



IJIRCCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 6, June 2021

ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 7.542



9940 572 462



6381 907 438



ijircce@gmail.com



www.ijircce.com

Smart Eye- Fishy URL Detection Using URL Features and CNN

Nayan Suresh Ambolikar, Atul Madan Thorat, Ashutosh Mahendra Sakhare, Prof. Vijay Sonawane

U.G. Students, Dept. of Computer Engineering, JSPM's BSIOTR, Pune University - Pune, Maharashtra, India

Professor, Dept. of Computer Engineering, JSPM's BSIOTR, Pune University - Pune, Maharashtra, India

ABSTRACT: Phishing is a criminal scheme to steal the user's personal data and other credential information. It is a fraud that acquires victim's confidential information such as password, bank account detail, credit card number, financial username and password etc. and later it can be misused by attacker.

First of all, the phisher has to create a phishing website to lure the victim which seems as legitimate one. Then, host the site on the internet for use of victim's secret information. If victim visits the phishing website, it convinces the victim to enter some confidential information. Phisher then acquires some entered data and later it can be misused by phisher.

We aim to use WhoIs features of URL as the basis of detecting phishing websites. We propose a novel solution, Phishing Detection using Soft Computing and Machine Learning, to efficiently detect phishing web pages using URL and WhoIs features. The convolutional Neural Network is used to train the network and finally detect the site is Phishing or not.

KEYWORDS: Fishy URL Detection, WHOIS features, URL Features, Convolutional Neural Network, URL Length, IP address, Avoiding Phishing Attacks

I. INTRODUCTION

Phishing is defined as the fraudulent acquisition of confidential data by the intended recipients and the misuse of such data. The phishing attack is often done by email. An example of Phishing; as if e-mail appears to be from known web sites, from a user's bank, credit card company, e-mail, or Internet service provider. Generally, personal information such as credit card number or password is asked to update accounts.

These emails contain a URL link that directs users to another website. This site is actually a fake or modified website. When users go to this site, they are asked to enter personal information to be forwarded to the phishing attacker.

PHISHING ATTACKS

The aim is to steal sensitive data such as credit card and login information or to install malicious software on the victim's machine.

Phishing is a common type of cyber-attack that everyone must learn to protect themselves from. Phishing starts with a fake e-mail or other type of transmission designed to attract a victim. In this type of attack, the message appears to come from a trusted source.

In a phishing attack, attackers can use social engineering and other public information resources, including social networks like LinkedIn, Facebook and Twitter, to gather background information about the victim's personal and work history, interests and activities. With this pre-discovery, attackers can identify potential victims' names, job titles and email addresses, information about the names of key employees in their colleagues and organizations.

Phishing is also used to learn someone's password or credit card information. With the help of e-mail prepared as if coming from a bank or official institution, computer users are directed to fake sites.

The common information that is stolen by a phishing attack is listed as follows:

- User account number
- User passwords and user name
- Credit card information

II. RELATED WORK/LITERATURE SURVEY

Jian Mao^{1*}, JingdongBian et al.[1] proposed a system where the key aim is to enable automated page-layout-based phishing detection techniques using machine learning techniques.

The given aggregation analysis mechanism decides page layout similarity, which is used to detect phishing pages. It evaluates four popular machine learning classifiers on their accuracy and the factors affecting their results.

SHAFTI MUHAMMAD ABDULHAMID et al.[2] uses soft computing approach called Artificial Neural Network (ANN) algorithm with confusion matrix analysis for the detection of e-banking phishing websites. The ANN algorithm produces a significant accuracy and reduced false positive rate during detection. This signifies that ANN algorithm with confusion matrix analysis can generate a competitive results that is fit for detecting phishing in e-banking websites.

Neda Abdelhamid et al.[3] experimentally compare large numbers of ML techniques on real phishing datasets and with respect to different metrics. The main purpose of the comparison is to disclose the advantages and disadvantages of ML predictive models and to show their actual performance when it comes to phishing attacks. The experimental results show that Covering approach models are more suitable as anti-phishing solutions, particularly for novice users, since of their simple yet effective knowledge bases in addition to their good phishing detection rate.

Longfei Wu, et al.[4] designed for web phishing attacks on PCs cannot efficiently point out the several phishing attacks on mobile devices. Henceforth, the author presented MobiFish, a novel automated lightweight anti-phishing scheme for mobile platforms. MobiFish verifies the validity of web pages, applications, and persistent accounts by comparing the actual identity to the claimed identity.

Mohammed Nazim Feroz et al.[5] describes an approach that categorizes URLs repeatedly based on their lexical and host based parameters. Clustering is used on the whole dataset and a cluster ID (or label) is calculated for each URL, which in turn is used as a predictive feature by the classification system.

Luong Anh Tuan Nguyen et al.[6] proposing a new approach to detect phishing site by using the features of URL. Mostly, we develop different components from URL and compute a metric for every component. So, the page ranking will be shared with the achieved metrics to decide whether the websites are phishing websites.

III. PROPOSED ARCHITECTURE

The phishing attack is often done by email. An example of Phishing; as if e-mail appears to be from known web sites, from a user's bank, credit card company, e-mail, or Internet service provider. Generally, personal information such as credit card number or password is asked to update accounts.

The system is solution for avoiding such phishing attacks.

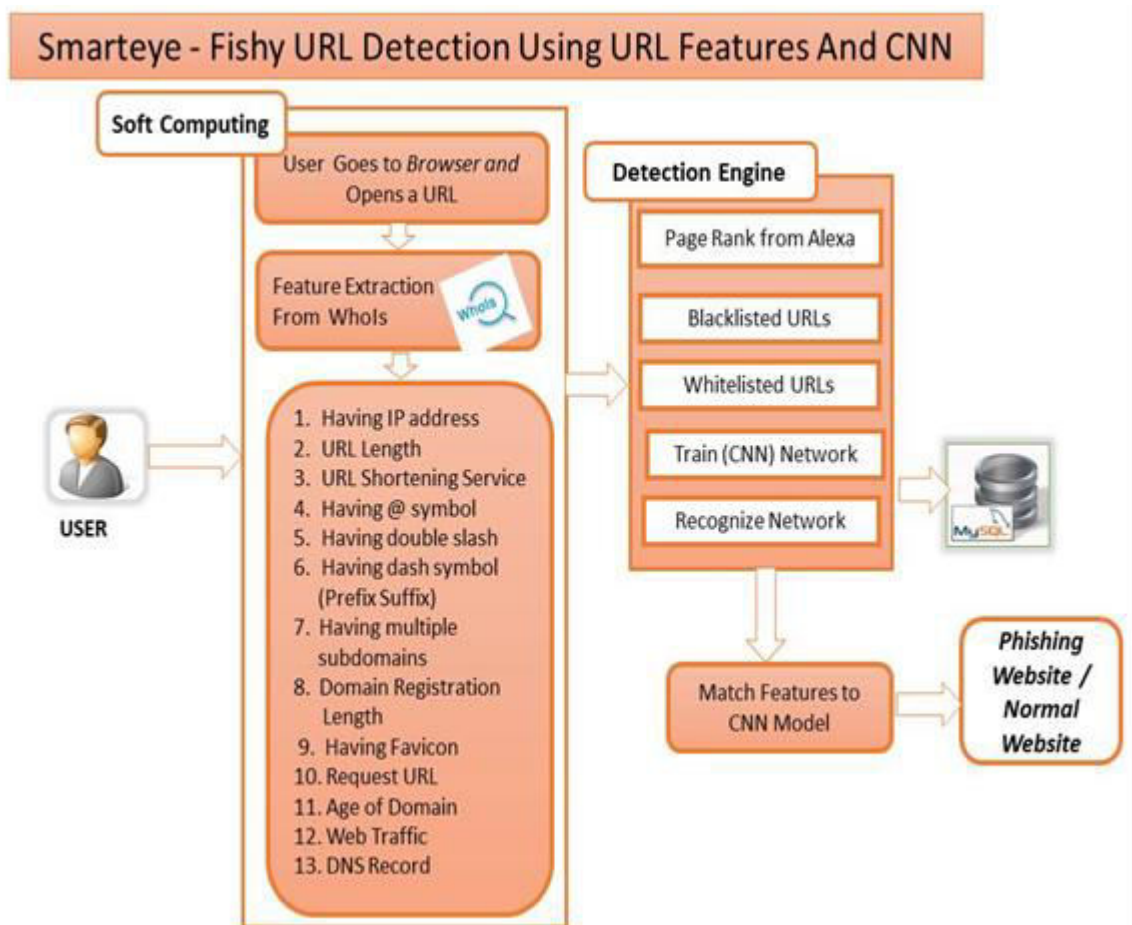


Figure3.1: System Architecture

Whois Domain: WHOIS is a question and reaction convention that is broadly utilized for questioning databases that store the enlisted clients or trustees of an Internet asset, for example, an area name, an IP address square or a self-governing framework, but on the other hand is utilized for a more extensive scope of other data. The convention stores and conveys database content in a comprehensible format. A WHOIS is a way for you to search the public database for information about a specific domain, such as the expiration date, current registrar, registrant information, etc. The following features are extracted from Whois from URLs:

Sr. No.	Features	Significance
1	Having IP Address	If IP address is used in domain name
2	URL Length	Legitimate URLs have length of nearly 75 characters, URL length more than 75 is Phishing sites.
3	Shortening Service	Shortening Service
4	Having @ Symbol	Websites having @ symbol are Phishy in general
5	Double slash redirecting	If there is '/1' then it can be categorized
6	Having Sub Domain	Legitimate Websites use only domain generally upto two level.
7	URL of Anchor	In legitimate websites the anchor tag is connected to the same domain as the source code, Phishy
8	Links in tags	Links in tags lead to some fraudulent websites
9	Abnormal URL	This feature is extracted from Who is Database, Legitimate websites' main identity is in the URL

10	Age of domain	Legitimate websites have an age of sixmonths; websites with more than this agecan be classified as Phishing.
11	Page Rank	Phishing websites will have low page rankdue to lack of links pointing to them.
12	Links Pointing to page	Phishing websites have links pointing tozip files that automatically get downloadedcontaining malware.
13	Favicon	Many existing user agents such as graphicalbrowsers and newsreaders show favicon asa visual reminder of the website identity in the address bar websites.
14	DNS (Domain Name System) Record	If the DNS record is empty or not found then the website is classified as “Phishing”,otherwise it is classified as “Legitimate”.
15	Web Traffic	Web traffic is the amount of data sent and received by visitors to a website. Phishingwebsites will create huge web traffic.
16	Website Traffic	This feature measures the popularity of thewebsite by determining the number of visitors and the number of pages they visit.

Table 3.1: Websites Features

Built Detection Model using Convolution Neural Network: The system can detect the phishing site using Convolution Neural Network (CNN) technique. A CNN consists of an input and an output layer, as well as multiple hidden layers. The hidden layers of a CNN typically consist of Convolutional layers, pooling layers, fully connected layers and normalization layers. CNN will be used to train the data analytics engine for recognizing the phishing site URL.

Avoiding phishing attacks: A whitelist in the context of phishing detection is simply a list of trusted websites.

– **The URL of the trusted site:**

The URL of the trusted site is used to periodically update the information in the database. This is the URL of the site such as “https://signin.ebay.com”.

– **The domain of the site:**

The domain of the trusted site is the domain of the URL such as “signin.ebay.com” and is used to determine whether the current page displayed in the browser is on the whitelist or not.

– **The title of the site:**

The title of the trusted site is the page title of the site such as “Welcome to eBay” and can be used to speed up the matching potential of phishing site titles with titles in the whitelist Database.

– **Alexa Ranking**

In case your site is ranked relative to other sites, changes in traffic to other sites affect your site’s rank. Every day, Alexa estimates the average daily visitors and page views to every site over the past 3 months. The site with the highest combination of visitors and page views over the past 3 months is ranked 1. As phishing websites live for a short period of time, they may not be recognized by the Alexa database (Alexa the Web Information Company., 1996). If the domain has no traffic or is not recognized by the Alexa database, it is classified as “Phishing”. Otherwise, it is classified as “Suspicious”.

IV. SIMULATION RESULTS

The simulation studies involve the deterministic Phishing Detection using Soft Computing and Machine Learning, to efficiently detect phishing web pages using URL and WhoIs features. The convolutional Neural Network is used to train the network and finally detect the site is Phishing or not.

The main modules involved in the system are :

1. Whois feature extraction of the input website.
2. Training
3. Avoiding phishing attacks

ALGORITHM DETAILS

Convolutional Neural Networks (CNNs)

Convolution Neural Network Traditional feature learning methods rely on semantic labels of images as supervision. They usually assume that the tags are evenly exclusive and thus do not point out towards the complication of labels. The learned features endow explicit semantic relations with words. We also develop a novel cross-modal feature that can both represent visual and textual contents. CNN is a method of categorizing the images as a part of deep learning. In which we apply a single neural network to the full image. The steps in CNN are as follows: convolution, subsampling, activation and full connectedness.

Step 1: Convolution it is the primary layers that accept an input signal are called convolution filters. Convolution is a procedure where the network tries to tag the input signal by referring to what it has learned in the past.

Step 2: Subsampling Inputs from the convolution layer can be smoothed to decrease the sensitivity of the filters to noise and variations. This smoothing procedure is labeled as sub-sampling, and can be attained by taking averages or considering the maximum over a sample of the signal.

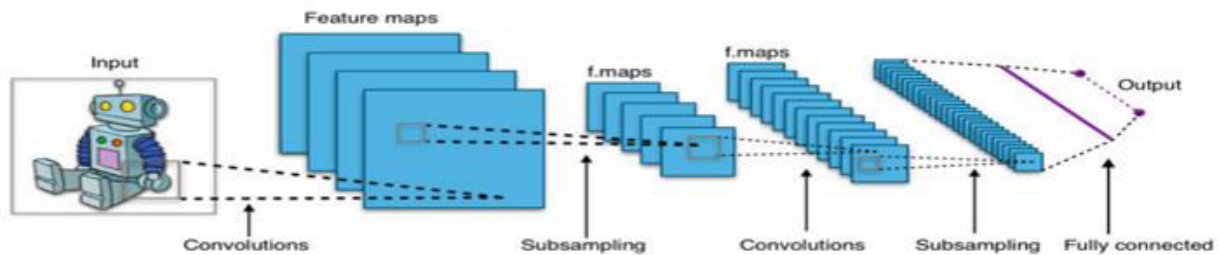


Figure 4.1: Working of Convolutional Neural Network Algorithm

Step 3: Activation the activation layer manages the signal flows from one layer to the subsequent Output signals which are strongly connected with past references would activate more neurons, enabling signals to be propagated more efficiently for identification.

Step 4: Fully connected the final layers in the network are fully connected, such that the neurons of preceding layers are connected to every neuron in subsequent layers. This imitates high Level reasoning where all feasible path ways from the input to output is measured.

V. CONCLUSION AND FUTURE WORK

Conclusion:-

Phishing is a criminal scheme to steal the user's personal data and other credential information. It is a fraud that acquires victim's confidential information such as password, bank account detail, credit card number, financial username and password etc. and later it can be misused by attacker.

We propose a novel solution, Phishing Detection using Soft Computing and Machine Learning, to efficiently detect phishing web pages using URL and CSS features. Features are extracted for Blacklisted and white listed URL features are used as dataset for machine learning algorithms.

Future Work:-

The software designed can also be integrated within Software Application, Web application or Plugins of the Web Browsers

REFERENCES

1. S. Jahan, "Human Resources Information System (HRIS): A Theoretical Perspective", Journal of Human Resource and Sustainability Studies, Vol.2 No.2, Article ID:46129,2014.
2. Neda Abdelhamid, Fadi Thabtah, Hussein Abdel-jaber "Phishing Detection: A Recent Intelligent Machine Learning Comparison based on Models Content and Features" IEEE 2017.
3. Longfei Wu, Xiaojiang Du, and Jie Wu "MobiFish: A Lightweight Anti-Phishing Scheme for Mobile Phones" IEEE 2014.
4. Longfei Wu, Xiaojiang Du, and Jie Wu, "Effective Defense Schemes for Phishing Attacks on Mobile Computing Platforms" IEEE 2015.
5. Guang-Gang Geng, Zhi-Wei Yan, Yu Zeng and Xiao-Bo Jin "RRPhish- Anti-Phishing via Mining Brand Resources Request" 2018 IEEE International Conference on Consumer Electronics (ICCE).
6. Sadia Afroz and Rachel Greenstadt "PhishZoo: Detecting Phishing Websites By Looking at Them" IEEE 2011.
7. Muhammet Baykara and Zahit Ziya Gurel "Detection of phishing attacks" IEEE 2018.
8. Mohammed Nazim Feroz, Susan Mengel "Phishing URL detection using URL Ranking" International Congress on Big Data 2015 IEEE.
9. Luong Anh Tuan Nguyen[†], Ba Lam To[†], Huu Khuong Nguyen[†] and Minh Hoang Nguyen*[†] Faculty of Information Technology "Detecting Phishing Web sites: A Heuristic URL-Based Approach" International Conference on Advanced Technologies for Communications 2013.
10. Ji Hua 1,2, Zhang Huaxiang 1,2 "Analysis on the Content Features and Their Correlation of Web Pages for Spam Detection" IEEE 2015.
11. Samuel Marchal, Jérôme François, Radu State, and Thomas Engel "PhishStorm: Detecting Phishing With Streaming Analytics" IEEE TRANSACTIONS ON NETWORK AND SERVICE MANAGEMENT, 2014.
12. Luong Anh Tuan Nguyen¹, Ba Lam To², Huu Khuong Nguyen¹ and Minh Hoang Nguyen³¹ Faculty of Information Technology "A Novel Approach for Phishing Detection Using URL-Based Heuristic" IEEE 2014.
13. Jian Mao^{1,2}, Pei Li¹, Kun Li¹, Tao Wei³, and Zhenkai Liang⁴ "BaitAlarm: Detecting Phishing Sites Using Similarity in Fundamental Visual Features" 5th International Conference on Intelligent Networking and Collaborative Systems 2013.
14. T. G. Gregory Paul and T. Gireesh Kumar, A Framework for Dynamic Malware Analysis Based on Behavior Artifacts. Singapore: Springer Singapore, 2017.
15. Mohammad, R., M., Thabtah, F., and McCluskey, L., 2014 Predicting phishing websites based on self-structuring neural network. Neural Computing and Applications.
16. Kaveh, A., "Cuckoo search optimization," Advances in Metaheuristic Algorithms for Optimal Design of Structures, 2017.



INNO  **SPACE**
SJIF Scientific Journal Impact Factor
Impact Factor: 7.542



ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

 **9940 572 462**  **6381 907 438**  **ijircce@gmail.com**



www.ijircce.com

Scan to save the contact details