International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-4S, April 2019

Machine Learning Based Malicious URL Detection

Divya Kapil, Atika Bansal, Anupriya, Nidhi Mehra, Aditya Joshi

Abstract: ***Today Internet technology has become an essential part of our life for education, entertainment, gaming, banking and communication. In this modern digital era, it is very easy to have any information by one click. But everything which has pros and cons, as we have any information at our tips but Internet is an attack platform also. When we use Internet to make our work easy same time many attacker try to steal information from our system. There are many means for attacking, malicious URL one of them. When a user visits a website, which is malicious then it triggers a malicious activity which is predesigned. Hence, there are various approaches to find dangerous URL on the Internet. In this paper, we are using machine learning approach to detect malicious URLs. We used ISCXURL2016 dataset and used J48, Random forest, Lazy algorithm and Bayes net classifiers. As performance metrics, we calculate accuracy, TPR, FPR, precision and recall.***

Keywords: Malicious URL, Classifiers, Random forest, Lazy algorithm, Bayes net, Machine Learning

I. INTRODUCTION

Cyber security is a very important requirement for users. The objective of cyber security to prevent the harm to software and hardware, network and it protects from attacker who can steal the valuable information. According to Google report in 2018, it finds many new malicious web pages daily which is thousand in numbers. These malicious links or web pages are created by compromising legitimate web pages. Hence the security of cyber is the major challenge for the researchers. There are many conventional techniques such as authentication, firewall, data encryption etc. to protect from many types of attacks. There are various challenges in cyber security, one of the major issues is, having malicious URL on the Internet. Malicious URL or malicious website is very serious issue for cyber security. URL stands for uniform resource locator which indicates the resources on www. There are two parts of the uniform resource locator: a) name of the resource which refers the domain name or IP address where the resource is situated. b) protocol identifier which refers which protocol is used. Figure 1 shows a uniform resource locator example. These characteristics are presented by sahoo et al. [1]. These components can be changed by the attackers to cheat the users, they use these malicious URLs or link which redirect them to malicious pages or resources where a malicious activity is triggered.

Manuscript published on 30 April 2019.
***** Correspondence Author (s)
Divya Kapil, School of Computing, CSE, Graphic Era Hill University, Dehradun, India. E-mail: divya.k.rksh@gmail.com
Atika Bansal, School of Computing, CSE, Graphic Era Hill University, Dehradun, India. E-mail: atika04591@gmail.com
Anupriya, School of Computing, CSE, Graphic Era Hill University, Dehradun, India. E-mail: anu.coerian@gmail.com
Nidhi Mehra, School of Computing, CSE, Graphic Era Hill University, Dehradun, India. E-mail: nidhigehu@gmail.com
Aditya Joshi, Graphic Era Deemed To Be University, Dehradun, India. E-mail: adi.joshi@geu.ac.in

Hence now-a-days detection of malicious URLs is interested research issue for the researchers. There are various traditional approaches such as blacklist method. Blacklist method is very simple and can provide better accuracy if lists are updated timely but this method can be less efficient to find malicious URLs which are newly generated. Nowadays Machine learning is used in various fields and cyber security one of them. In modern techniques of malicious URLs detection, machine learning is playing very important role to identifying the malicious URLs. In this paper we focus on machine learning based approaches. Download links which appears safe can have the malicious URLs which are hidden. There are many detection malicious URLs techniques such as phishing, spam, social engineering and drive-by Download.

In the detection of malicious URLs, there are various issues such as large amount of data which is very problematic.

- There is a large number more than thirty trillion of URLs which are unique Sullivan et al. [2].
- Other issue is feature selection for the quality performance of machine learning algorithms. Features collection can be time consuming as the techniques use features which are based on page contents and host based.
- It is very difficult to trace the botnets which are made by malwares due to the large size of web. There is requirement of resources for searching the web page which is infected.
- Under sampling is can be cause of imbalanced classification issue.

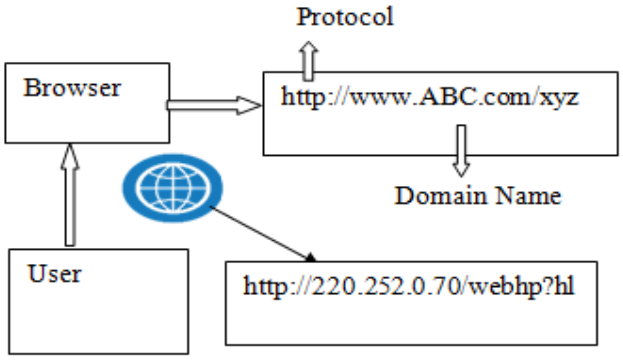

Figure 1

We categories this paper is categorized into five sections. Section II we discuss the related work of malicious URLs and compare the techniques. In section III, we presented our methodology in which we discussed about the dataset and the used classifiers. In section IV we presented the results and in section V, we concluded our paper and also future directions.

II. RELATED WORK

In related work we present the techniques for malicious URL techniques and compare these techniques.

Table 1 shows a comparison chart of the techniques. Figure 2 [6] shows Detection of Malicious URL Model using Machine Learning.

McGrath et al. [2] shows the difference between phishing and benign URLs. They used the features domain length and URL.

Kan et al. [3] used URL based classification of web page, not content based. The URL is divided in various tokens and from the token, features for classification are extracted. It improves the rate of classification.

Frank et al. [4] discussed the malicious URLs detection as problem of binary classification. They used SVM, Naive Bayes, Decision Trees, Multi-Layer Perceptron, Random Forest, and KNN classifiers. They show Multi-Layer Perceptron and Random Forest has the highest accuracy.

Zhao et al. [5] evade the imbalance of class in the detection of malicious URL. They used their CSOAL (Cost-Sensitive Online Active Learning) framework.

Cho Do Xuan et al. [6] proposed detection method which has a set of URLs behaviours and features and big data technique and machine learning technique. They show the results that URL attributes and behaviour which are proposed, these can improve detection of malicious URL.

Anjali Bet al. [7] presented many aspects of process of Uniform Resource Locator classification that identifies the website is benign or malicious. They used Naïve Bayes algorithm for malicious URLs detection and automated classification. They show that Naïve Bayes gets the better accuracy than SVM.

Yong Shi et al. [8] proposed a malware detection technique for APT attacks and this technique is machine learning based. They used Extreme Learning Machine which is a neural network technique which has fast speed for learning and accuracy. They show more than 95 % accuracy.

Baojiang Cui et al. [9] proposed a novel detection technique which is based on machine learning. They combined two techniques, for feature extraction they used sigmoidal threshold and statistical analyses which is based on gradient learning. They used decision tree, Support vector machine and Naïve Bayes classifiers to get the accuracy and efficiency of the proposed technique. They show the accuracy rate 98.7%.

CHUN-MING WU et al. [10] proposed a method which is based on static feature of URLs to detect website which is malicious. They used statistic for extraction of features which are used in classification. They achieved better efficiency and accuracy.

Chong et al. [11] presented a method which is machine learning based to detect malicious URLs. They combined JavaScript source features, payload size and URL lexical features. They used Support Vector Machine with a polynomial kernel and get 0.81 accuracy and 0.74 F1 score.

Justin Ma et al. [12] trained dataset with the features such as separating character of subdomain count, overall URL length and host-name length. They combined the host information with lexical features and achieved 95% accuracy rate.

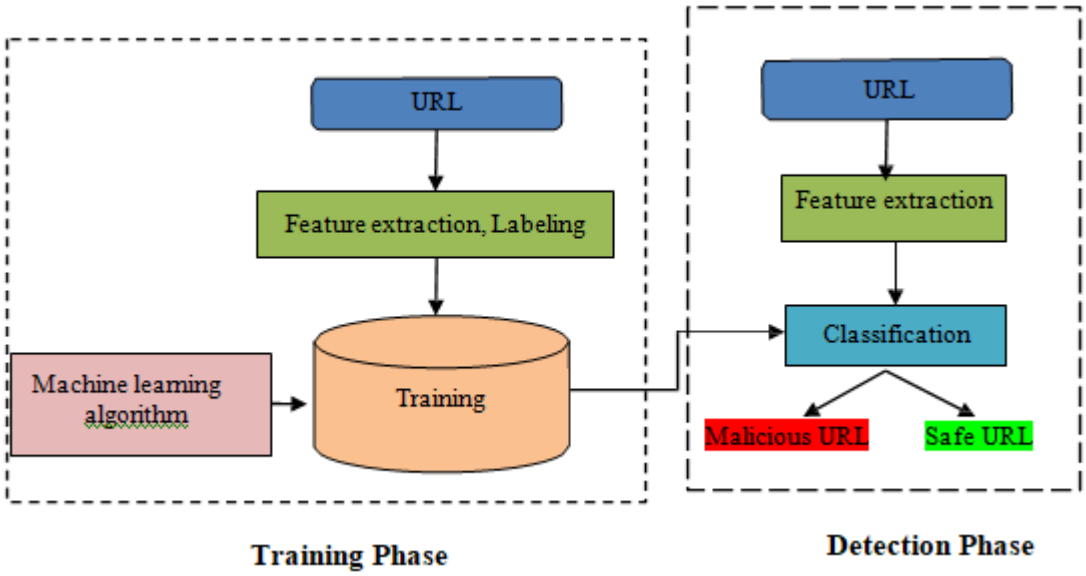

Figure 2 Detection of Malicious URL Model using Machine Learning [6]

Table1 Comparison Chart of Techniques

Paper	Technique	Features	Classifiers	Result
McGrath et al. [2]	-	domain length and URL	-	difference between phishing and benign URLs.
Kan et al. [3]	URL based classification of web page	From token	-	improves the rate of classification.
Frank et al. [4]	-	-	SVM, Naive Bayes, Decision Trees, Multi-Layer Perceptron, Random Forest, and KNN	Multi-Layer Perceptron and Random Forest has the highest accuracy
Zhao et al. [5]	Cost-Sensitive Online Active Learning) framework.	-	-	evade the imbalance of class
Cho Do Xuan et al. [6]	detection method with machine learning and Big data technique	-	-	improve detection of malicious URL
Anjali Bet al. [7]	identifies the website is benign or malicious	-	Naïve Bayes	Better accuracy
Yong Shi et al. [8]	Extreme Learning Machine	-	-	fast speed for learning and 95 % accuracy
Baojiang Cui et al. [9]	Machine learning based technique	-	decision tree, Support vector machine and Naïve Bayes	98.7 % accuracy
CHUN-MING WU et al. [10]	-	Static Feature of URL	-	better efficiency and accuracy.
Chong et al. [11]	Machine learning based technique	JavaScript source features, payload size and URL lexical features.	Support Vector Machine with a polynomial kernel	0.81 accuracy and 0.74 F1 score.
Justin Ma et al. [12]	combined the host information with lexical features	separating character of subdomain count, overall URL length and host-name length.	-	95% accuracy rate.

III. METHODOLOGY

In our methodology, we discuss ISCXURL2016 dataset and Random Forest, J48 and Lazy algorithm classifier. For our experiment we use the Weka tool.

[A] Dataset: We use ISCX-URL-2016 [13] dataset for our experiment. There are five types of URLs in this dataset. There are number of samples of different types of URLs. For our experiment we take random samples for each URLs. In table 2, we show the various types of URLs and their sample.

Table 2 URLs in dataset

URL	Number of samples	Source
Malware	11,500	DNS-BH
Benign	35,300	Alexa top websites
Defacement	45,450	Alexa ranked trusted websites
Spam	12,000	Publicly available WEBSPAM-UK2007 dataset.
Phishing	10,000	Open-Phish

[B] Weka Tool: WEKA [14] is open source and a data mining and machine learning based tool that build models. In Weka tool, there are many classifiers or machine learning algorithms which can be used directly on dataset and implement model for preprocessing of data, regression, classification, clustering and association rules and also has a visualization feature.

[C] Pre-Processing: In our experiment, we take random number of samples. Sample dataset contains 47 attributes. We have rejected some features so that overfitting problem can be avoided. The dataset is in 'Malware', 'Spam', 'Benign', 'Defacement' and 'Phishing' form. We divided dataset into 80% for training and 20% for testing.

Table 3 Pre-processing on Dataset

Number of features	47
Total Number of samples	4999
Number of Classes	5

[D] Classifiers

i. **J48 Decision Tree:** Decision tree is a classifier which uses tree like structure which gives a clear distribution. This algorithm was developed by Quinlan in 1993. It is a supervised algorithm and it uses divide and conquer approach. In decision tree algorithm, every node refers to test on attribute and every edge represents result of test and child (leaf) nodes denotes class which comes after all decisions.

ii. **Random Forest:** Random forest is machine learning algorithm which is used for classification and regression. Random forest is ensemble learning technique, Bagging and boosting are also ensemble learning algorithms. Using random forest, a big dataset can be executed efficiently.

iii. **Lazy Classifier:** Lazy classifier can deal with many problems simultaneously and solve them successfully [15]. Lazy classifiers need big space to store training dataset.

iv. **Bayes-Net:** Bayes-Net is also known as Bayesian networks which represents probabilistic relationship among set of random variables graphically.

[E] Experiment: For experiment we use ISCXURL2016 dataset and J48, Random forest, Bayes-Net and lazy classifiers. We load the sample dataset which has 47 attributes and 4999 samples of various URLs. We build the model using J48, Random forest, Bayes-Net and lazy algorithms and after that give test data to these classifiers. Finally, we obtain the result, we take TPR, FPR, F-measure, Recall, and Precision as a measurement metrices. Figure 3 shows the flow Chart of Experiment.

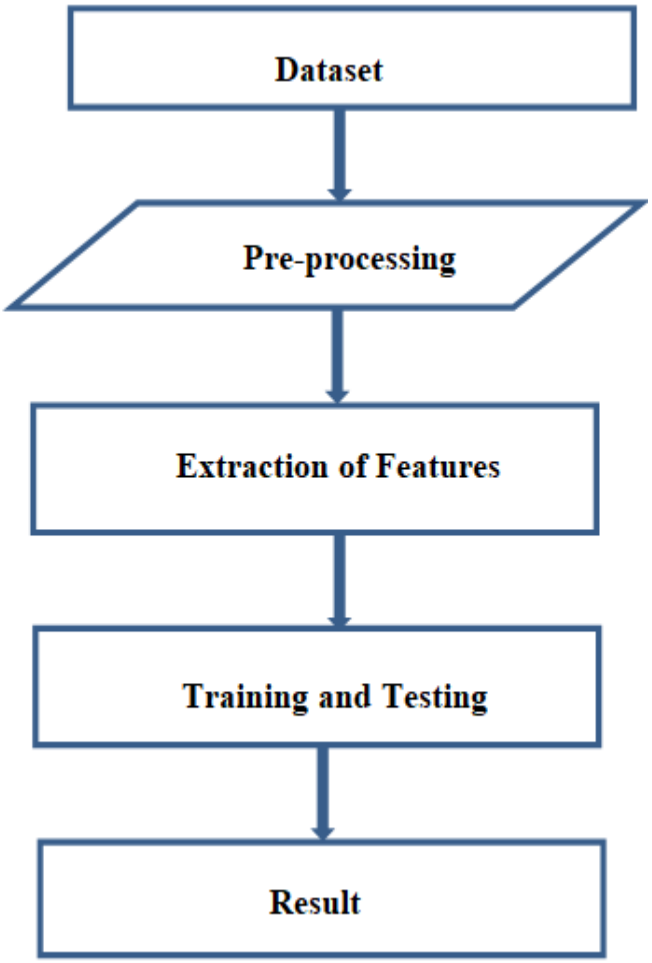

Figure 3 Experiment Flow Chart

IV. RESULT

We use performance metrices True positive rate, false positive rate, Precision, Recall, and F-measure. Table 4 shows the result of our experiment. Figure 4 shows the classification of different types of URLs.

Precision = True Positive / False Positive +True Positive

Recall= True Positive / True Positive +False Negative

F-MEASURE =2* recall* precision / recall + precision

Table 4 Result

Classifier	TPR	FPR	Precision	Recall	F-Measure
J48	0.944	0.032	0.944	0.944	0.944
Random Forest	0.961	0.027	0.961	0.961	0.961
Bayes-Net	0.921	0.051	0.922	0.922	0.922
Lazy	0.954	0.028	0.954	0.954	0.954

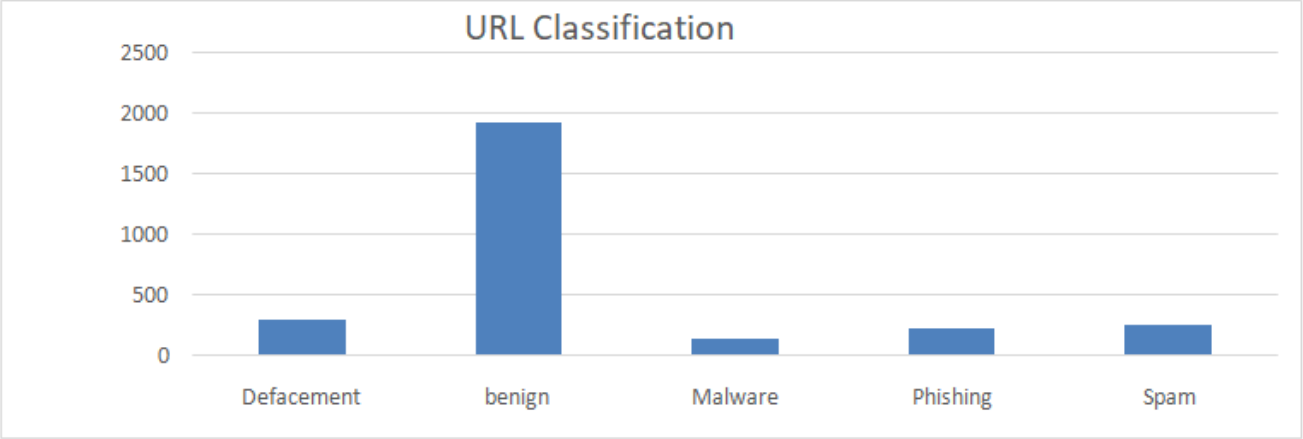

Figure 4 URLs Classification

V. CONCLUSION

Malicious URLs are the biggest threat for the cyber security. In this paper, we discuss about the malicious URLs detection techniques and compare these techniques. We used ISCX-URL-2016 dataset and use J48, random forest, Bayes-Net and lazy classifiers. We show the result using performance metrices TPR, FPR, Precision, Recall and F-measure. Malware URLs is a critical issue for the researchers and machine learning technique is very helpful in various areas. Malicious URLs detection using machine learning better idea than conventional techniques. In future work we will propose a machine learning based model for classification.

REFERENCES

1. D. Sahoo, C. Liu, S.C.H. Hoi, “Malicious URL Detection using Machine Learning: A Survey”, CoRR, 2017.
2. McGrath, D. Kevin, and Minaxi Gupta, "Behind Phishing: An Examination of Phisher Modi Operandi", LEET 8 (2008): 4.
3. Kevin, M.D., Gupta, M, “Behind phishing: an examination of phisher modi Operandi”, In: Proceedings of the 1st Usenix Workshop on Large-Scale Ex-ploits and Emergent Threats (2008)
4. Vanhoenshoven, Frank, et al. "Detecting malicious URLs using machine learning techniques.", IEEE Symposium Series on Computational Intelligence (SSCI), 2016.
5. P. Zhao and S. C. Hoi, “Cost-sensitive online active learning with application to malicious URL detection,” in Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2013, pp. 919–927.
6. Cho Do Xuan, Hoa Dinh Nguyen, Tisenko Victor Nikolaevich, “Malicious URL Detection based on Machine Learning”, International Journal of Advanced Computer Science and Applications, Vol. 11, No. 1, 2020
7. Sayamber, Anjali B., and Arati M. Dixit, "Malicious URL detection and identification", International Journal of Computer Applications 99.17 (2014): 17-23.
8. Shi, Y., Chen, G. & Li, J, “Malicious Domain Name Detection Based on Extreme Machine Learning”, Neural Process Lett 48, 1347–1357 (2018).
9. Baojiang Cui, Shanshan He, Xi Yao, “Malicious URL detection with feature extraction based on machine learning”, “International Journal of High-Performance Computing and Networking”, Volume 12, Issue 2.
10. WU, Chun-ming, "Malicious website detection based on URLs static features.", DEStech Transactions on Computer Science and Engineering mso (2018).
11. Chong, Christophe, Daniel Liu, and Wonhong Lee. "Malicious url detection." (2009).
12. Ma, Justin, Lawrence K. Saul, Stefan Savage, Geoffrey M. Voelker,” Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs.” , http://www.cs.berkeley.edu/jtma/papers/beyondbl-kdd2009.pdf.
13. Mohammad Saiful Islam Mamun, Mohammad Ahmad Rathore, Arash Habibi Lashkari, Natalia Stakhanova and Ali A. Ghorbani, "Detecting Malicious URLs Using Lexical Analysis", Network and System Security, Springer International Publishing, P467--482, 2016.
14. Deshmukh J.J. And Tated R.R.,” Weka - Open Source Technology, Its Implementation and Benefits”, World Research Journal of Computer Architecture, Volume 1, Issue 1, 2012, pp.-01-05.
15. Kaushik H. Raviya, Biren Gajjar,” Performance Evaluation of Different Data Mining Classification Algorithm Using WEKA”.