International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.089|

||Volume 9, Issue 4, April 2020||

Malicious/Phishing Webpage Detection using Machine Learning

Chitrashree Kurtkoti[1], Fazeelat Nandgadi[2], Vishal Adiyandra[3], Annapoorna Thingalaya[4], Pawan C. B. [5]

Assistant Professor, Department of Computer Science and Engineering, S.G.Balekundri Institute of Technology, ShivabasavaNagar, Belagavi, India[1]

U.G. Student, Department of Computer Science and Engineering, S.G.Balekundri Institute of Technology, Shivabasava Nagar, Belagavi,India[2,3,4,5]

ABSTRACT: The Malicious URL, a.k.a. Malicious Website, is a common and serious threat to cybersecurity. Malicious/Phishing URLs host unsolicited content and are used to perpetrate cybercrimes. Malicious/Phishing URLs may lead to illegitimate access to the user data by adversary.It is imperative to detect them in a timely manner. Traditionally, this is done through the usage of blacklists, Heuristic Classification etc., which cannot be exhaustive, and cannot detect newly generated Malicious/Phishing URLs. Furthermore, these approaches also fall short in detecting the modern URLs such as short URLs, dark web URLs. Even though we are currently interested in just machine learning techniques, but out of all (such as CNN, Random Forest etc.), K-Nearest Neighbors(KNN) and Support Vector Machine (SVM) provided the better results this is because of the effective learning rate and quite suitable for the feature extraction. The proposed classification model is built on sophisticated machine learning methods that not only takes care about the syntactical nature of the URL, but also the semantic and lexical meaning of these dynamically changing URLs. The proposed approach is expected to outperform the existing techniques.

KEYWORDS:Malicious URLs, Black-Listing, Machine Learning, URL Features, Cyber Crime, KNN, SVM.

I. INTRODUCTION

The advent of new communication technologies has had tremendous impact in the growth and promotion of business spanning across many applications including online-banking, e-commerce, and social networking. In fact, in today's age it is almost mandatory to have an online presence to run a successful venture. As a result, the importance of the World Wide Web has continuously been increasing.Unfortunately, the technological advancements come coupled with new sophisticated techniques to attack and scam users. Such attacks include rogue websites that sell counterfeit goods, financial fraud by tricking users into revealing sensitive information which eventually lead to theft of money or identity, or even installing malware in the user's system.

There are a wide variety of techniques to implement such attacks, such as explicit hacking attempts, drive-by download, social engineering, phishing, watering hole, man-in-the middle, SQL injections, loss/theft of devices, denial of service, distributed denial of service, and many others.Considering the variety of attacks,potentially new attack types, and the innumerable contexts in which such attacks can appear, it is hard to design robust systems to detect cyber-security breaches.Adversaries who try to get unauthorized access to the confidential data may use malicious URLs and present it as a legitimate URL to naive user. Such URLs that act as a gateway for the unsolicited activities are called as malicious URLs.These malicious URLs can cause unethical activities such as theft of private and confidential data, ransomware installation on the user devices that result in huge loss every year globally.

The phishing website exactly looks the same as an original website. Phishing is an attempt to steal a user's personal information typically through a fraudulent email or website. Almost all phishing sites include the functionality in which users enter sensitive information, such as their personal identification, password, and/or account number.

So, the best way to prevent ending up at Malicious/Phishing site is by integrating supervised learning algorithms such as SVM (Support Vector Machine) and KNN (K-Nearest Neighbors). These algorithms help to detect the authenticity of the website. It performs dynamically by adding Malicious/Phishing URL once captured. The feature set is composed of some features, such as token count, average path token, largest path, largest token, etc.

Malicious/Phishing URL Detection through machine learning typically comprises two steps: first to obtain an appropriate feature representation from the URL, and second, to use this representation of the URL to train machine learning based prediction models.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.089|

||Volume 9, Issue 4, April 2020||

II. RELATED WORK

Malicious URLs host unsolicited content and are used to perpetrate cybercrimes. It is imperative to detect them in a timely manner. Traditionally, this is done through the usage of blacklists, which cannot be exhaustive, and cannot detect newly generated malicious URLs [1]. To address this, recent years have witnessed several efforts to perform Malicious URL Detection using Machine Learning [1]. Damage caused by phishing attacks that target personal user information is increasing. Phishing involves sending an email to a user or inducing a phishing page to steal a user's personal information [2]. This type of attack can be detected by blacklist-based detection techniques; however, these methods have some disadvantages and the numbers of victims have therefore continued to increase. In this paper, we propose a heuristic-based phishing detection technique that uses uniform resource locator (URL) features [2]. The results demonstrate that the proposed technique can detect more than 98.23% of phishing sites [2]. **URL Structure:** A URL is a protocol that is used to indicate the location of data on a network. The URL is composed of theprotocol, subdomain, primary domain, top-level domain (TLD), and path domain. In this study, the subdomain, primary domain, and TLD are collectively referred to as the domain.

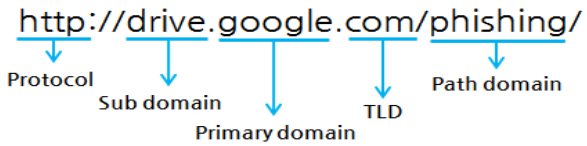

Figure 1: Depicts the individual components of a URL

Malicious URL, a.k.a. malicious website, is a common and serious threat to cybersecurity. Malicious URLs host unsolicited content (spam, phishing, drive-by downloads, etc.) and lure unsuspecting users to become victims of scams (monetary loss, theft of private information, and malware installation), and cause losses of billions of dollars every year [3].This article aims to provide a comprehensive survey and a structural understanding of Malicious URL Detection techniques using machine learning. We present the formal formulation of Malicious URL Detection as a machine learning task, and categorize and review the contributions of literature studies that address different dimensions of this problem (feature representation, algorithm design, etc.) [3].

III. METHODOLOGY

A variety of approaches have been attempted to tackle the problem of Malicious URL Detection. According to the fundamental principles, we categorize them into: (i) Blacklisting or Heuristics, and (ii) Machine Learning approaches.Blacklisting or Heuristic Approaches:Blacklisting approaches are a common and classical technique for detecting malicious URLs, which often maintain a list of URLs that are known to be malicious. Whenever a new URL is visited, a database lookup is performed. If the URL is present in the blacklist, it is considered to be malicious and then a warning will be generated; else it is assumed to be benign.Heuristic approaches are a kind of extension of Blacklist methods, wherein the idea is to create a "blacklist of signatures". Common attacks are identified, and a signature is assigned to this attack type. Intrusion Detection Systems can scan the web pages for such signatures, and raise a flag if some suspicious behaviour is found. These methods have better generalization capabilities than blacklisting, as they have the ability to detect threats in new URLs as well.

Disadvantages of existing system:

- They are not resistant to 0-day attacks.
- Websites may not launch an attack immediately after being visited, and thus may go undetected.

Machine Learning approaches try to analyse the information of a URL and its corresponding websites or webpages, by extracting good feature representations of URLs, and training a prediction model on training data ofboth malicious and benign URLs. There are two-types of features that can be used - static features, and dynamic features. In static analysis, we perform the analysis of a webpage based on information available without executing the URL (i.e., executing JavaScript, or other code). The features extracted include lexical features from the URL string, information about the host, and sometimes even HTML and JavaScript content. Since no execution is required, these methods are safer than the Dynamic approaches.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.089|

||Volume 9, Issue 4, April 2020||

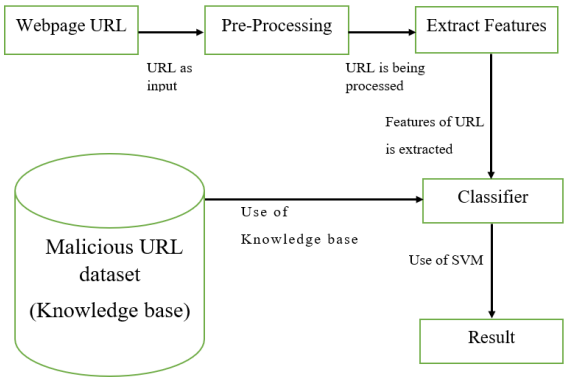

Figure 2: Workflow of Detecting Malicious/Phishing URL

Table 1: Features of the URL

Malicious/Phishing URL detection using machine learning can have immense real-world applications. However, it is

SL. No.	URL Based Features	
	Feature Name	**Description**
1.	Length of host URL	This association rule indicates that if the length of the URL is more than 75, the URL is categorized as a malicious or phishing URL.
2.	Number of Dots and Slashes	The number of the dots in the domain part indicates the number of subdomains used in the URL. Therefore, if the number of the sub domains is more than 3, the URL is classified as malicious/phishing.
3.	Using special character	This rule indicates that when special characters such as dash, underscore, comma, semicolon are part of the input URL, then it is a malicious/phishing URL.
4.	Using IP Address	If an IP Address is part of the URL, it is an indication of someone is trying to access the sensitive information through a phishing attack. So, if a URL contains IP Address, the system will mark it as malign else legitimate.
5.	Age of Domain	The age of the domain can be extracted from a WHOIS database. Therefore, the domain of all legitimate URLs has an age greater than 6 months which indicates that the URL is legitimate.
6.	Average Domain Token Length	Domain registration length should be at least one year as we are aware of the fact that trustworthy domains are regularly paid for several years in advance.
7.	Alexa page Rank	The web page rank is a key indicator which is governed by the factors such as the number of pages visited by the web users as well as the number of the visitors to a website or URL. Therefore, the less the page rank of a web site, the more legitimates the URL.
8.	DNS records	If a URL has to be classified as a legitimate one, it should have a DNS record else it is treated as malign URL.

nontrivial to make it work in practice. Several researchers have proposed architectures for building end-to-end malicious/phishing URL Detection systems and deployed them for real-world utilities. Any machine learning technique typically comprises of two steps: one is to obtain the appropriate feature representation that it could provide the determining insights in finding the Malicious/Phishing URLs, and the second is to use this representation to train a

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.089|

||Volume 9, Issue 4, April 2020||

learning-based prediction mechanism. In the proposed approach, we have provided the feature representation of the URLs. Analogically, Blood of this Process is the features and heart is the machine learning mechanism. Every time the blood passes through the heart the refining will happen. In the same manner features of the URLs will pass through the machine learning engine and then based on the previous learning the classification develops. In our case, we clearly followed the Lexical Analysis Features in addition to the 3rd party feature, geo ranking, altogether we gathered the feature set as shown in the TABLE:1.

Our model is implemented as a Plugin in the Browser. When the user enters the URL into the Browser, pre-processing (tokenization) of the URL takes place, later the features are extracted from the URL.

- The extracted features are fed to the Classifier.
- The Classifier (SVM) uses the Knowledge Base Dataset and analyses the given URL and determines whether the given URL is Malicious or Legitimate.
- Dataset link(https://archive.ics.uci.edu/ml/datasets/phishing+websites)

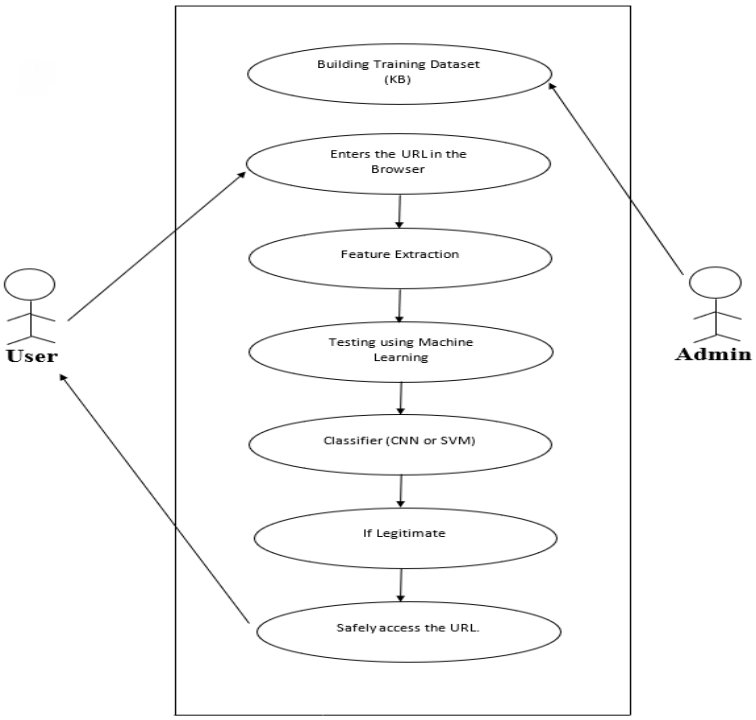

Figure 3: Use Case Diagram

IV. EXPERIMENTAL RESULTS

The presented work is still in its early state. The motive of this paper is to provide a brief about our approach. One assumption is that malicious URLs could be detected by extracting the lexical features. For doing the basic conduction we performed the Classifying method based on the TF - IDF word association. We can support the features that are extracted from the URL bigrams and term frequency and inverse term frequency will give the minimal classifying environment. But the classifying that uses the proposed features is the main task and we completed the preprocessing state. The presented work is an early effort in malicious URL detection, we will be covering the post process the Feature set and give the classifying coefficients which are used as the separating parameters as a future work.To conduct classifier training and evaluation through an experimental dataset, we collected the URLs of phishingand legitimate sites. We gathered 3,000 phishing site URLsfrom PhishTank and 3,000 legitimate site URLs fromDMOZ. The evaluation was conducted using *k*-fold crossvalidation. *K*-fold cross validation divides the input data into*k*; $k - 1$ datasets are used for training, and the remaining is used for validation.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.089|

||Volume 9, Issue 4, April 2020||

V. CONCLUSION

In this paper, we aimed to find the highest performance model using the smallest number of features and the smallest structural parameters (tree number, max tree depth and maxi-mum leaf size) in order to find the least complex but high per-forming classifier. In the social networks, the spam detection task requires a fast response to the new emerging techniques and tricks that hacker's use. Features in this field need to be re-ranked and evaluated periodically since machine learning models need to be built on reliable and validated feature sets to achieve high-quality performance. For our future work, we aim to conduct a systematic analysis to evaluate the feature selection procedure and Support Vector Machine and Deep Neural Network (DNN) classifiers. Malicious URL detection plays a critical role for many cybersecurity applications, and clearly machine learning approaches are a promising direction. Machine learning algorithms and determined that the best classifier was random forest. It showed a high accuracy of 98.23% and a low false-positive rate. The proposed technique can provide security for personal information and reduce damage caused by phishing attacks because it can detect new and temporary phishing sites that evade existing phishing detection techniques, such as the blacklist-based technique. In this work, we have described how a machine can able to judge the URLs based upon the given feature set. Specifically, we described the feature sets and an approach for classifying the given the feature set for malicious URL detection. Certain efforts have to be made in that direction so as to come up with the more robust feature set which can change with respect to the evolving changes.

REFERENCES

[1] Hung Le, Quang Pham, Doyen Sahoo, Steven C.H. Hoi proposed,"URL Net: Learning a URL Representation with Deep Learning for Malicious URL Detection", 2018.
[2] Immadisetti Naga Venkata Durga Naveen, Manamohana K, Rohit Verma, "Detection of Malicious URLs using Machine Learning Techniques", 2019.
[3] Jin-Lee Lee, Dong-Hyun Kim, ChangHoon, Lee proposed, "Heuristic based Approach for Phishing Site Detection Using URL Features", 2015
[4] Doyen Sahoo, Chenghao Liu, Steven C.H. Hoi proposed, "Malicious URL Detection using Machine Learning: A Survey", 2016.
[5] Kurt Thomas, Justin Ma, Vern Paxson, Dawn Song, Chris Grier, "Design and Evaluation of a Real-Time URL Spam Filtering Service", 2014.
[6] Chai-Mei Chen, D.J. Guan, Qun-Kai Su, "Feature set identification for detecting suspicious URLs using Bayesian classification", 2018.
[7] Farhan Douksieh Abdi and Lian Wenjuan, "Malicious URL Detection using Convolutional Neural Network", 2017.
[8] Sadia Afroz and Rachel Greenstadt, "Phishzoo: Detecting phishing websites by looking at them. In Semantic Computing (ICSC)", 2011.