JURNAL EMACS
[Engineering, MAthematics and Computer Science] Vol.5 No.2 May 2023: 63-67
e-ISSN: 2686-2573
DOI: 10.21512/emacsjournal.v5i2.9951

Phishing Site Detection Classification Model Using Machine Learning Approach

Yohan Muliono[1*], Muhammad Amar Ma'ruf[2], Zakiyyah Mutiara Azzahra[3]
[1-3] Cyber Security Program, Computer Science Department, School of Computer Science,
Bina Nusantara University,
Jakarta, Indonesia 11480
ymuliono@binus.edu; muhammad.maruf@binus.ac.id;
zakiyyah.azzahra@binus.ac.id

*Correspondence: ymuliono@binus.edu

Abstract – *Phishing has been a cybercrime that has existed for a long time, and there are still many people who are victims of this attack. This research attempts to prevent phishing by extracting the attributes found on phishing websites. This study uses a hybrid method by combining allowlist and denylist as part of a classification system. This research utilizes 18 features to identify a phishing site in terms of address bar, abnormal request, and source code (HTML and JavaScript). Where in each feature the author determines the benchmark. This study validates the status code and detects 52 URL shortening service domains and then evaluates these abnormalities with a binary classification system. Algorithms that have good results are Decision Tree and K Nearest Neighbor (KNN). After evaluating the performance of the algorithm in terms of Precision, Recall, and F-Measure. As a result, the Decision Tree algorithm has the highest accuracy of 97.62% and the fastest computation time of 0.00894 seconds. So that the Decision Tree is superior in terms of accuracy and computation time in detecting phishing URLs.*

Keywords: *Phishing; Machine Learning; Cyber Crime; KNN; Decision Tree*

I. INTRODUCTION

Phishing is a form of cybercrime which takes advantage of the victim's negligence (Saha et al, 2020) in accessing a link on a website, so that the victim enters sensitive data in a fake link. According to (Ketaren, 2017), cybercrime is a criminal act that violates the law by using computer technology as a means of crime. In 2020, cybercime is responsible for losses of $13.3 billion or 190 trillion rupiah throughout 2020. (FBI, 2020).

According to an investigation conducted by Verizon in its Data Breach Investigation Report (DBIR) (Philippe Langlois, 2020), more than 30% of data breaches that occurred in 2021 were caused by human negligence in dealing with attacks regardless of the layer of security that may have been applied, human negligence can always happen.

The Anti-Phishing Working Group (APWG) (Aaron, 2021) noted that throughout 2021 the trend of using HTTPS on phishing websites was carried out by 82% of total phishing cases, while the use of HTTP on phishing websites continued to decrease below 20% per 2021 HTTPS itself is used to secure communication by encrypting data sent between the browser and the website it visits. HTTPS is an important key to knowing the security of a site. With the widespread use of HTTPS in phishing sites, it's becoming increasingly difficult to tell the real site apart. Therefore, other features are needed to identify phishing sites.

According to (Ansari et al, 2022), phishing attacks can be detected and prevented using an AI-based model, which in this study will use a decision tree and K-Nearest neighbor as an experiment.

II. METHODS

The model proposes in this research is a hybrid model where this model will utilize allowlist, denylist then using machine learning techniques simultaneously. Take advantage of the use of allowlist and denylist techniques to minimize the possibility of false positives in the classification

system. This research uses 18 features to identify phishing sites, with benchmarks that used in previous researches. The model that will be proposed does not use third party services, this is done to reduce computational time in the identification process. To reduce computation time, the author's model also includes a filter that does not process phishing sites that are no longer active, so that the site's address does not go through the feature extraction process, significantly reducing computation time.

As of data collection method, this research collecting data in the form of phishing and non-phishing sites. Data collected via the Internet and references from previous studies. The source from the internet in question comes from *http://www.alexa.com* for non-phishing sites, while the list of phishing sites is obtained from PhishTank which is addressed at http://phishtank.com. As a result, data obtained approximately 4,000 site URLs, each consisting of 2,000 non-phishing site URLs and 2,000 phishing site URLs.

The URLs collected will be checked for its activeness, if the website is active and accessible, the website will continue to enter another validation system. If not, then the website will pass through the system and not enter into the feature extraction and other validation processes so as to streamline computer work and computation time. After going through the status code checking process.

The allowlist function is as an access control over what URLs can or may be accepted by the user. Usually this concept is implemented in corporate security networks to prevent employees from accessing URLs other than those specified. This research takes advantage of this allowlist concept and uses it to allow trusted domains to enter without passing through the author's system thereby increasing the effectiveness of the system and reducing the false positive rate, i.e. where actual non-phishing URLs are predicted as phishing by the system mean while denylist function is as access control to deny permission to certain URLs that are in a denylist. The author uses the denylist concept to check whether the website is included in a phishing website. If yes, then the URL no longer needs to be entered into the system for the extraction feature to be carried out, the URL will be automatically issued as phishing. Thus increasing the effectiveness of the system's work and reducing the false negative rate (false positive rate), namely where the URL is actually phishing but is predicted as non-phishing by the system.

As many phishing URLs comes in a form of shortened URL, this classification system also has a filter to detect if a URL uses a URL shortening service. If the URL uses the service, the system will automatically issue a suspicious or suspicious value. This is done to reduce the false negative rate where the phishing URL is predicted to be non-phishing by the system or the prediction is inaccurate in the system because the features of a website that are extracted on the author's system are the websites that are displayed first, if the perpetrator uses a shortening service. URL, then what is detected on the author's system is the shortening service, not the phishing website itself and the whole system defined as Figure 1.

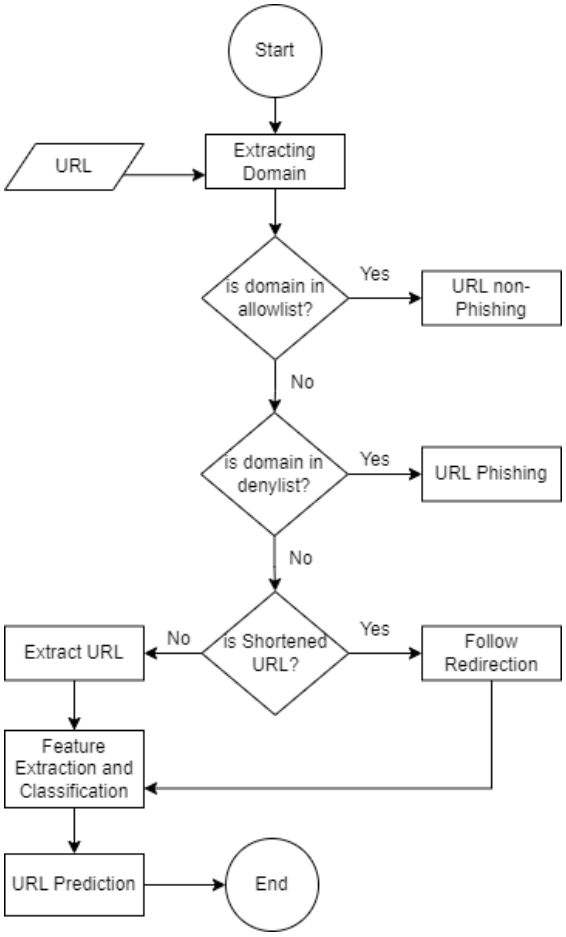

Figure 1. Classification Model

The proposed feature extraction method will focus site URLs to identify phishing websites from non-phishing websites. 18 features are used to effectively determine phishing and non-phishing URLs. The model in this research prioritizes a URL-based approach, because this approach is able to combine and evaluate detection features in a domain. Below are the features extracted to feed into the classification model:

2.1. IP Address

In general, legitimate websites will provide their domain with their brand name. When logged into the system, the hostname of the URL will be checked to see if the URL contains an IP address. Based on previous research, (Alshahrani et al, 2022) it was found that most of phishing URLs contained IP addresses, while out of 1200 non-phishing URLs did not contain IP addresses.

2.2. URL Length

URL is a string used by internet users to identify a source from a URL. The URL string consists of three elements, namely network protocol, domain and path. For a given URL, the length of the URL will be checked down to its subdirectories.

2.3. Slashes in URL

Phishers always Phishers always try to trick web users by imitating the appearance of URLs to make them look legitimate. One technique used in phishing is to add a slash in the URL. Therefore, the authors consider the number of slashes in the URL as an identifying feature. The slashes will be checked and counted. Based on previous

research (Jeeva et al, 2016) the average number of slashes in phishing URLs found is more or equal to 5. And the average number of slashes in non-phishing URLs found is an average of 3. The data collected is 1200 URLs phishing and 200 non-phishing URLs.

2.4. Prefix and Suffix

This feature counts the number of hyphen characters ("-") in a hostmame which usually signifies a prefix and suffix. According to previous research conducted by (Jeeva et al, 2016), showed that out of 1200 phishing URLs and 200 non-phishing URLs, the average number of hyphens in phishing URLs was more than one. While the average number of non-phishing hyphens is 1.

2.5. Subdomain

The Security Week (Gupta et al, 2021) reports that phishing attacks are increasing with the use of subdomains. Phishers trick users by adding sub domains to make the URL look legitimate. Adding a subdomain to a URL makes cyberspace users believe that the URL belongs to the real website. From the results of the author's observations, the author found that 36.76% of the 10,000 phishing URL data had at least 1 subdomain.

2.6. Favicon

A favicon is an icon that appears in the browser address bar or browser tab or next to a website name. Figure 1 shows an example of an Internet Explorer browser displaying the favicon of a PayPal website address. A favicon represents a website's identity as a 16×16 pixel image file. It is also available in several different image sizes, such as 32 × 32, 48 × 48, or 64 × 64 pixels.

To access the favicon, we add string "favicon.ico" to the website's domain name. For example, the URL name that we want to check is https://www.paypal.com/, the model extracts only the domain name, namely, paypal.com. and adding favicon.ico to the end of the domain name, so it becomes paypal.com/favicon.ico. The newly formed URL will be entered into a Google search by the image engine to get information related to the favicon. Based on observations (Jeffrey et al, 2018) the favicon feature can reduce the false positive rate by 0.57%. However, it can increase the false negative rate from 3.00% to 3.03%.

2.7. Non Standard Port Usage

This feature has a function to validate if a URL requests access for a connection on a certain port. To prevent phishers from getting connections through important ports some requests to ports are blocked. However, some services, such as firewall servers, proxies, and Network Address Translation (NAT), by default block all or most of the ports and only open selected ports. If all ports are open, phishers can run almost any service they want.

2.8. Transport Layer Security

URL can be divided into 3 components namely Domain, Top Level Domain, and Path. The URL uses Transport Layer Security to determine whether the URL is encrypted data in the process of being sent. The presence of the HTTP Protocol is very necessary when sensitive information is transferred across the network. Therefore the type of Transport Layer Security (TLS) will be checked. If the URL uses TLS in the form of HTTPS then it is given the value non-phishing otherwise it is phishing. The results of the analysis carried out (Jeeva et al, 2016) by analyzing the phishtank data set, 99.16% of URLs were found without HTTPS

2.9. Special Character "@"

The "@" character in URLs allows URLs before the "@" character to not be processed. Where non-phishing URLs will be placed after the "@" character while non-phishing URLs will be placed on the left before the "@" character. Example:

https://www.xyzbank.com@login.phishing.xyzbank.com

The browser will automatically process "login. phishing.xyzbank.com" and not process "www. xyzbank. com". This technique uses official websites as disguises to make the URL appear official at first. In research conducted by (Maher et al, 2010) the use of the "@" character was found in 20% of the 1,000 phishing URLs they studied.

2.10. Abnormal URL Request

The Abnormal URL request feature will check whether the website makes a request to an external domain to retrieve its assets contained in web pages such as images, videos and sounds. On non-phishing websites, the website address and most of the assets are retrieved from the same page source. In a phishing website, most of the assets are copied or loaded from the site it is impersonating. This is done because the attacker intends to reduce production costs to create a phishing website, this is usually because phishing websites are produced on a large scale covering various sectors, so that many assets are loaded from external systems, in order to simplify and cut production costs. issued to memory. In research conducted by (Aljofey et al, 2022) the use of external URL requests was found in 100% of the 1,000 phishing URLs studied. In the author's observations, the average phishing URL making requests to external domains is as much as 41%.

2.11. URL of Anchor

Not only assets such as images, video or audio are taken from external websites. Phishers try to make websites as similar as possible to the official website they are copying, so most of the assets and media shown on phishing pages are from the original official website. Unlike phishing websites, legit websites don't solicit assets from external domains. The source will look something like src = /asset/img/logo.jpg compared to a phishing website with the source in the image loaded from an outside domain src =

https://www.legitimatesite.com/asset/img/logo.jpg.

The <a> tag is treated exactly like the "Abnormal URL Request" feature but in this feature we will check:

- If the <a> tag makes a request to an external domain
- If the <a> tag does not link the request to any web page:

  a. <a href="#">
  b. <a href="#content">
  c. <a href="#skip">
  d. <a href="JavaScript ::void(0)">

Observations made by (Aljofey et al, 2022) the use of the <a> tag making a request to an external domain on a phishing website was found in 233 of the 1,000 phishing URLs studied. The proportion of <a> tags that make requests to external domains is found with an average proportion of 37% on a phishing URL from 6854 active data.

2.12. Link in <Script> and <Link>

Not only the <a> tag, this research also observes and examines tags that allow a phishing website to make requests to external domains. Like the href and src attributes in the <script> and <link> tags. Observations made by (Aminu et al, 2019) use of the <script> and <Link> tags that make requests to external domains on are included in the top 7 features in detecting phishing websites. In my observations, the proportion of links in <script> and <Link> that make requests to external domains is found to be an average proportion of 49% in a phishing URL out of 6854 active data.

2.13. Server Form Handler

Server Form Handler with an empty string or "about:blank" can be considered suspicious because every form filled out must have an action afterwards on the information submitted. Also, if the domain name on the Server Form Handle is different from the domain name of the web page, this indicates that the web page is suspicious because the information submitted is rarely handled by external domains. In research conducted by (Maher et al, 2010) the use of an external Form Handler Server on a phishing website found 100% of the 1,000 phishing URLs studied.

2.14. Info Submit through E-mail

A form on a web page can serve to send a user's personal information to a server for processing. A fraudster might redirect user information to his personal email. For that, scripts on the server side can use the "mail()" function. And on the client side there will probably be a "mailto:" function in the script to send private information to phisher emails. Research conducted by (Aljofey et al, 2022) used the "mailto:" function on a phishing website to find as many as 20% of the 1,000 phishing URLs studied or 200 phishing URLs.

2.15. Website Forwarding

One of the things that differentiate phishing websites from legitimate ones is the number of site redirects that are made when accessing a. In my data set, I found that legitimate websites make a maximum of one redirect to another website. On the other hand, a phishing website that does at least 4 redirects. The goal is to make phishing websites more difficult for anti-phishing to detect.

2.16. Mouseover

In some cases, phishers are getting smarter in tricking victims. In this feature phishers imitate a link so that the link looks like it is from an official site

2.17. Disable Right-Click

Phishers can use JavaScript to disable the right click on the mouse, this is done so that the user cannot see and save the source code of the website, this is suspect, official websites generally don't hide anything from their users. For that the author will look for the scripts "event.button==2" and ".preventDefault()" in the website page source code and check if right click is disabled.

2.18. Pop-up Window

An official website rarely asks users to submit their personal information via a pop-up window. In contrast, this pop-up window feature is mostly used by legitimate websites to alert users about fraudulent activity or notify announcements.

III. RESULTS AND DISCUSSION

4,000 site URLs, each consisting of 2,000 non-phishing site URLs and 2,000 phishing site URLs. 80% will be used for training, and 20% for testing for each phishing and non-phishing URLs, the algorithm use to classify will be Decision Tree and K-Nearest Neighbor.

The following is the test results of the Decision Tree algorithm with a total data of 4000 URLs. The results of the Decision Tree algorithm trial produce an average accuracy of 97.62%, show in Table I.

Table I. Decision Tree Result

Class	Precision	Recall	F1
Non-Phishing	0.99	0.96	0.97
Phishing	0.96	0.99	0.97

After going through the train and test with the K-Nearest Neighbors algorithm with a total data of 4000 URLs. The results of the K-Nearest Neighbor algorithm trial produced an accuracy of 97.25% with a training time of 381 ms. After making predictions on the dataset, the authors conducted an accuracy rate test, namely by finding the correct K value for the proposed classification system based on precision, recall and F1-score. The following is a graph of the accuracy value of the k value, show in Figure 2.

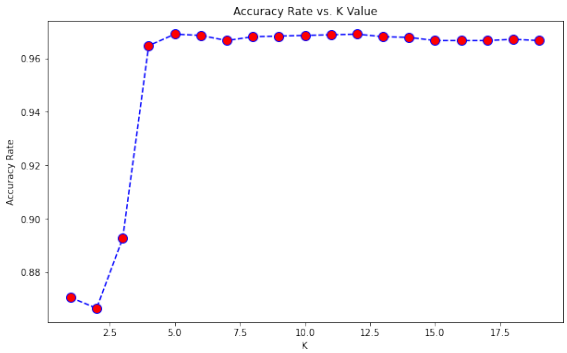

Figure 2. Accuracy: k-value

After running some trials. This research found that the value of k has an accuracy of 12 with an accuracy of 96.79%. When compared with precision, recall and F1 score. Then k = 12 can be accepted as the best accuracy for the K-Nearest Neighbor algorithm. After the next k value is determined, the two k values are compared in terms of Precision, Recall and F1-Score (F1). The following is the performance evaluation matrix of the K-Nearest Neighbor algorithm with a value of k = 5 and k = 12, show in Figure 3 and 4.

	precision	recall	f1-score	support
0.0	0.9948	0.9268	0.9596	410
1.0	0.9345	0.9953	0.9640	430
accuracy			0.9619	840
macro avg	0.9646	0.9611	0.9618	840
weighted avg	0.9639	0.9619	0.9618	840

Wall time: 205 ms

Figure 3. KNN Result with k value 5

	precision	recall	f1-score	support
0.0	0.9948	0.9390	0.9661	410
1.0	0.9448	0.9953	0.9694	430
accuracy			0.9679	840
macro avg	0.9698	0.9672	0.9678	840
weighted avg	0.9692	0.9679	0.9678	840

Wall time: 156 ms

Figure 4. KNN Result with k value 12

From the performance evaluation matrix above, it can be concluded that the average accuracy obtained with a value of k = 5 is 96.19%. While the average accuracy obtained with a value of k = 12 is 96.79%. So it can be concluded that the best K value is 12 with a training time value of 156 ms. The following is the final matrix of the K-Nearest Neighbor algorithm trial in terms of Precision, Recall and F1-Score (F1) with a value of K = 12, show in Table II.

Table II. KNN Result with k value 12

Class	Precision	Recall	F1
Non-Phishing	0.99	0.93	0.96
Phishing	0.94	0.99	0.96

IV. CONCLUSION

The aim of this research is to leverage advances in computer science in a way that contributes to cybercrime prevention efforts. The author focuses on cybercrime phishing, by extracting the attributes found on phishing websites. The author uses a hybrid method by combining allowlist, denylist, code status checking, and URL shortening as part of the author's classification system. This research utilize 18 features to identify phishing sites in terms of address bar, abnormal requests, and source code (HTML and javascript) etc. then identify the URL as Phishing and Non-Phishing using Decision Tree and K-Nearest Neighbor. As a result, the Decision Tree algorithm proved to be more suitable in detecting phishing URLs with an average accuracy of 97.62%. Finally, this research could have improved further using additional features that will be used by phisher in the future and keep update the features and technique used by phisher in the future.

REFERENCES

Aaron, G. (2020). Phishing Activity Trends Report. Anti Phishing Working Group.

Aljofey, A., Jiang, Q., Rasool, A., Chen, H., Liu, W., Qu, Q., & Wang, Y. (2022). An effective detection approach for phishing websites using URL and HTML features. Scientific Reports, 12(1), 8842.

Alshahrani, S. M., Jeeva, S. C., & Rajsingh, E. B. (2022). URL Phishing Detection Using Particle Swarm Optimization and Data Mining. *CMC J*, *73*, 5625-5640.

Aminu, A., Abdulkarim, A., Aliyu, M., Yahaya, A., & Maigari, A. (2019). Detection of Phishing WebsitesUsing Random Forest and XGBOOST Algorithms. Frontiers of Knowledge Journal Series.

Ansari, M. F., Sharma, P. K., & Dash, B. (2022). Prevention of phishing attacks using AI-based Cybersecurity Awareness Training. Prevention.

Federal Bureu Of investigaion, I. C. 2020). Internet crime report 2020: *https://www.ic3.gov/*

Gupta, B. B., Yadav, K., Razzak, I., Psannis, K., Castiglione, A., & Chang, X. (2021). A novel approach for phishing URLs detection using lexical based machine learning in a real-time environment. Computer Communications, 175, 47-57.

Jeeva, S. C., & Rajsingh, E. B. (2017). Phishing URL detection-based feature selection to classifiers. International Journal of Electronic Security and Digital Forensics, 9(2), 116-131.

Ketaren, E. (2017). CYBERCRIME, CYBER SPACE, DAN CYBER LAW. Article, 35–42.

Langlois, P. (2020). 2020 Data Breach Investigations Report.

Saha, I., Sarma, D., Chakma, R. J., Alam, M. N., Sultana, A., & Hossain, S. (2020, August). Phishing attacks detection using deep learning approach. In *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)* (pp. 1180-1185). IEEE.