International Journal for Innovative Engineering and Management Research

PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL

www.ijiemr.org

IJIEMR Transactions, online available on 07th Sept 2023. Link :http://www.ijiemr.org/downloads.php?vol=Volume-12&issue=Issue 09

10.48047/IJIEMR/V12/ISSUE 09/05

Title DETECTION OF MORE HARMFUL PHISHING WEBSITES WITH MACHINE LEARNING METHODS

Volume 12, ISSUE 09, Pages: 36-45

Paper Authors **RAMIREDDY HIMABINDU, N SURENDRA, V SUBHASINI**

International Journal for Innovative Engineering and Management Research

PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL

www.ijiemr.org

DETECTION OF MORE HARMFUL PHISHING WEBSITES WITH MACHINE LEARNING METHODS

RAMIREDDY HIMABINDU[1], N SURENDRA[2], V SUBHASINI[3]

[1]M.Tech Student, Dept of CSE, MJR College of Engineering & Technology, Piler, AP, India.
[2]Assistant Professor, Dept of CSE, MJR College of Engineering & Technology, Piler, AP, India.
[3]Assistant Professor & HOD of CSE, MJR College of Engineering & Technology, Piler, AP, India.

ABSTRACT: Cybercriminals have become increasingly sophisticated in their methods of phishing attacks. An attacker uses social media platforms or emails to send fake messages as part of a social engineering attack. Users' information can be stolen or malicious software installed when a phishing attack is conducted. A phishing message can appear to be legitimate to a user, making it difficult to detect. A phishing URL could be included in this message, making it possible for even an expert to be victimized. This URL can be used by attackers to create a fake website that steals a victim's login and payment details. Phishing attacks can be detected without the help of experts thanks to advance research and engineering. There is no comprehensive survey of the methods for detecting phishing in HTML and URLs, despite the fact that many papers discuss these methods. With six dissimilar classification procedures based on eleven prearranged structures, suggest a novel method to detect phishing websites via Internet URLs and domain names. As a result of the proposed method, feature extraction is simplified, and processing overhead is reduced while URLs and domain names are also considered as part of feature extraction, which improves overall performance. A Random Forest algorithm was used to illustrate which classification results had the maximum correctness percentage out of six possible classification consequences. In this article, a dataset of 33,918 data points is used, of which 12,134 data points are free of phishing internet sites and 20,614 data points contain phishing websites. The data points are labeled using eleven specified attributes. The proposed method is capable of detecting phishing websites with an accuracy of 99.20% according to our experimental results. In this study, RF descriptors with SVM representations were shown to accurately mark phishing web pages.

Keywords: Phishing attacks, HTML, URL, Artificial Intelligence, Machine Learning, Natural Language Processing, Intelligent Detection.

INTRODUCTION

Phishing tricks have emerged as a serious concern for both online finance and e-commerce users. The act of phishing involves fraudulent attempts to deceive individuals into close-fitting delicate data such as usernames, passwords, and credit card facts [1]. According to recent data from Google, the number of registered phishing sites has reached a staggering 2,135,214 as of January 18 2022 [2]. Related on 2021, this rate mounted through 29%.

Furthermore, according to IBM's report, phishing is identified as the another most costly basis of information breaches. A breach prompted by the occurrence can have significant financial implications for firms, with an average cost of $4 [3].

Furthermore, despite considerable recent advancements in technologies for the identification and prevention of phishing websites, this issue still results in enormous losses every year [4], [9]. The first category includes methods that focus on detecting and blocking phishing websites.

The first is a list of questionable websites that can be found on URL blacklists maintained by hosting companies, computer antivirus developers, or other reputable organizations. [10], [15]. Every time a web browser reads a page, the Uniform Resource Locator (URL) blacklists mechanism checks blacklists to determine if the URL being used is included. You will be informed and the required steps will be made if your URL is found on any blacklists.

The second category involves the utilization of machine learning algorithms on websites, rather than simply relying on a predetermined list of criteria. [16], [17]. These mechanisms, including URLs, Hypertext Markup Language code (HTML), and page content, are crucial for accurately identifying phishing websites.

The need for further exploration and refinement of categorization and tagging features is warranted in order to improve the effectiveness and accuracy of these techniques. Because attackers modify URLs and HTML, it creates inconsistencies between genuine websites and phishing websites. In order to detect phishing websites, researchers have proposed various methods that focus on analyzing the URL and the HTML content of a website.

One of the main problems with options based on artificial intelligence is the lack of attributes for classifying and tagging URLs when classifying data. Additionally, there are seldom any freely obtainable training datasets that contain phishing URLs.

Taking all of the existing research and knowledge into consideration, a promising approach to detecting phishing websites is through the utilization of machine learning techniques. These approaches include blacklists, heuristics, visual similarity, machine learning, deep learning, and text-based approaches using NLP algorithms. This analysis will provide valuable insights into the building and association of websites, facilitating a better understanding of the online landscape. Based on our assumption and testing, we have determined that the building and association of fraud websites are more distinct compared to non-phishing internet site.

The two submissions mentioned below comprise the majority of this study:

- This project will compile a dataset of fraudulent website addresses using the most recent intelligence sources. Future studies can make advantage of this data collection. The important list it produced, which distinguished it apart from pre-existing data sets, was established by secure organization professionals and incorporated information from national sources.
- We evaluated the URLs and web addresses of the domains on the provided data set by categorizing them through six distinct artificial intelligence methods and eleven specified characteristics. With the information at hand, we attempted to evaluate which machine learning procedure would produce additional precise outcomes.

II. RELATED WORK

Techniques to notice fake websites include utilizing machine-learning algorithms to classify and label URLs and domain names based on identified features. Host-based characteristics and lexical features can both be extracted. Host-based features reveal the domain name of the website's location, its administrator, and the server from which it was uploaded. Lexical aspects are used to characterize the textual characteristics of the URL. URLs may evaluate a website's reliability according to its file format and other elements like protocol and hostname.

These approaches utilize various algorithms such as Random Forest and SVM to assess the characteristics of URLs within emails, enabling the classification of emails as potential phishing attacks. Some of the research in this field focused on identifying features of URLs and domain names to classify them effectively.

According to Ludl et al., phishing websites are classified based on only the HTML and URL information on phishing websites. The data set used for the research includes 678 fake websites and 4,049 secure pages. It has a test performance of 84.09%, depending on the outcome of the test. DOM-based, and URL-based methods that rely entirely on HTML. On the other hand, their success has been modest. due to the fact that hackers are able to change the URL and HTML format the Document Object Model

Phishing attacks can be detected using machine learning, according to Kulkarni et al. Based on a dataset containing 1,353 safe URLs of phishing sites, the suggested method can identify potentially fraudulent websites. As classifiers, we used decision trees, Nave Bayesian approaches, support vector machines (SVMs), and neural networks. Approximately 90% of real-world websites were classified accurately by the classifiers, according to the study.

In the study Fette et al. conducted, they identified phishing attacks using a technique known as PILFER that they developed to help categorize URLs. These features were essential for accurately identifying and categorizing URLs in order to effectively detect phishing attacks. Similarly, the study by Cantina focused on machine learning techniques and identified six out of fifteen features that were efficient for phishing identification. A total of 876 phishing emails and 6,643 non-phishing emails were included in the research dataset. Support Vector Machines (SVMs) are used as the classifier in this application. An accuracy of 92° was achieved after they trained and evaluated the classifier using 10-fold cross-validation. The simplest spam filter that is usually used, Spam Assassin, is claimed to be inferior to the proposed PILFER method due to its success rate. In addition to its small sample size and low success rate, the study is controversial.

Garena et al. classified fake URLs into quatern categories [25]. They used a dataset of 2,507 URLs for the investigation. The training showed an exactness of 94%. The blockbuster percentage is questionable because assailants can easily alter the URL.

III. MATERIALS AND METHODS

The first portion of this section introduces our dataset. In the next step, the proposed technique is well-defined. Then, the investigational outcomes of the used classification procedures and priority classification procedures are offered.

A. DATASET

Almost all data-driven studies require an accurate and well-organized dataset. The literature review indicates that a limited number of data sets have been used to study phishing attacks that employ machine learning methods; Fette et al., 850 phishing emails, and 6850 non-phishing 7820 [10], Zhang et al., 3,100 phishing websites [11], Xiang et al., 8,118 phishing, 4,783 genuine total 13,101 web pages [12]. However, we identified two key issues with these datasets. The first is that there isn't enough data to categorize the feature. The need for more particular details in these databases is the second issue. The lists obtained through the search are out-of-date and are handled more speculatively. Because of this, the available data sets are insufficient for our study.

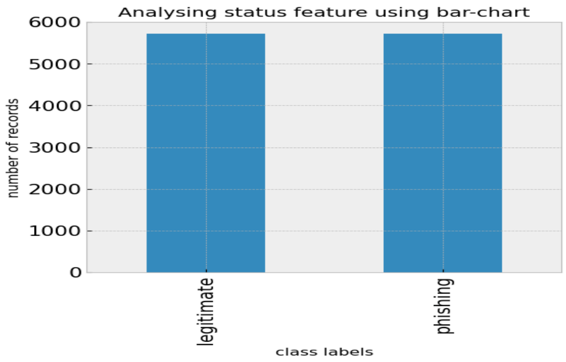

Fig1: Exploratory Data Analysis for Dataset

Various open-source data sources were used to create the dataset, including the TR-CERT website, which developed and shared ways of preventing potential cyber-attacks. Several phishing websites samples are collected by the organization's specialized team and method. The detection of malicious links took place between October 24, 2021, and September 17, 2022. Based on the data set provided, it was determined that out of the total records, 33,134 belonged to the phishing category. In the context of phishing detection, a feature is classified as either belonging to the phishing category and assigned a value of 1, or not belonging to the phishing category and assigned a value of -1. Data distrusted of phishing is specified by a value of 0 in the heuristic-based detection system, as per its defined characteristics. The careful selection of the dataset ensures that it contains a balanced representation of both phishing and legal websites. The research study aimed to generate findings that were valid and representative by utilizing a diverse and comprehensive dataset. Based on the available information, Table 1 provides a comprehensive breakdown of the distribution of values in the data analysis set, specifically focusing on the presence or absence of phishing domains.

Table 1. Data Circulation

Total Data	Phishing Data	Non-Phishing Data	Data Rate
32,928	20,614	12,314	0.626

After data segmentation, the dataset is divided into three sets: training, validation, and testing.

Table 2. Circulation of training and test data.

Total Data	Phishing Data	Non-Phishing Data	Data Rate
Training-23049	14,429	8,620	0.626
Test-9879	6,185	3,694	0.626

B. PREPARATION OF DATASET

A malicious cybercriminal uses incoming emails, notifications, SMS messages, or a different communication channel to detect the source page for the phishing attack to steal corporate account information.

A study of current research on the detection of phishing sites found that URLs and queries based on the URLs were used frequently in detecting phishing sites. In addition, obfuscation, and manipulation techniques are not usually possible with URL, and query-base data.

A URL analysis is defined as a collection of IP addresses, subdomains, prefixes, and suffixes, as well as the length of the URL. In the case of query-based data, examples include Google Search, Page Rank Checker, Web Traffic, WHOIS Query, and Statistics Report.

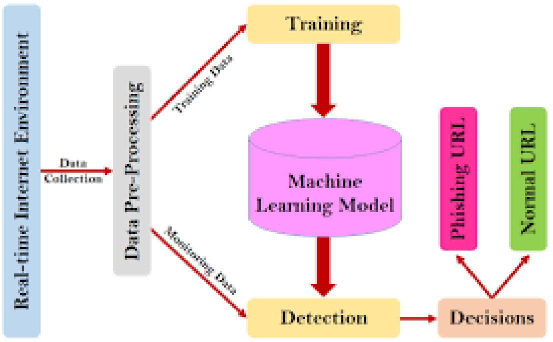

Fig 2: URL based Phishing Detection from Dataset

The creation of a dataset is depicted in Figure 1. The TR-CERT hazardous list's URL-based static analysis are used to obtain the first four indications. Exists a URL with a known IP address. Is there more than two subdomains in the second stage. Is the third stage denoted by a negative sign. Finally, it was decided if the URL included more characters than 30. Second, a four-step process was used to examine how to create query-based dynamic analytics. Has the shady website been indexed.

C. WITH URL’S TO HIDE DOUBTFUL PORTION

When an IP address is used in place of the name of the website in the URL, as in the example "http://38.38.231.15/phishg.html," it is likely an effort is being made to steal confidential data.

D. WITH LONG URLS TO HIDE DOUBTFUL PORTION

Phishers can conceal the suspicious portion in the web browser's address bar by using lengthy URLs. If the URL is longer than 30 characters that are used such as http://bitcelltr.co.in.bg/6f/aje/abc71e5e369e56502g417dre47b783a75e, it is considered to be a fraudulent page.

E. A PREFIX OR SUFFIX WITH A MINUS SYMBOL IS PRESENT IN THE DOMAIN NAME

In trustworthy URLs, the dash sign is rarely used. To give visitors the impression that they are dealing with a trustworthy website, phishers frequently attach prefixes or suffixes to domain titles that are detached by (-). Consider the website http://www.gogaccount-name.com. Such applications are frequent in phishing campaigns.

F. VARIOUS SUB-DOMAINS

Top-level website names or domain names, also known as extensions or domain suffixes, are composed of a domain name (also known as an IP address), a top-level domain, and an additional sub-domain. The firm name and well-known login credentials are used by scammers in an attempt to trick clients. If there are a couple of dots, the domain will have numerous sub-domains, and the website is considered to be a phishing site. As an illustration, http://login.domainname.xyz.com.

G. DNS RECORD

Websites are labeled as "Phishing" or "Legitimate" depending on whether the appropriate web page is predictable through WHOIS database or if histories have been produced for the website's domain name, i.e. the record in the DNS is blank or couldn't be located.

H. GOOGLE INDEX

Websites that Google has indexed show up in search engine outcomes. Numerous fraudulent websites don't appear in the Google index for the reason fraudulent websites are accessible only for a brief period of time. Phishing sites are those that Google has not indexed.

IV. USE OF MACHINE LEARNING ALGORITHMS

Detecting phishing websites can be difficult without machine learning. The learning algorithms can distinguish between genuine and deceitful websites after training on a large dataset of legitimate and fraudulent websites. The development of such systems can lead to the detection of potentially dangerous websites that can be automatically identified and warned to users.

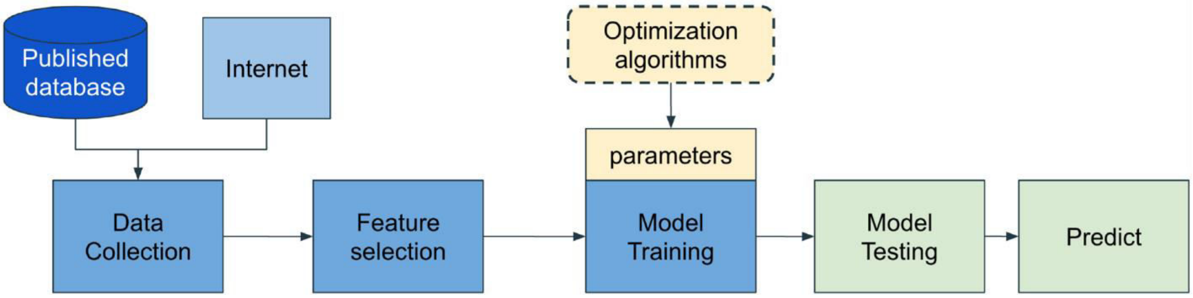

Fig 3: implementation of Machine Learning Algorithm

V. RANDOM FOREST ALGORITHM

A machine learning approach for detecting fraud is called random forest. The program, which uses unsupervised learning, isolates abnormalities in the information to find corruption. Both training data and test data are separated from the dataset. To create training data, the dataset is first shrunk by 83%, then test data is created by further reduced the trained data by 27%. Phishing URL identification is one of the numerous applications of the widely utilized machine learning technique Random Forest.

First, several decision trees are constructed in the algorithm to create a "forest" of trees. A random selection of features is used in the training of each decision tree to help prevent overfitting. The URLs that are going to be used to detect phishing have to be classified as legitimate or phishing. In addition to the URL size, you can get information about the subdomains and keywords present in the URL. Based on these characteristics, Random Forest models are developed.

VI. PREDICTION

A trained Random Forest model can be used to detect phishing URLs by following the steps outlined below

Preprocess the new URL: Extracting relevant URL features such as length, number of subdomains, and the incidence of specific keywords can offer valued insights into the nature and characteristics of a website or webpage.

Feed the features into the Random Forest model: The Random Forest model will use the extracted features of the new URL as input for its training. The random forest model uses the input features of positions and velocities of the galaxies to create a set of questions based on these features.

Get the predicted label: Based on the results of the Random Forest model, the output will be a prediction classifying the new URL as either phishing, doubtful, or legitimate.

Interpret the predicted label: The model predicts that the new URL is phishing if the predicted label is 1. This indicates that the model has predicted the URL to be legitimate if the predicted label is 0.

VII. CONCLUSION

To presumption whether an internet page was organized for phishing, data from internet pages was used in the learning. In the study, the random forest model was preferred due to its high accuracy and ability to handle complex data relationships. The model was useful to a effective dataset, and the estimate rates twisted suitable results, confirming the reliability and validity of the model. On untrained data, the model consistently demonstrates a high level of accuracy in its predictions. One popular technique for classifying websites is to extract features from the URL and webpage content, such as domain name, IP address, presence of suspicious keywords or phrases, and the structure of the webpage. The use of URL-based research has been shown to enhance the speed of detection. However, the choice of algorithm depends on several factors such as the size and nature of the dataset, the computational possessions presented, and the detailed necessities of the classification task at hand.

In order to achieve this goal, researchers propose the use of feature selection algorithms to assess the legitimacy or maliciousness of a website. Random Forest proved to be highly effective in achieving a detection accuracy of 98.14%, with the additional benefit of minimizing false positives. Moreover, the improvement in classifier performance is directly proportional to the increase in the amount of training data utilized. As well, the results show that more data in training gives better results for classifiers.

VIII. REFERENCES

[1] I.Vayansky and S.Kumar, ``Phishing: Challenges, and solutions,'' Comput. Fraud, Secur., vol. 2018, pp. 1520, Jan. 2018.

[2] (2022). Tessian. [Online]. Available: https://www.tessian.com/blog/ phishing-statistics-2020/

[3] (2021). IBM. [Online]. Available: https://www.ibm.com/security/databreach

[4] D.-J. Liu, G.-G. Geng, X.-B. Jin, and W. Wang, ``An efficient multistage phishing website detection model based on the CASE feature framework: Aiming at the real web environment,'' Comput. Secur., vol. 110, Nov. 2021,Art. no. 102421.

[5] R. S. Rao and A. R. Pais, ``Detection of phishing websites using an efficient feature-based machine learning framework,'' Neural Comput.,Appl., vol. 31, pp. 38513873, Aug. 2019.

[6] Y. Cao, W. Han, and Y. Le, ``Anti-phishing based on automated individual white-list,'' in Proc. 4th ACM Workshop Digit. Identity Manage. (DIM), 2008, pp. 5160.

[7] K. L. Chiew, E. H. Chang, S. N. Sze, and W. K. Tiong, ``Utilisation of website logo for phishing detection,'' Comput. Secur., vol. 54, pp. 1626, Oct. 2015.

[8] W. Zhang, H. Lu, B. Xu, and H. Yang, ``Web phishing detection based on page spatial layout similarity,'' Informatica, vol. 37, no. 3, pp. 231-244, 2013.

[9] L. J. P. van der Maaten and G. E. Hinton, ``Visualizing high-dimensional data using t-SNE,'' J. Mach. Learn. Res., vol. 9, pp. 25792605, Nov. 2008.

[10] A. A. Ubing, S. Kamilia, A. Abdullah, N. Jhanjhi, and M. Supramaniam, ``Phishing website detection: An improved accuracy through feature selection and ensemble learning,'' Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 1,p. 2019, 2019.

[11] A. Valjarevic and H. S. Venter, ``Introduction of concurrent processes into the digital forensic investigation process,'' Austral. J. Forensic Sci., vol. 48, no. 3, pp. 339357, May 2016.

[12] A. Georgiadou, S. Mouzakitis, and D. Askounis, ``Detecting insider threat via a cyber-security culture framework,'' J. Comput. Inf. Syst., vol. 62,no. 4, pp. 706-716, Jul. 2022.

[13] M. C. A. M. R. Damodaram and L. Valarmathi, ``Phishing website detection, and optimization using particle swarm optimization technique,'' Int.J. Comput. Sci., Secur., vol. 5, no. 5, p. 477, 2011.

[14] A. V. Bhagyashree and A. K. Koundinya, ``Detection of phishing websites using machine learning techniques,'' Int. J. Comput. Sci., Inf. Secur.,vol. 18, no. 7, 2020.

[15] J. Jang-Jaccard and S. Nepal, ``A survey of emerging threats in cybersecurity,'' J. Comput. Syst. Sci., vol. 80, no. 5, pp. 973993, Aug. 2014.

IX. AUTHORS PROFILE

1. **RAMIREDDY HIMABINDU** Pursuing M.Tech at MJR College of Engineering & Technology, Department of Computer Science & Engineering, Piler, Andhra Pradesh.
 Email: himabindu67@gmail.com
2. **N SURENDRA** Working as an Assistant Professor in MJR College of Engineering & Technology, Department of CSE, Piler, Andhra Pradesh.
3. **V SUBHASINI** Working as an Assistant Professor & HOD in MJR College of Engineering & Technology, Department of CSE, Piler, Andhra Pradesh.