and resource name, which specifies the IP address or the domain name where the resource is located. It can be seen that each URL has a specific structure and format. Attackers often try to change one or more components of the URL's structure to deceive users for spreading their malicious URL. Malicious

URLs are known as links that adversely affect users. These URLs will redirect users to resources or pages on which attackers can execute codes on users' computers, redirect users to unwanted sites, malicious website, or other phishing site, or malware download. Malicious URLs can also be hidden in download links that are deemed safe and can spread quickly through file and message sharing in shared networks. Some attack techniques that use malicious URLs include [2, 3, 4]: Drive-by Download, Phishing and Social Engineering, and Spam. According to statistics presented in [5], in 2019, the attacks using spreading malicious URL technique are ranked first among the 10 most common attack techniques. Especially, according to this statistic, the three main URL spreading techniques, which are malicious URLs, URLs, there are two main trends at present as malicious URL detection based on signs or sets of rules, and malicious URL detection based on behavior analysis techniques[1,2].The method of detecting malicious URLs based on a set of markers or rules can quickly and accurately detect malicious URLs.

However, this method is not capable of detecting new malicious URLs that are not in the set of predefined signs or rules. In our research, machine learning algorithms are used to classify URL Machine learning algorithms are a part of the whole malicious URL detection system. Two supervised machine learning algorithms are used, Support vector machine (SVM)and Random forest(RF).

II. RELATED WORK

2.1 Signature based Malicious URL Detection

Studies on malicious URL detection using the signature sets had been investigated and applied long time ago[6,7,8].Most of these studies often use lists of known malicious URLs.; otherwise URLs will be considered as safe. The main disadvantage of this approach is that it will be very difficult to detect new malicious URLs that are not in the given list.

2.2 Machine Learning based Malicious URL Detection

There are three types of machine learning algorithms that can be applied on malicious URL detection methods, including supervised learning, unsupervised learning, and semi supervised learning. And the detection methods are based on URL behaviors. The behaviors and characteristics of URLs can be divided into two main groups, static and dynamic. In their studies [9, 10, 11] authors presented methods of analyzing and extracting static behavior of URLs, including Lexical, Content, Host, and Popularity-based. The machine learning algorithms used in these studies are Online Learning algorithms and SVM. Malicious URL detection using dynamic actions of URLs is presented in [12, 13]. In this paper, URL attributes are extracted based on both static and dynamic behaviors. Some attribute group sare investigated, including Character and semantic groups; Abnormal group in websites and Host-based group; Correlated group.

2.3 Malicious URL Detection Tools

• URLVoid: URL Void is a URL checking program using multiple engines and blacklists of domains. Some examples of URL Void are Google Safe Browsing, Norton Safe Web and My WOT.

The advantage of the Void URL tool is its compatibility with many different browsers as well as it can support many other testing services. The main disadvantage of the Void URL tool is that the malicious URL detection process relies heavily on a given set of signatures.

• Dr. Web Anti-Virus Link Checker: Dr. Web Anti Virus Link Checker is an add-on for Chrome, Firefox, Opera, and IE to automatically find and scan malicious content on a download link on all social networking links such as Facebook,Vk.com, Google+. Comodo Site Inspector: This is a malware and security hole detection tool. This helps users check URLs or enables webmasters to setup daily checks by

• downloading all the specified sites. And run the mina sandbox browser environment.

• Some other tools: Among aforementioned typical tools, there are some other URL checking tools, such as UnShorten. it, Virus Total, Norton Safe Web, Site Advisor (by McAfee), SucuriBrowser Defender, Online

Link Scan, and Google Safe Browsing Diagnostic. From the analysis and evaluation of malicious URL detection tools presented above, it is found that the majority of current malicious URL detection tools are signature-based URL detection systems. Therefore, the effectiveness of these tools is limited.

III. PROPOSED SYSTEM

For detection of Malicious URLs traditional filtering mechanism like Black-Listing, Heuristic Classification etc. was used. These old and conventional mechanisms are based on URL syntax matching and URL Keyword matching. Therefore, these older mechanisms cannot effectively deal with newly evolving URL technologies and also fail in detecting the modern URLs such as Embedded Links, Short URLs and Dark Web URLs. In the proposed classification approach machine learning algorithm is used in detection of malicious URLs. Figure 1. shows the model which contains two stages i.e., training stage and detection stage.

Training Stage: From the Dataset of URLs (Good URLs and Bad URLs), features are extracted and each URL has label '0' if it is non-Malicious and '1' if it is Malicious. The features that are extracted are Address Bar based features, Domain based features and HTML and JavaScript based features. This URLs are trained using Machine Learning algorithm.

Detection Stage: User inputted URL will be taken and then features are extracted from the URL and will classify as 'Good' (safe URL) or 'Bad' (Malicious URL). Here it will be testing the accuracy of the model depending upon the prediction made by it.

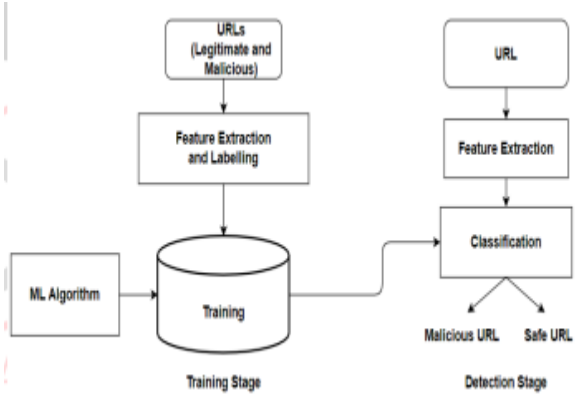

Figure 1. Malicious URL Detection Model using Machine Learning

There are two steps of machine learning technique. The first step is to assign correct feature through feature extraction so that it could give the deciding bits of knowledge (Legitimate-0, Phishing-1) in finding Malicious URLs and the next step is to use the above features to train a Machine Learning Algorithm. Here we will be classifying the URLs based on the features extracted from them. In this paper we have used Address Bar based features, Domain based features and HTML and JavaScript based features. The reason of using Address Bar based features, Domain based features and HTML and JavaScript based features is that even most of the new evolving URLs generated today should follow the same structure like the existing system. In Figure 2., we have discussed the flow and working of our model. The first phase of our model is collection of Benign (Open-Source Dataset) and Malicious URLs (Phistank Dataset) to form dataset. Total 10,000 URLs are used to form dataset. The complete dataset is stored using CSV format.

Malicious URLs: - Randomly 5,000 Malicious URLs are collected from open source service called phish tank to train ML models.

Legitimate URLs: -Randomly 5,000 Legitimate URLs are collected from open datasets of University of New Brunswick.

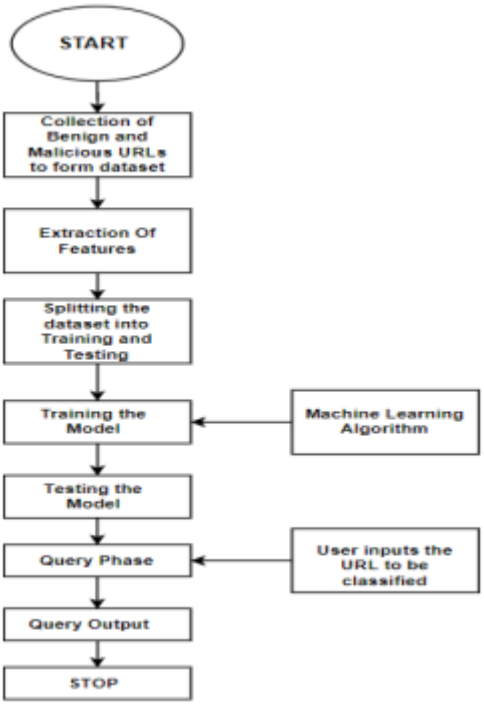

Figure 2. Work Flow

The second phase of the workflow is extraction of features. Total 15 features of URLs are extracted. Table 1. shows 15 features along with feature group and data type. Eight Address Bar based, four Domain based and three HTML and JavaScript based features are extracted. All features except depth of URL is of Boolean data type whereas depth of URL is of numeric data type. Malicious URLs are labelled as '1' and Legitimate URLs are labelled as '0'.In the third phase of the workflow the labelled data collected which consists of Benign and Phishing URLs undergoes the process of feature extraction where the various features are extracted, and then, the data are divided into Training dataset and the Testing dataset. We have divided training and testing data in two ratios 80:20 and 70:30.

In the next phase, the model is trained by passing the training data through various models such as XGBoost and Random Forest Algorithms. In this paper we have used two Supervised Machine Learning Algorithm XGBoost and Random Forest which is discussed in literature. Then, the Trained model is tested using the Testing dataset.

In the query phase user inputs, the URL that has to be classified. Input URL's feature is extracted and the output is '1' if the URL is Malicious and '0' if the URL is Legitimate.

Table 1. List of URL feature

Sr. No	Feature group	Feature	Data type
1	Address Bar	IP Address in URL	Boolean
2		Prefix or Suffix "-" in Domain part of URL	Boolean
3		"//" in URL	Boolean
4		"@" in URL	Boolean
5		Tiny URL	Boolean
6		"https://" in URL	Boolean
7		"http/https" in Domain Name	Boolean
8		Depth of URL	Numeric
9	Domain	DNS (Domain Name System) Record	Boolean
10		Website Traffic	Boolean
11		Age of Domain	Boolean
12		End Period of Domain	Boolean
13	HTML and JavaScript	Website Forwarding	Boolean
14		Iframe Redirection	Boolean
15		Status Bar Customization	Boolean

Figure 3. and Figure 4. shows the feature importance graph of RF which tells that RF considers "https//:" in URL part feature as an important.

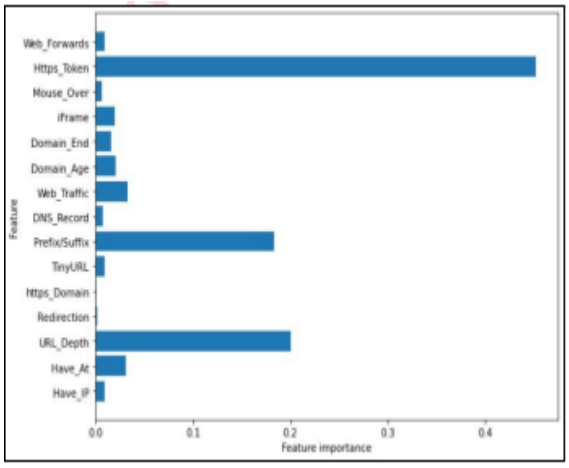

Figure 3. RF (70:30)

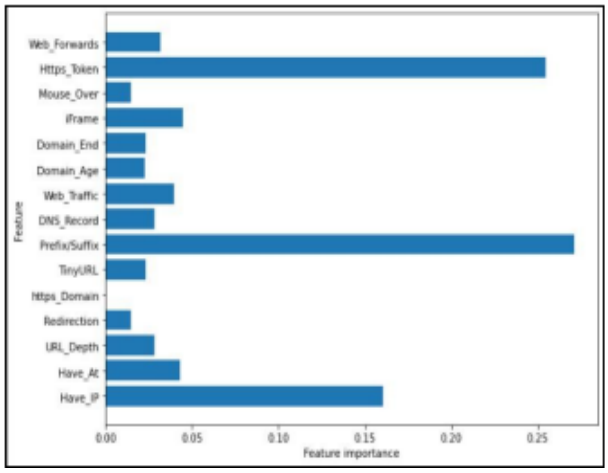

Figure 4. RF (80:20)

Figure 5. and Figure 6. shows the feature importance graph of XGBoost which tells that XGBoost considers prefix/suffix '-' in Domain part of URL feature as an important.

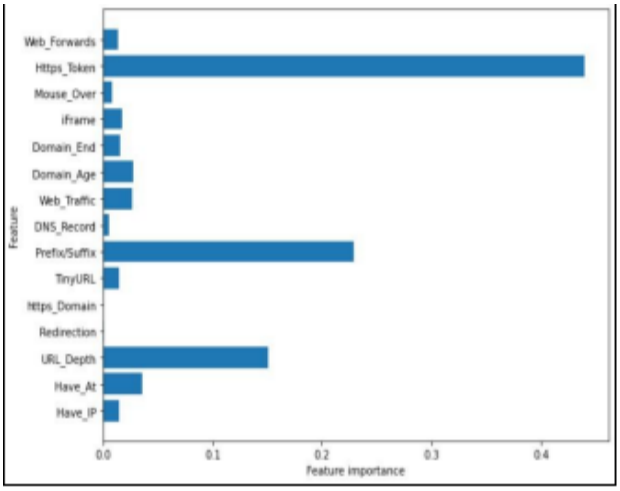

Figure 5. XGBoost (70:30)

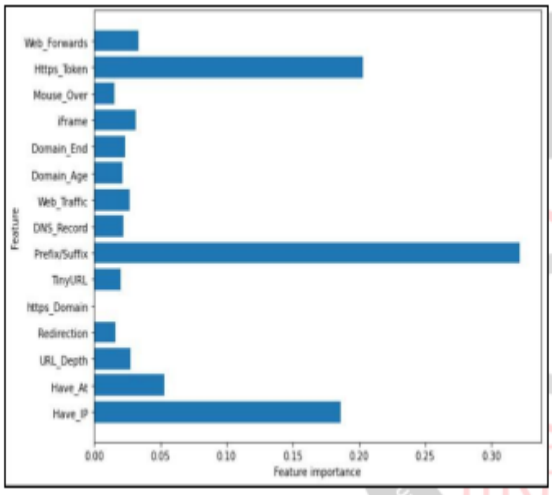

Figure 6. XGBoost (80:20)

IV. RESULTS AND DISCUSSION

Table 1. shows the Training Data Accuracy, Testing Data Accuracy and Combined Testing time taken for the both along with Split Ratios of the dataset. We had applied Random Forest Algorithm and Extreme Gradient Boosting (XGBoost) Algorithm. Results shows that XGBoost gives more accuracy than Random Forest specifically with the split ratio of 80:20.In the Figure 3. and Figure 4., the feature importance graph of Random Forest shows that HTTPs token is important feature among the features set which is used to classify the URL. Figure 5. and Figure 6. shows the feature importance graph for the Extreme Gradient Boost Algorithm. Here Prefix/Suffix is seen to be the most prominent feature for classification followed by HTTPs token and Have IP having their importance as well. These 3 important features give rise to much more accurate and improved results for classification of URL.

Table 2. Accuracy of Machine Learning Methods

Algorithm	Split Ratio	Training Data Accuracy	Test Data Accuracy	Training +Testing Time(s)
Random Forest	70:30 Ratio	0.784	0.777	0.089
	80:20 Ratio	0.778	0.776	0.091
XGBoost	70:30 Ratio	0.840	0.823	0.064
	80:20 Ratio	0.838	0.833	0.064

V. CONCLUSION AND FUTURE SCOPE

In this work, we have presented how we can train a Machine Learning model to make it classify the URL into Malicious or Genuine URL based on the features of the URL. When the traditional methods fail to detect the newly evolving URLs our method of classification can surely come up with the improved results. We also compared the accuracy of many Machine Learning Algorithms to classify the URL, out

of which we found that XGBoost gave the best results among the algorithms.

The Future Scope of this work would be training the Machine Learning model with more data and also with more features of the URL for more accurate and improved results. Model can be further trained to detect the Dark Websites. Moreover a Browser Extension can also be made for this so that the process can run in the background continuously to filter the Malicious Websites dynamically.

VI. REFERENCES

[1]. Cho Do Xuan,Hoa Dinh Nguyen, Tisenko Victor Nikolaevich,(2020) "Malicious URL Detection based on Machine Learning", International Journal of Advanced Computer Science and Applications.

[2]. Eint Sandi Aung, Hayato Yamana, (2020) "Malicious URL Detection:A Survey", Department of Computer Science and Communication Engineering, Graduate School of Fundamental Science and Engineering.

[3]. Ripon Patgiri, Hemanth Katari, Ronit Kumar and Dheeraj Sharma, (2020) "Empirical Study on Malicious URL Detection Using Machine Learning", International Conference, ICDICT.

[4]. Tie Li, Gang Kou, Yi Peng (2020) "Improving Malicious URLs Detection via Feature Engineering: Linear and nonlinear Space Transformation Methods", Information Systems (Elsevier).

[5]. Immadisetti Naga Venkata Durga Naveen, Manamohana K, Rohit Verma, (2019) "Detection of Malicious URLs using Machine Learning Techniques", International Journal of Innovative Technology and Exploring Engineering (IJITEE).

[6]. Vanitha N and Vinodhini V, (2019) "Malicious URL Detection using Logistic Regression Technique", International Journal of Engineering and Management Research.

[7]. Lekshmi A R, Seena Thomas (2019) "Detecting Malicious URLs Using Machine Learning Techniques: A Comparative Literature Review", International Research Journal of Engineering and Technology (IRJET).

[8]. Yasin Sonmez, Turker Tuncer, Huseyin Gokal, Engin Avci (2018)"Phishing Web Sites Features Classification Based on Extreme Learning Machine", 6th International Symposium on Digital Forensic and Security (ISDFS)

[9]. G.Sai Chaitanya Kumar, Dr.Reddi Kiran Kumar, Dr.G.Apparao Naidu, "Noise Removal in Microarray Images using Variational Mode Decomposition Technique " Telecommunication computing Electronics and Control ISSN 1693-6930 Volume 15, Number 4 (2017), pp. 1750-1756

[10]. G. S. C. Kumar, D. Prasad, V. S. Rao and N. R. Sai, "Utilization of Nominal Group Technique for Cloud Computing Risk Assessment and Evaluation in Healthcare," 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), 2021, pp. 927-934, doi: 10.1109/ICIRCA51532.2021.9544895

[11]. V. S. Rao, V. Mounika, N. R. Sai and G. S. C. Kumar, "Usage of Saliency Prior Maps for Detection of Salient Object Features," 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2021, pp. 819-825, doi: 10.1109/I-SMAC52330.2021.9640684

[12]. YAHYA, Ammar; AHMAD, R.Badlishah; MOHD YACOB, Yas min; MOHD WARIP, Mohd Nazri Bin. Lightweight phishingURLs detection using N-gram features. 2016, vol. 8, pp. 1563–1570.

[13]. VERMA, Rakesh; DAS, Avisha. What's in a URL: Fast FeatureExtraction and Malicious URL Detection. In: 2017, pp. 55–63.

[14]. VERMA, Rakesh; DYER, Keith. On the Character of Phishing URLs: Accurate and Robust Statistical Learning Classifiers. CO DASPY 2015 - Proceedings of the 5th ACM Conference on Data andApplication Security and Privacy. 2015.

[15]. PAO, H.; CHOU, Y.; LEE, Y. Malicious URL Detection Based onKolmogorov Complexity Estimation. In: 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology. 2012.

Cite this article as :

N Sudha Laxmaiah, Kandoju Praveshika, Pallerla Ruthvika, "Malicious URL Detection based on Machine Learning", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 10 Issue 4, pp. 42-48, July-August 2023.
Journal URL : https://ijsrset.com/IJSRSET23103157