JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Detection OF Phishing Websites using ML

Sairaj Jadhav
CS, G.H.Raisoni College of Engineering and Management.
{Affiliated To SPPU}
Pune, India
sairaj.jadhav.cs@ghrcem.raisoni.net

Safin Tamboli
CS, G.H.Raisoni College of Engineering and Management.
{Affiliated To SPPU}
Pune, India
safin.tamboli.cs@ghrcem.raisoni.net

Shoeb Akthar Shah
CS, G.H.Raisoni College of Engineering and Management.
{Affiliated To SPPU}
Pune, India
shoeb.shah.cs@ghrcem.raisoni.net

Sarthak Pawar
CS, G.H.Raisoni College of Engineering and Management.
{Affiliated To SPPU}
Pune, India
sarthak.pawar.cs@ghrcem.raisoni.net

Prof. Amol Rindhe
CS, G.H.Raisoni College of Engineering and Management.
{Affiliated To SPPU}
Pune, India
amolrindhe45.cs@gmail.com

Abstract— **Since mobile devices have become so common, there is a trend toward moving practically all offline activity online. Due to the anonymity of the Internet, this breaks several security laws even though it simplifies our daily lives. The simplest method for obtaining sensitive information from unwitting users is through phishing attacks. Phishers seek to get private data, including usernames, passwords, and bank account details. Cybersecurity experts are searching for consistent and dependable methods of detecting phishing websites. In this research, numerous properties of both genuine and phishing URLs are extracted and analyzed in order to detect phishing URLs. Phishing websites can be recognized using decision trees, random forests, and support vector machine algorithms.**

***Keywords*— cybersecurity, phishing, machine learning, website classification**

1. INTRODUCTION

In our daily life, we do a lot of work on digital platforms. In many ways, the use of computers and the Internet make our work and personal lives easier. This allows us to quickly complete our processes and operations in areas such as trade, health, education, communication, banking, aviation, research, engineering, entertainment and public services. We live in a technical world and with more and more advances in technology, we face some serious problems such as external phishing or hackers getting hold of users or customer information by creating fake websites that have a general resemblance to the original website. These attackers can steal bank credentials and various data formats related to users' mail and devices. Since phishing attacks are more successful due to the lack of user awareness, they are more difficult to counter, so it is necessary to develop phishing techniques. Phishing attacks are the easiest way to get sensitive information from innocent users. Phishers aim to obtain sensitive information such as usernames, passwords and bank account information. Everyone is now looking for a reliable and consistent detection method to identify phishing websites.

This project ideally deals with machine learning technology to detect phishing URLs by extracting and analyzing various features of legitimate and phishing URLs. Decision trees, random forests, and support vector machines are algorithms used to identify phishing websites.

The main goal of this paper is to find an effective way to prevent real-time phishing attacks. It shows the basic life cycle of a phishing attack as an entry point when a user clicks on a phishing link and uses technical techniques to detect phishing links and alert users. In addition to commonly used blacklist recognition and matching techniques, this paper provides an in-depth description of machine learning-based URL detection technology. This paper presents state-of-the-art solutions, compares and analyzes the challenges and limitations of each solution, and provides research directions and ideas for future solutions. The major aspects of this paper are as follows:

- The fishing life cycle to specifically address the problem of phishing.
- Search major databases and information sources for phishing detection websites.
- It is a machine learning-of-the-art based solution for detecting phishing websites.

Researchers and data analysts have been using machine learning for years because of its comparable performance in terms of data accuracy and precision. analysis. In addition, ML-based algorithms are more intuitive due to the simplicity of tracking how data is generated and its inner workings. Made by hand the feature is risky and highly database dependent. So, recently, researchers have focused on database features that extract features based on URL text. Simply put, researchers adapt neural networks to extract rich characters/words from URLs to show valuable information. Our research focuses on data-based features by using neural network-based models that consider domain- and path-based features. Then, we compare our results with previous papers and summarize ideas for better detection systems.

Statistics from previous work have shown that there is fishing URL is getting more attention lately. However, URL parsing is not an easy search area because most URLs are generated randomly and are informative but difficult to research. Therefore, our research focuses on finding phishing URLs to gather as much information as you can find feature-rich information. The presented system works in two phases where we extract different features of URLs and then using these features a web application is developed for the users to detect any URL that they may think is phishing.

2. LITERATURE SURVEY

This section discussed some of the techniques based on lists, rules, visual similarity, and machine learning.

A. A. List-Based Phishing Detection Systems

This system uses two lists to classify phishing and non-phishing websites. These are called whitelists and blacklists. The white list includes websites that are safe and legitimate, while the black list includes websites that are classified as phishing. Researchers use whitelists to identify fishing grounds. In the search, access to the website is only possible if the URL is whitelisted. Another method is blacklisting. In addition to programs such as Google Safe Browsing API and Phish Net, there are also several studies using blacklists in the literature. In a blacklist-based system, the URL is checked against the list, and if it is not on the list, the URL can be accessed. The biggest weakness of this system is that small changes in URLs prevent them from matching in the list. In addition, the latest attacks, known as zero-day attacks, cannot be caught by this defense system.

Several techniques have been developed for detecting phishing websites. Some of the most common techniques are:

1. Blacklisting: This technique involves maintaining a list of known phishing websites and blocking access to them. This is a reactive approach to phishing detection, as it relies on detecting known phishing websites. However, this approach is not effective against new or unknown phishing websites.
2. Machine Learning: Machine learning techniques can be used to analyze website features and identify phishing websites. This approach involves training a machine learning model on a dataset of known phishing and legitimate websites and then using the model to classify new websites as phishing or legitimate. This approach has shown promising results, but it requires a large amount of data and can be computationally expensive.
3. URL Analysis: URL analysis involves analysing the URL of a website to detect phishing attempts. This technique looks for characteristics such as the use of IP addresses instead of domain names, the presence of suspicious keywords, and the use of long and complex URLs. This approach can be effective in detecting phishing websites, but it is not fool proof.
4. Content Analysis: Content analysis involves analysing the content of a website to detect phishing attempts. This technique looks for characteristics such as the presence of login forms, the use of logos and branding, and the presence of suspicious links. This approach can be effective in detecting phishing websites, but it can be time-consuming and may require manual analysis.

B. Machine Learning-Based Phishing Detection Systems

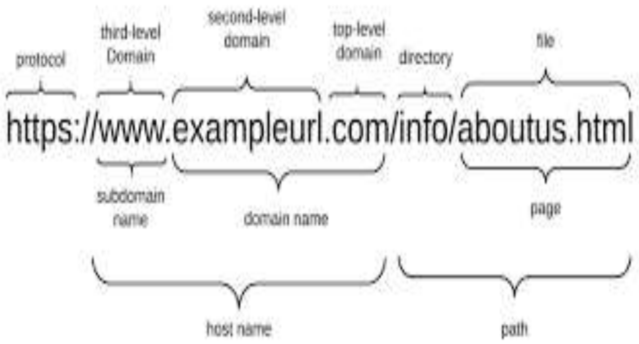
In the machine learning-based phishing detection system, the detection of phishing websites is based on the classification of special features using several artificial intelligence techniques. Author [4 & 5] URL attribute, domain name, website attribute or website content, etc. It is created by collecting in different categories like User Security has become popular because of its dynamic structure, especially to detect anomalies on the web page. In the paper written by the author [6], it was observed that a higher level of accuracy could be achieved by reviewing previous studies and using different features. Unlike previous studies, the new study is based on features selected and coded from a larger number of features. 58 features were identified by URL analysis. Accuracy rate and training time of different algorithm models compared with machine learning methods.

Machine learning techniques have shown great potential in detecting phishing websites by analysing website features and identifying patterns that distinguish them from legitimate websites.

1. Feature-based Classification: Feature-based classification involves extracting features from website content, such as URL structure, website content, HTML tags, images, and other attributes, and using these features to train a machine learning model to classify websites as either phishing or legitimate. This approach has shown promising results in research studies, with accuracy rates ranging from 80% to 99%. However, this approach requires a large amount of training data and feature engineering expertise.
2. Text-based Classification: Text-based classification involves analysing the textual content of a website, such as the text of a login form or a pop-up message, to detect phishing attempts. This approach has been shown to be effective in detecting phishing websites that use social engineering techniques to trick users into providing sensitive information. However, this approach can be susceptible to text obfuscation techniques used by attackers to hide the true intent of the website.
3. Neural Networks: Neural networks are a subset of machine learning algorithms that have been shown to be effective in detecting phishing websites. Neural networks can be used to analyse website features and classify websites as phishing or legitimate based on patterns in the data. This approach has been shown to be highly accurate, with some studies reporting accuracy rates of over 99%. However, this approach requires significant computational resources and expertise in training and tuning neural network models.
4. Ensemble Learning: Ensemble learning involves combining multiple machine learning models to improve the accuracy and robustness of phishing detection. Ensemble learning approaches can be used to combine multiple feature-based or text-based classification models or to combine different machine learning algorithms, such as neural networks and decision trees. This approach has been shown to be effective in improving detection accuracy and reducing false positives.

3. RESEARCH METHODOLOGY

In this post, we aim to implement a phishing detection system by analyzing website URLs. URL is a complex string that represents a syntactic and semantic expression for a resource on the Internet. In more detail, the URL structure is shown in Figure 3. In its most basic form: <protocol>: //<hostname> <URL> is detailed as follows in its complex form.



A.Fig. 3. **From [3]**

B. Domain, subdomain, Top Level Domain (TLD), protocol, directory, filename, path, and request fields allow you to create different URLs. These related fields in phishing URLs are different from legitimate fields on web pages. Therefore, the URL plays an important role in detecting phishing attacks, especially for quickly classifying a website.

In the literature review by the author [7], it was observed that effective features extracted from URLs improve classification accuracy. In addition, the use of third-party services, site layout, CSS, content, meta data, etc. feature can also improve accuracy. However, these features will increase the classification time of new websites that need to be classified. It is expected that the proposed model, which is only trained with URL derived features, will cluster faster than other models. Given this information, only URL analysis is planned in the study. Therefore, in machine learning, the results of feature classification obtained by different algorithms are compared. In addition, the results of another study with the same database compared to the results of the current study.

B. Datasets

1) Phistank.com is a site where phishing URLs are found and accessed via API calls. This is an organization whose data is used by companies such as Yahoo Mail, McAfee, APWG, Mozilla, Opera, Kaspersky and Avira. In the literature review, it was observed that the phishing data used in the machine learning method was generally obtained from Phistank.com. Does the classification need the previous website address? It also provides information on positive/negative classification (phishing/non-phishing). However, it does not store website content; so it's a good resource for URL-based analysis.

2) This article uses open source and accessible databases. We prefer open databases for comparative studies. Three databases are used in this paper, and the researcher named the system Catch Phish [8]. The first of these databases: legitimate sites from the Alexa database and phishing sites from Phish Tank. Second: The legitimate site of public surfing and fishing from Phish Tank. Third: [9] Legitimate sites from regular crawling and Alexa database, Phishing site from Phish Tank. The number of URLs in this database is given in Table I.

TABLE I.

Categories	Dataset-1	Dataset-2	Dataset-3
Phishing	40,666	40,768	40,678
Not Phishing	43,178	42,820	85,809
Total	83,865	82,188	128,07

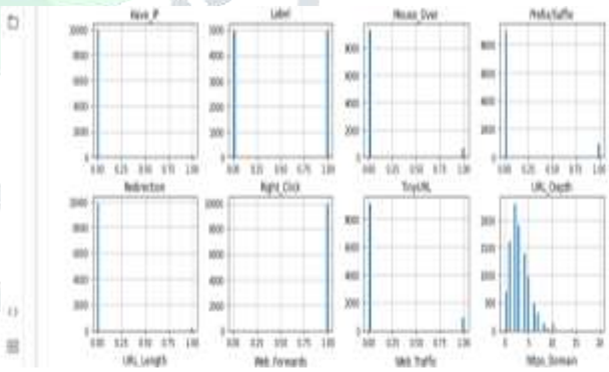
4. DATA ANALYSIS

Address Bar-Based Features:

Many features can be extracted that can be considered as address bar base features. Out of them, below mentioned were considered for this project.

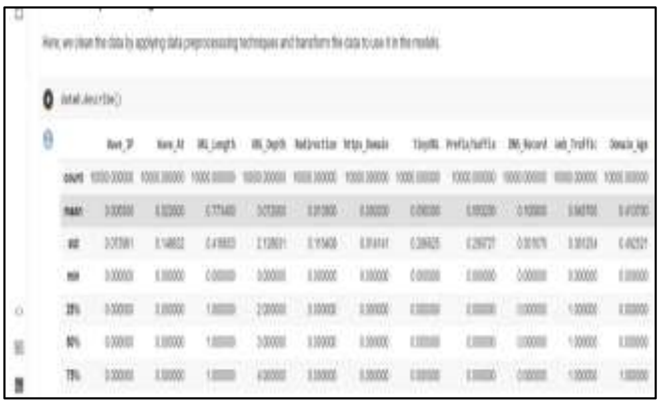
1. Domain of URL
2. IP Address in URL
3. "@" Symbol in URL
4. Length of URL
5. Depth of URL
6. Redirection "//" in the URL
7. "http/https" in the Domain name
8. Using URL Shortening Services “Tiny URL”
9. Prefix or Suffix "-" in Domain

Visualizing the Data-



Data Preprocessing and EDA-

Applying Data Preprocessing and transforming the data to use in the model:



Machine Learning Models & Training-

From the dataset above, it is clear that this is a supervised machine-learning task. There are two major types of supervised machine learning problems, called classification and regression. This data set comes under a classification problem, as the input URL is classified as phishing (1) or legitimate (0). The supervised machine learning models (classification) considered to train the dataset in this notebook are:

1. Decision Tree
2. Gradient Classifier
3. Random Forest
4. Multilayer Perceptrons
5. XGBoost
6. Autoencoder Neural Network
7. Support Vector Machines

Final Results and Outcomes –

On the final side of the project, we used Gradient Classifier to develop a web application that is able to deliver a quick URL detection of whether the site is phishing or not and also whether is it safe to use.

At last, we divided our web application into various web pages as follows:

1. Login Page
2. Dataset Uploading
3. Training Dataset and Preview
4. URL Prediction
5. Results
6. Performance Analysis
7. Chart

The following are the final outcomes of the project presented-

5. FUTURE SCOPE

- Creating a safe user-friendly environment that can detect illegitimate activities.
- It is possible to report and block a hacker using phishing website URLs and tracing the location of such anonymous hackers.
- Awareness can be created among users by displaying a certain type of Phishing URLs available or causing more harm to our system like zero-hour phishing websites.
- An AI System can be developed to detect phishing URLs.

6. CONCLUSIONS

In recent years, due to the evolving technologies on networking not only for traditional web applications but also for mobile and social networking tools, phishing attacks have become one of the important threats in cyberspace. Although most security attacks target system vulnerabilities, phishing exploits the vulnerabilities of human end-users. Therefore, the main defense form for the companies is informing the employees about this type of attack. However, security managers can get some additional protection mechanisms that can be executed either as a decision support system for the user or as a prevention mechanism on the servers.

In this paper, we aimed to implement a phishing detection system by using some machine learning algorithms specifically Random Forest Algorithm and RNN. The proposed systems are tested with some recent datasets in the literature and reached results are compared with the newest works in the literature. The comparison results show that the proposed systems enhance the efficiency of phishing detection and reach very good accuracy rates. As future works, firstly, it aims to create a new and huge dataset for URL-based Phishing Detection Systems to create a safe, user-friendly environment that can detect illegitimate activities.

It is possible to report and block a hacker using a phishing website URL and tracing the location of such anonymous hackers as suggested by **Author [10]**. Awareness can be created among users by displaying a certain type of Phishing URLs available or causing more harm to our system like zero-hour phishing websites.

REFERENCES

[1] Jain A.K., Gupta B.B. "PHISH-SAFE: URL Features-Based Phishing Detection System Using Machine Learning", Cyber Security. Advances in Intelligent Systems and Computing, vol. 729, 2018, https://doi. org/10.1007/978-981-10-8536-9_44

[2] Anti-Phishing Working Group (APWG), https://docs.apwg.org//reports/apwg_trends_report_q4_2019. Pdf

[3] Purbay M., Kumar D, "Split Behaviour of Supervised Machine Learning Algorithms for Phishing URL Detection", Lecture Notes in Electrical Engineering, vol. 683, 2021, https://doi.org/10.1007/978-981- 15-6840-4_40

[4] Gandotra E., Gupta D, "An Efficient Approach for Phishing Detection using Machine Learning", Algorithms for Intelligent Systems, Springer, Singapore, 2021, https://doi.org/10.1007/978-981-15-8711-5_ 12

[5] Hung Le, Quang Pham, Doyen Sahoo, and Steven C.H. Hoi, "URL Net: Learning a URL Representation with Deep Learning for Malicious URL Detection", Conference'17, Washington, DC, USA, arXiv:1802.03162, July 2017.

[6] Hong J., Kim T., Liu J., Park N., Kim SW, "Phishing URL Detection with Lexical Features and Blacklisted Domains", Autonomous Secure Cyber Systems. Springer, https://doi.org/10.1007/978-3-030-33432- 1_12.

[7] J. Kumar, A. Santhanavijayan, B. Janet, B. Rajendran, and B. S. Indhumathi, "Phishing Website Classification and Detection Using Machine Learning," 2020 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2020, pp. 1–6, 10.1109/ ICCCI48352.2020.9104161w

[8] G. Karatas, O. Demir and O. K. Sahingoz, "Deep Learning in Intrusion Detection Systems," 2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT), ANKARA, Turkey, 2018, pp. 113-116, doi: 10.1109/IBIGDELFT.2018.8625278

[9] S. C. Jeeva and E. B. Rajsingh, "Intelligent phishing URL detection using association rule mining," Human-centric Computing and Information Sciences, vol. 6, no. 1, Oct. 2016.

[10] R. Islam and J. Abawajy, "A multi-tier phishing detection and filtering approach," Journal of Network and Computer Applications, vol. 36, no. 1, pp. 324–335, 2013.

[11] I. Arnaldo, A. Arun, and S. Kyathanahalli, "Acquire, adapt and anticipate: continuous learning to block malicious domains," Proc. International Conference on Big Data, 2018. DOI:10.1109/BigData.2018.8622197

[12] M. Trivesan, and I. Drago, "Robust URL classification with generative adversarial networks," in Journal of ACM SIGMETRICS Performance Evaluation Review, vol.46, no.3, pp. 143-146, 2018. DOI:10.1145/3308897.3308959

[13] A. Anand, K. Gorde, J. R. A. Moniz, N. Park, T. Chakraboty, and B. Chu, "Phishing URL detection with oversampling based on text generative adversarial networks," Proc. International Conference on Big Data, pp.1168-1177, 2018. DOI:10.1109/BigData.2018.8622547

[14] S. Shivangi, P. Debnath, K. Sajeevan, and D. Annapurna, "Chrome extension for malicious URLs detection in social media applications using artificial neural networks and long short term memory networks," Proc. 18t h International Conferences on Advances in Computing, Communications and Informatics, pp.1993-1997, 2018. DOI:10.1109/ICACCI.2018.8554647

[15] A. Vazhayil, R. Vinayakumar, and K. P. Soman, "Comparative study of the detection of malicious URLs using shallow and deep networks," Proc. 9t h International Conference on Computing, Communication and Networking Technologies, pp.1-6, July, 2018. DOI:10.1109/ICCCNT.2018.8494159

[16] A. C. Bahnsen, I. Torroledo, D. Camacho, and S. Villegas, "DeepPhish: simulating malicious AI," in APWG Symposium on Electronic Crime Research, 2018.

[17] Y. Shi, G. Chen, and J. Li, "Malicious domain name detection based on extreme machine learning," in Journal of Neural Processing Letters, vol.48, pp.1347-1357, 2018. DOI:10.1007/s11063-017-9666-7

[18] H. Le, Q. Pham, D. Sahoo, and S. C. H. Hoi, "URLNet: learning a URL representation with deep learning for malicious URL detection," in ArXiv, vol.abs/1802.03162, 2017.

[19] A. C. Bahnsen, and E. C. Bohorquez, "Classifying phishing URLs using recurrent neural networks," in APWG Symposium on Electronic Crime Research, pp.1-8, 2017. DOI:10.1109/ECRIME.2017.7945048

[20] D. Sahoo, C. Liu, and S. C. H. Hoi, "Malicious URL detection using machine learning : a survey," in ArXiv, vol.abs/1701.07179v1, 2017.

[21] A. Hodzic, J. Kevric, "Comparison of machine learning techniques in phishing website classification," Proc. International Conference on Economic and Social Studies (ICESoS'16), vol.3, pp.249-256, 2016.

[22] M. Dadkhah, S. Shamshirband, A. Wahab, "A hybrid approach for phishing web site detection," in the Electronic Library, vol.34, no.6, pp.927-944, 2016.

[23] M. N. Feroz, S. Mengel, "Phishing URL detection using URL ranking," Proc. 2015 IEEE International Congress on Big Data, pp.635-638, 2015.

[24] S. Marchal, J. Francois, R. State, and T. Engel, "PhishStorm: detecting phishing with streaming analytics," in IEEE Transactions on Network and Service Management, vol.11, no.4, pp.458-471, 2014.

[25] E. Sorio, A. Bartoli, E. Medvet, "Detection of hidden fraudulent URLs within trusted sites using lexical features," Proc. 2013 International Conference on Availability, Reliability and Security, 2013.

[26] P. Prakash, M. Kumar, R. R. Kompella, and M. Gupta, "Phishnet: predictive blacklisting to detect phishing attacks," Proc. IEEE INFOCOM, 2010, pp.1-5, 2010.

[27] Y. Cao, W. Han, and Y. Le, "Anti-phishing based on automated individual white-list," Proc. the 4t h ACM Workshop on Digital Identity Management, pp.51–60, 2008.

[28] M. Sharifi and S. H. Siadati, "A phishing sites blacklist generator," Proc. IEEE/ACS International Conference on Computer Systems and Applications, pp. 840-843, 2008.

[29] J. Kang and D. Lee, "Advanced white-list approach for preventing access to phishing sites," Proc. International Conference on Convergence Information Technology (ICCIT 2007), pp.491-496, 2007.

[30] L. Wenyin, G. Huang, L. Xiao Yue, Z. Min, X. Deng, "Detection of phishing webpages based on visual similarity," in Special interest tracks and posters of the 14th International Conference on World Wide Web, pp. 1060- 1061, 2005.