International Journal for Research in Engineering Application & Management (IJREAM)
ISSN : 2454-9150 Vol-07, Issue-05, AUG 2021

Malicious URL Detection using Machine Learning

[1]Siddhant Hosalikar, [2]Saikumar Iyer, [3]Ankit Limbasiya, [4]Prof. Suvarna Chaure

[1,2,3]B.E. Student, Dept. of Computer Engineering, SIES GST, Navi Mumbai, India

[1]hosalikarsiddhant07@gmail.com,[2]sai.iyer.98031@gmail.com, [3]ankitpatel518@gmail.com

[4]Assistant Professor, Dept. of Computer Engineering, SIES GST, Navi Mumbai, India

[4]suvarna.kendre@siesgst.ac.in

Abstract: URL is the heart of any browser and without the URL, browser is of no use. Sometimes URL is misused in order to perform tasks which are unethical such as Phishing, Hacking etc. Using a Malicious URL is nowadays one of the easiest ways to deceive the end users and retrieve sensitive and confidential information from them. So now it is important to counter such activities in order to protect the user from being exposed while on the Internet. Black Listing, Heuristic Classification is some of the traditional methods of URL Classification, but these methods fail when it comes to classifying Short URLs, Embedded URLs. Our method of using Machine Learning to detect Malicious URL by extracting features from URL is much safe and better since there is no need to access the website or click the website. The method proposed in this paper will try to detect Malicious URL effectively.

Keywords — URL, Black listing, Heuristic classification, Malicious URL detection, Feature extraction, Machine learning.

I. INTRODUCTION

URL is the heart of any browser and misuse of it has now become a very common issue. It is quite obvious that without a URL nobody can access a browser. But now many people with have started using URL to perform unethical activities. Today's trend is that the Malicious URL is either embedded inside a Legitimate URL or the user gets redirected to a Malicious Website after clicking in the link. Even clicking on such links can enable the intruder enter the system and wreak havoc. So now it is the need of the hour in order to counter such malicious activities and safeguard the user information from getting exposed.

Attackers attack the ordinary users through different media and mostly the common media are web, mail and network. In web medium the targeted users are unaware of checking URL link and the link seem to be realistic which has name similar to legitimate and the user is directed to phishing website. In mail medium there are two types of targets individual and organization. The phishers trick the superiors of organizations through fake trust-worthy information. The victim receives an email in which they are asked to update or confirm their information which is personal by clicking on link given within the email. In network medium phishing attack takes place when router is hijacked and addresses of DNS server is changed.

Whenever the user is surfing anything in the browser, it is necessary that every URL that is present on the webpage should be thoroughly checked for any Malicious or Phishing content. If present the user should be prompted before the user has clicked the link. This should be like the background process in a browser to constantly check the whole webpage.

The remainder of this paper is organized as follows. Reviews related works in Section 2. Proposed system in Section 3. Methodology in Section 4. Results in Section 5. Conclusion and future scope in Section 6.

II. LITERATURE REVIEW

The proposed detection system in [1] consists of URLs features, a machine learning algorithm and bigdata technology where Random Forest and Support Vector Machine Algorithm is used and RF with 100 trees gives the best predictive result when compared to RF with 10 trees while SVM considers whole features of URL and that affects in whole performance and RF with a relatively small number of samples gives good results.

Machine Learning methods and Neural Network methods is discussed in survey [2]. In Machine Learning – Naive Bayes, Support Vector Machine (SVM), KNearest Neighbor (KNN), Decision Trees, Random Forest, Gradient Boosting, XGBoost, AdaBoost and Logistic Regression is studied whereas in Neural Network -Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Generative Adversarial Network (GAN) and Neural

 DOI : 10.35291/2454-9150.2021.0420

Network Architecture is studied. The result of the survey was that Machine Learning Algorithm has objective of only detection whereas Neural Network methods is used more as it exposes Malicious URLs beforehand.

Host Based Features, Lexical Features and Site Popularity Features of URLs are extracted in [3]. Support Vector Machine and Random Forest Algorithm is used and the whole dataset was divided into mainly 3 types where ratios were 80:20,70:30 and 60:40 and accuracy of Support Vector Machines and Random Forest was calculated for ratio 80:20, 70:30 and 60:40 and it was found that average accuracy of Support Vector Machines was less than Random Forest.

The paper [4] uses five space transformation models singular value decomposition, distance metric learning, Nystrom methods, DML-NYS and NYS-DML. These five models are used to generate new features.

In the paper [5] detection of Malicious URLs is done by extracting 18 features. The paper shows the importance of feature extraction as new and existing Malicious URLs have the same structure which will surely outperform the traditional and conventional methods of classifying the URLs. These traditional methodologies also are not able to detect modern URLs such as embedded URLs, short URLs and dark web URLs.

The edifice of the whole model proposed in [6] is data-sets and hence there is sufficient and appropriate amount of data for Benign as well as Phishing URLs existing in the model so that it can be trained and tested for the classification. Three machine learning algorithms is used Logistic Regression, Naive Bayes and Random Forest. Here, the model is trained with three split ratio 1:1,4:1 and 10:1. Logistic Regression obtained maximum learning accuracy when compared to Naive Bayes and Random Forest.

Survey in [7] has discussed different types of methods used for detecting malicious URLs. Machine Learning is the best method for detecting malicious URLs among all the methods discussed.

Naive Bayes (NB), Support Vector Machine (SVM), Artificial Neural Networks (ANN) classification methods is used in [8]. Total 30 features of URLs were extracted which included 12 Address Bar based features, 6 abnormal based Features, 5 HTML and 7 JavaScript based Features and Domain based Features. ELM (Extreme Machine Learning Algorithms) achieved higher performance as compared to other methods Support Vector Machine (SVM) and Naive Bayes in terms of performance and speed.

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.

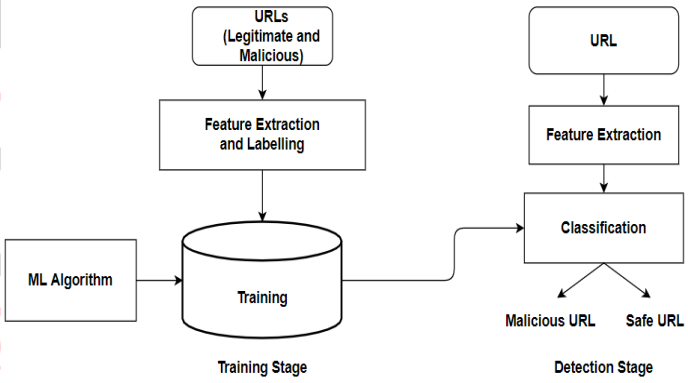

Figure 1. Malicious URL Detection Model using Machine Learning

IV. METHODOLOGY

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

 DOI : 10.35291/2454-9150.2021.0420

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 opensource 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.

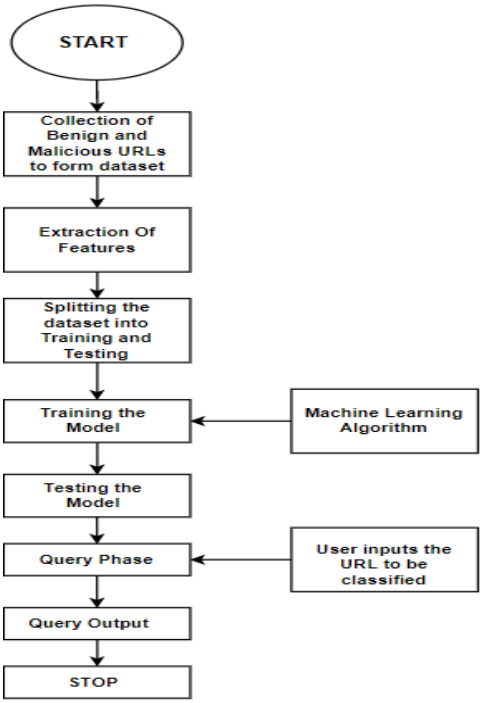

Figure 2. Work Flow

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.

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

Table 1. List of URL feature

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.

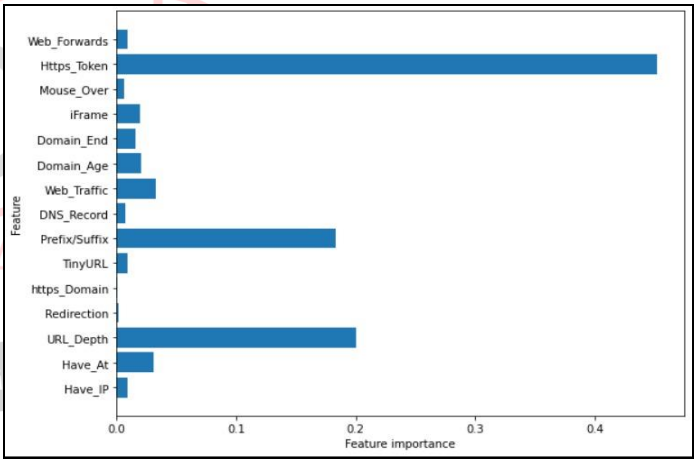

Figure 3. RF (70:30)

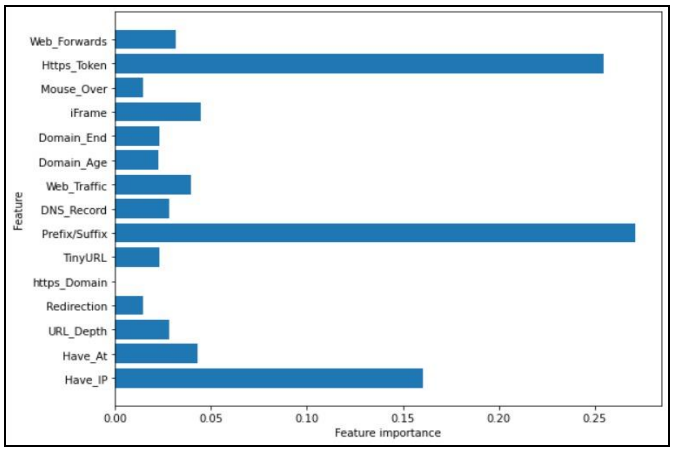

Figure 4. RF (80:20)

 DOI : 10.35291/2454-9150.2021.0420

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.

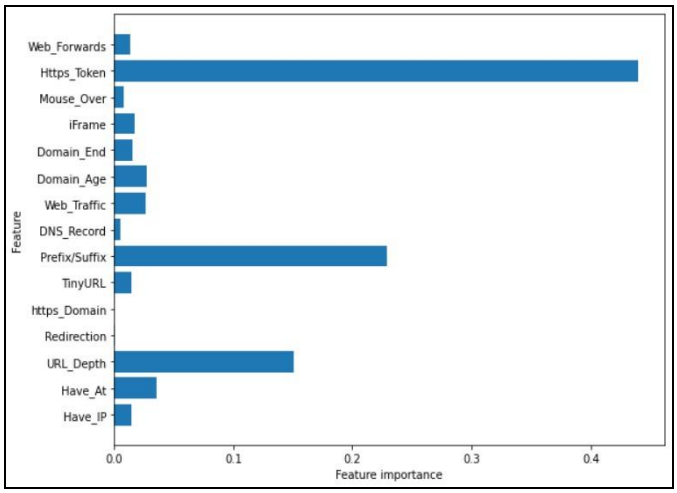

Figure 5. XGBoost (70:30)

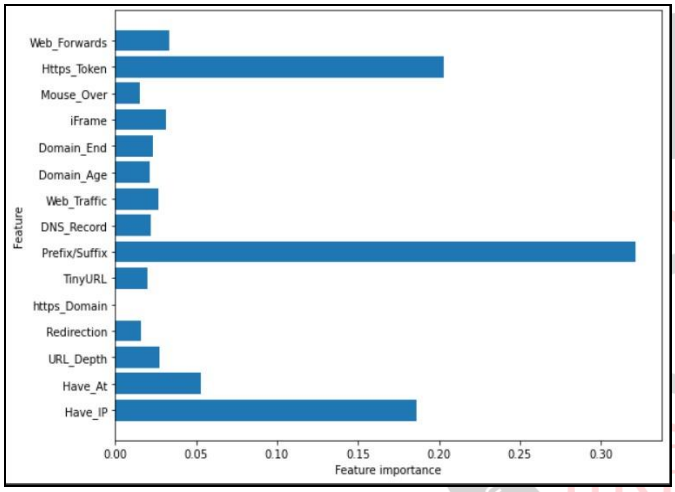

Figure 6. XGBoost (80:20)

V. RESULTS

Table 2. 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.

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

Table 2. Accuracy of Machine Learning Methods

VI. 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.

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).*

 DOI : 10.35291/2454-9150.2021.0420