# Webpage Classification for Safer Browsing

[1,] Abhay Pratap Singh , [2]Aniket Kumar Singh, [3]Gangadhar Kannaujiya and [4] Mrs. Surbhi Verma

[1,2,3] Pursuing bachelors degree in Computer Science in Dr. A.P.J Abdul Kalam Technical University, India

[4,]Currently working as an assistant professor at ABES Engineering College, India

Abhay Pratap Singh- Email- abhay.19b121046@abes.ac.in
Aniket Kumar Singh- Email-aniket.19b121061@abes.ac.in
Gangadhar Kannaujiya- Email- gangadhar.19b121020@abes.ac.in
Mrs. Surbhi Verma- Email-surbhi.verma@abes.ac.in

***Abstract*** - Phishing attack is a simplest way to obtain sensitive information from innocent users. Aim of the phishers is to acquire critical information like username, password and bank account details. Cyber security persons are now looking for trustworthy and steady detection techniques for phishing websites detection. This paper deals with technology for detection of phishing URLs by extracting and analyzing various features of legitimate and phishing URLs . Aim of the paper is to detect phishing URLs

*Keyword*
Phishing Attack.

## I. Introduction

Nowadays Phishing becomes a main area of concern for security researchers because it is not difficult to create the fake website which looks so close to legitimate website. Experts can identify fake websites but not all the users can identify the fake website and such users become the victim of phishing attack. Main aim of the attacker is to steal banks account credentials. In United States businesses, there is a loss of US $2billion per year because their clients become victim to phishing

1. In 3rd Microsoft Computing Safer Index Report released in February 2014, it was estimated that the annual worldwide impact of phishing could be as high as $ 5 billion 2. Phishing attacks are becoming successful because lack of user awareness. Since phishing attack exploits the weaknesses found in users, it is very difficult to mitigate them but it is very important to enhance phishing detection techniques. The general method to detect phishing websites by updating blacklisted URLs, Internet Protocol (IP) to the antivirus database which is also known as "blacklist" method. To evade blacklists attackers uses creative techniques to fool users by modifying the URL to appear legitimate via obfuscation and many other simple techniques including: fast-flux, in which proxies are automatically generated to host the web-page algorithmic generation of new URLs; etc. Major drawback of this method is that, it cannot detect zero-hour phishing attack. Heuristic based detection which includes characteristics that are found to exist in phishing attacks in reality and can detect zero-hour phishing attack, but the characteristics are not guaranteed to always exist in such attacks and false positive rate in detection is very high

3. To overcome the drawbacks of blacklist and heuristics based method, many security researchers now focused on machine learning techniques. Technology consists of a many algorithms which requires past data to make a decision or prediction on future data. Using this technique, algorithm will analyze various blacklisted and legitimate URLs and their features to accurately detect the phishing websites including zero- hour phishing websites

## II. FEATURE EXTRACTION

We have implemented program to extract features from URL. Below are the features that we have extracted for detection of phishing URLs.

1) **Presence of IP address in URL:** If IP address present in URL then the feature is set to 1 else set to 0. Most of the benign sites do not use IP address as an URL to download a webpage. Use of IP address in URL indicates that attacker is trying to steal sensitive information.

2) **Presence of @ symbol in URL**: If @ symbol present in URL then the feature is set to 1 else set to 0. Phishers add special symbol @ in the URL leads the browser to ignore everything preceding the "@" symbol .

3) **Number of dots in Hostname**: Phishing URLs have many dots in URL. For example http://shop.fun.amazon.phishing.com,in this URL phishing.com is an actual domain name, whereas use of "amazon" word is to trick users to click on it. Average number of dots in benign URLs is 3. If the number of dots in URLs is more than 3 then the feature is set to 1 else to0.

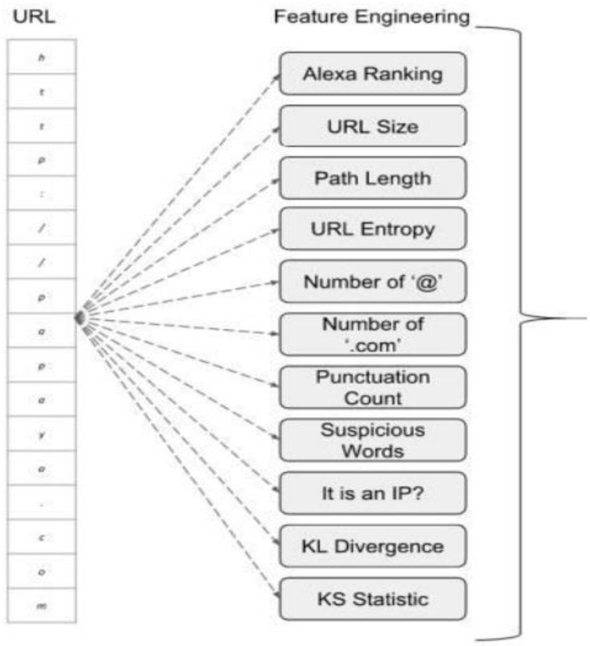
4) **Prefix or Suffix separated by (-) to domain:** If domain name separated by dash (-) symbol then feature is set to 1 else to 0. The dash symbol is rarely used in legitimate URLs. Phishers add dash symbol (-) to the domain name so that users feel that they are dealing with a legitimate webpage. For example Actual site is http://www.onlineamazon.combut phisher can create another fake website like http://www.online-amazon. combut confuse the innocent users.

5) **HTTPS token in URLs:** If HTTPS token presenting URL then the feature is set to 1 else to 0. Phishers may add the "HTTPS". S represent 'Secure' token to the domain part of a URL in order to trick users. For example, http://https-wwwpaypal- it-mpp-home.soft-hair.com [4].

6) **Length of Host name:** Average length of the benign URLs is found to be a 25, If URL’s length is greater than 25 then the feature is set to 1 else to0

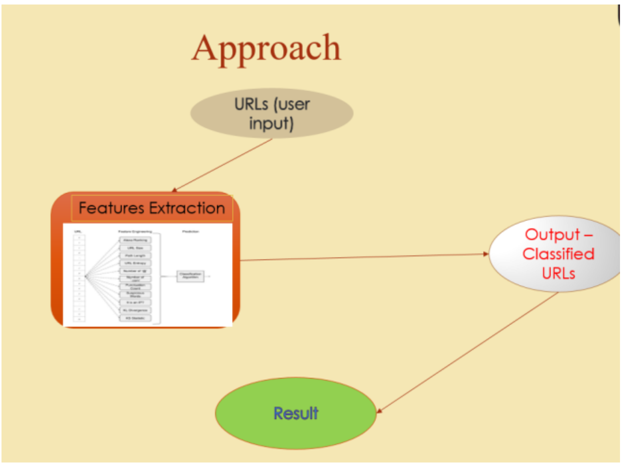
8) **Presence of sensitive words in URL:** Phishing sites use sensitive words in its URL so that users feel that they are dealing with a legitimate webpage. Below are the words that found in many phishing URLs :- 'confirm', 'account', 'banking', 'secure', 'ebyisapi', 'webscr', 'signin',
'mail', 'install', 'toolbar', 'backup', 'paypal', 'password',
'username',etc;

7) **Number of slash in URL:** The number of slashes in benign URLs is found to be a 5; if number of slashes in URL is greater than 5 then the feature is set to1else to 0.



# IV. Methodology

We have used HTML,CSS on Frontend of our project and Java Script as Backend for validation to extract the feature. When the user enter the url’s in the search bar than it will take command to backend(code) from where it will classify whether the url is phishing or bening on the basic of feature extracted from program.



## IV. Conclustion

This paper aims to enhance detection method to detect phishing websites using  technology. Also result shows that classifiers give better performance when we used more data as training data. We are not able to reach 100% accuracy, thus we end up by creating this system, with adequate time and data and tried to get very close to the goal. For future purpose more improvement can be found in the and to find the real time fraud.

## V. References

1. Eloquent javaScript:A modern introduction to programming Author-marjin haverbeke

2. Javascript & Jquery:Interactive front-End web devolpment Author-Jon Duckett

3. www.javatpoint.com

4. www.W3School.com