Real-Time Phishing Detection Using Statistic Database Check, DNS and Who Is Check, Verifying ASCII Content of the URL and Visual Similarity

Uthkarsh Sanjay(✉), Pushkar Ananad, Adith A. Danthi, G. R. Akshay, and P. Ravi

Department of Computer Science and Engineering, Vidyavardhaka College of Engineering, Mysore, India
uthkarshsanjay@gmail.com, ravip@vvce.ac.in

Abstract. Phishing happens to be a severe cyber-crime that affects tens of lakhs of people every day. Cyber-attacks are now getting to end consumers, exploiting the weakest security component. So, to rectify these types of issues we need to develop different types of phishing detection techniques. We proposed a variety of phishing strategies for detection, rectification, and prevention, all of which are necessary to identify phishing. We detect phishing attacks using these four methods, Statistic Database check, DNS and who is check, Verifying Ascii content of the URL and Visual similarity and we have derived impressive results out of these methods.

Keywords: Phishing · Statistic Database · DNS · ASCII content and visual similarity

1 Introduction

Phishing attacks always aim the weak spots that exist because to manual error at the expense of confidential info such as credit-card information, social security numbers, employment details and bank account numbers. These con artists create fake websites and fictitious e-mail addresses in order to defraud individuals who have participated in secret financial transactions by collecting credentials. Innocent users trust the facts they obtain on the internet, and phishers utilize email/website/URL redirection to carry out injection assaults.

Phishing techniques are becoming more common, and one of them is projecting a login screen that allows phishers to reproduce the same website. The scammer sends an email with a Hyperlink that redirects to a clean website that claims to be legit. However, authentic account information, such as official websites, may be requested. As a result, it is evident that phishers use deceptive methods to entice visitors, such as suspicious URLs, emails, iframes, suspicious scripts, and pictures.

D. S. Guru et al. (Eds.): ICCR 2021, CCIS 1697, pp. 332–339, 2022.
https://doi.org/10.1007/978-3-031-22405-8_26

By employing a feature selection technique, the General Phishing-Detection improves accuracy. The algorithm selects a subset of the dataset's properties that are essential in forecasting the outcome. Unnecessary features have no bearing on the system's accuracy. Furthermore, Ensemble Learning is used to train the system. Because the result of using many models to make predictions is impartial, it is directed that the results from several of the models are regarded to represent the major part.

For example, if the majority of the models warn that the particular website is phishing, the ensemble's inference is that the site is phished.

2 Related Work

The methods used in this paper are, Linear SVC classifier, K-Nearest Neighbor and One Class SVM, (DT)Decision tree classifier, and (RF)Random Forest Classifier split into 2 stages generation and Prediction [1]. The Random Forest Classifier on a Dataset approach was utilized in this article, and the results showed that this method performed better than the others, with the greatest accuracy of 97.36% [2]. The mechanism employed in this study is TF-IDF weights to terms that are comparable to the hostname, path, and filename URLs. Which are run on WHOIS search to see whether there existed any difference between the real and the chosen domain name, with the following outcome: If the query and owner domain name are different, a phishing website can be illustrious [3]. The MFPD method is achieved in this paper by CNN to fetch local correlation features from UR.LSTM network dependency from a character sequence, and SoftMax to categorize select features. The results show that the MFPD method is more effective than other methods [4]., Rule Evaluation, Aggregation of Rule Outputs, Fuzzification and Defuzzification are the methodologies employed in this work, and the result obtained is layer one of the fuzzy website phishing systems demonstrated the relevance of the phishing website principle, demonstrating that a website might be phishy even if the rest of the characteristics are present and accurate [5]. This research investigates the identification of a segmented website logo using Google Image Database. To match the identification in Google Picture Search, a context-based image retrieval technique is utilized, and the accuracy attained is On the Google picture database, detection accuracy improves by up to 93% [6]. PhishNet is one of the approaches used in the paper are Predictive Blacklisting, DNS- Based Blacklist, Google Safe Browsing API, Automated Single White-List, and the precision observed are as follows: Blacklists are regularly renewed list of phished URLs and protocols that have been identified as phishing [7]. This technique performs by comparing the safety percentages among two website pages' codes for real and false websites, and fetching some phishing features from the W3C standards, and the conclusion is that a high percentage indicates a secure site, while the others indicate the website is almost certainly phished [8]. This approach compares the closeness among two web pages by contrasting the content of the two websites and calculating the accuracy gained. This approach finds phished website pages with a precision of 0.96 and a false-rate of less than 0.105 [9]. The approach described in this study works as mentioned below: keywords in URLs are translated into normal images, and then image signatures with attributes such as main color categories and centroid coordinates are presented to determine the similarity of two Web sites. It does not take the code into account if it is only aesthetically similar [10]. In this work, many differentiating methods such as Linear Discriminant, Nave Bayesian, and K-Nearest Algorithm are used, and this technique has a true accuracy of 85% to 95% and a false rate of 0.43% to 12% [11]. Once the malicious web page's target domain is discovered, a third-party DNS search is done, and the two IP addresses are compared in the article. The findings reveal that this method properly detected 99.85% of the domains [12]. The procedures utilized to check in this study include unusual anchors, unusual server form handlers, unusual request URLs, unusual cookies, unusual certificates, unusual URLs, and unusual DNS records in SSL, and the false-positive and miss-rates are exceedingly less [13]. The TF-IDF Algorithm is applied to identify phishing, and the Robust Hyperlinks is used to find the owner of such brands. The results show that the TF-IDF approach can predict 97% of fraudulent websites with just 6% false positives [14]. The method used in the paper is heuristic-based, and it checks many attribute of a website to detect phishing. This test yielded a phishing detection rate of 98% [15]. This work employs various features such as type, domain, page and word based features, and they discovered that on a single day, approximately 777 unique phishing pages were discovered, with 8.24% of users viewing phishing pages being classified as potential phishing victims [16]. In this publication, the approach is broken down into four steps: Get a list of possible phishing sites. Workers are sent the URL, they evaluate the potential phishing site, and the Task Manager aggregates the results. IE7 was the only platform that could accurately detect 60% of false URLs out of all the tools provided, however it misclassified 25% of the APWG created URLs and 32% of the phishtank.com URLs [17]. In this work, several normal data proportions for training and testing. The SVM is better than NN in detection; in terms of the not-true alarm rate and prediction for Probe, Dos, U2R, and R2Lattacks, only NN could outperform the SVM in terms of prediction [18]. In this proposed work, the steps to stay alert from phished websites are stated in detail so that everyone is alert about the basic guidelines to follow before providing personal information to a false website. As a result, users are expected to double-check the website before providing personal or sensitive information [19].

3 Proposed Work

In this section, we presented four different methods to identify phishing techniques, they are Statistic Database check, DNS and who is check, Verifying Ascii content of the URL and Visual similarity. The Fig. 1 shows the architecture of the proposed system and in this architecture, the working of our system that is all the four steps are shown in detail.

When the user enters any of the websites, first the URL is extracted to test whether it's a phishing website. The first test will be a database checks here frequently updated lists of previously detected phishing URLs will be stored in the database as a blacklist. So whenever the user tries to access the phishing website, the URL will be checked with the blacklisted URLs. So if the URL is present in the database then a popup will be raised notifying it as a phishing website. If the URL is not present in the database then the next check is ASCII check. URLs may also contain ASCII characters that are visually similar but not the same, which would in many cases very difficult to distinguish from the original, thereby redirecting you to a phishing or a scam site. To avoid this URL

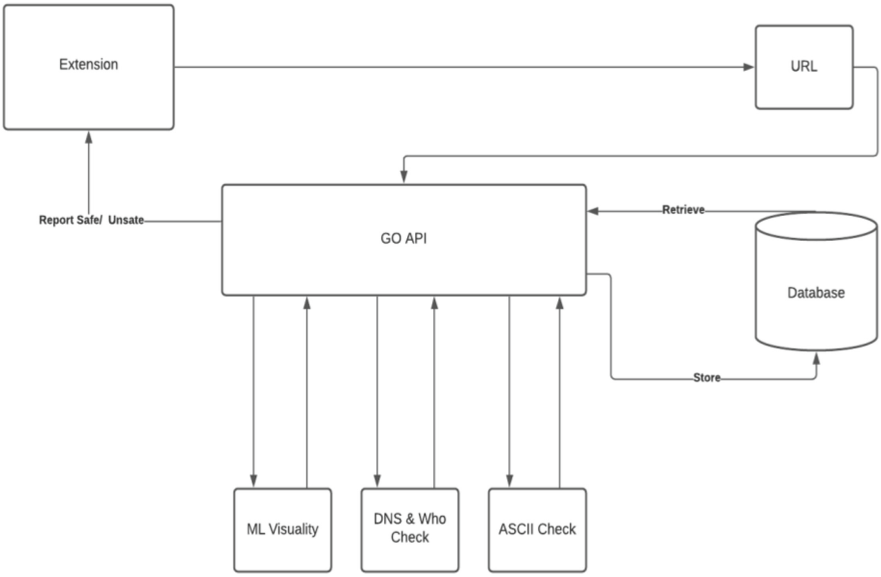


Fig. 1. Architecture of proposed system

match is done to flag any such URLs. If it's flagged then a popup will be raised notifying as a phishing website with a reason why and where it's being flagged and that particular URL is added to the database as blacklisted for faster detection.

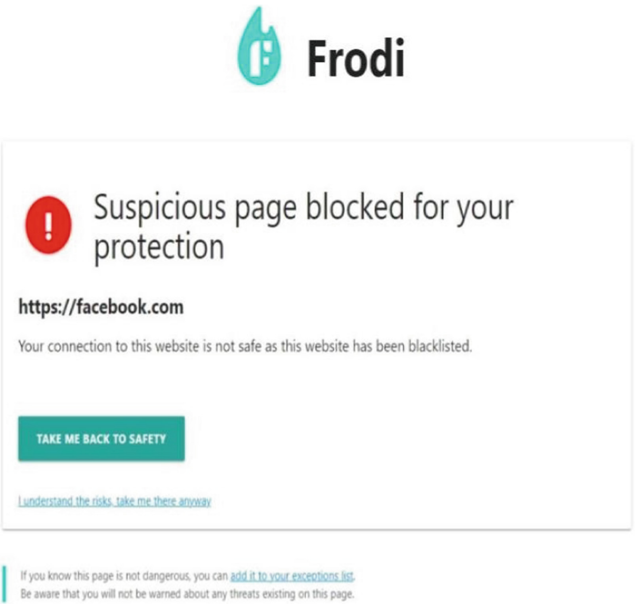
If the ASCII check is passed, then the next check will be DNS who is check. Here we verify the website's owner details to ensure the website's authenticity. Ensure the resolved IP for a website is within the actual IP range of the organization/company. If it the check is failed then a popup will be raised notifying as a phishing website and that particular URL is added to the database as blacklisted for faster detection. If the DNS check is passed then the next and powerful check will be visual similarity checked. Phishing websites generally create a clone of the UI of the original website.

Our model will check for the visual similarity between the snapshot of the website visited and the original website and will flag accordingly. If this check is failed a popup will be raised notifying as a phishing website and that particular URL is added to the database as blacklisted for faster detection. If all the cases are run successfully then the user can visit the particular website. Even if the user wants to visit the website for other reasons knowing it has a phishing website then there is an option to go to the website by clicking the option he can navigate.

4 Experimental Results

Method- 1: Static Database
We used Redis for storing the URLs, here reguraly renewed list of recently detected phished URLs will be stored in the database as a blacklist. So, whenever a user attempts

to access the phishing website, the URL will be checked with the blacklisted URLs. In the context of phishing, a blacklist is a list of untrusted URLs or, more simply, a list of prohibited websites that are known to have harmful intent as shows in Fig. 2. For Example, if a user tries to visit some phishing website, then extension will check the website URL or IP address with the blacklisted URLs. If ever it is a phishing website, then it will display a warning to the user.



Suspicious page blocked for your protection

https://facebook.com

Your connection to this website is not safe as this website has been blacklisted.

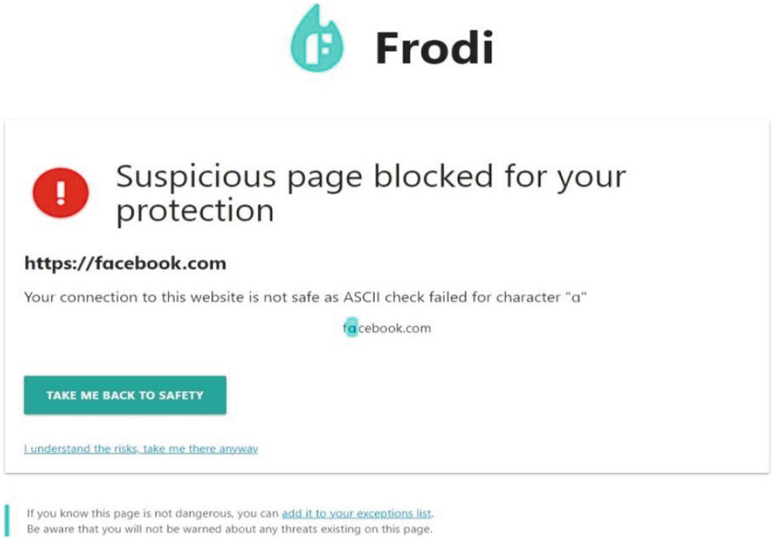
I understand the risks, take me there anyway

If you know this page is not dangerous, you can add it to your exceptions list.
Be aware that you will not be warned about any threats existing on this page.

Fig. 2. Blacklisted URL check

Method - 2: Verifying ASCII content of the URL

URLs may also contain ASCII characters that are visually similar but not the same, which would in many cases be very difficult to distinguish from the original, thereby redirecting you to a phishing or a scam site. To avoid this a URL match is done to flag any such URLs. For example, Rho(‘ρ’) and ‘p’ which look similar, but they are not. So, by checking the ASCII value of a character whether it lies within 97 to 122. If it lies in the range then it is the actual website otherwise it is a phished site, then those websites are added to the database as a blacklist. Presence of @ symbol at the URL: If @ symbol present in URL then the feature is set to malicious else set to legitimate as shows in Fig. 3.



Suspicious page blocked for your protection

https://facebook.com

Your connection to this website is not safe as ASCII check failed for character "ɑ"

fɑcebook.com

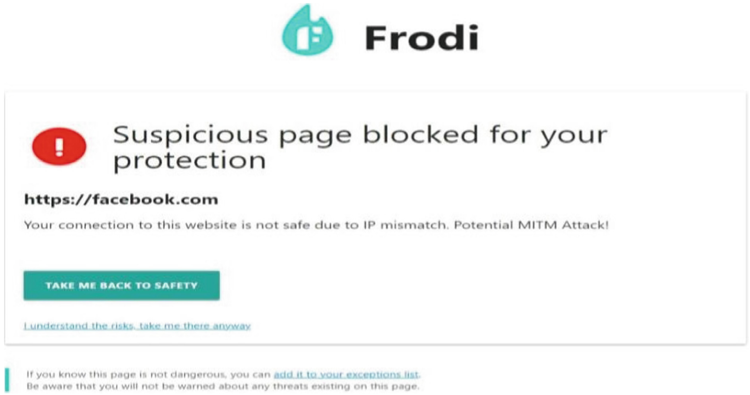
I understand the risks, take me there anyway

If you know this page is not dangerous, you can add it to your exceptions list.
Be aware that you will not be warned about any threats existing on this page.

Fig. 3. ASCII content check

Method - 3: DNS and Who is Check

Verify the website's owner details to ensure website's authenticity and ensure the resolved IP for a website is within the actual IP range of the organization/company. Man in the middle assault happens when a criminal inserts himself into a conversation between a user and a programmer, either to eavesdrop or to mimic one of the parties, giving the impression that a regular flow of information is taking place. The result of DNS and Who is check is shown in Fig. 4.



Suspicious page blocked for your protection

https://facebook.com

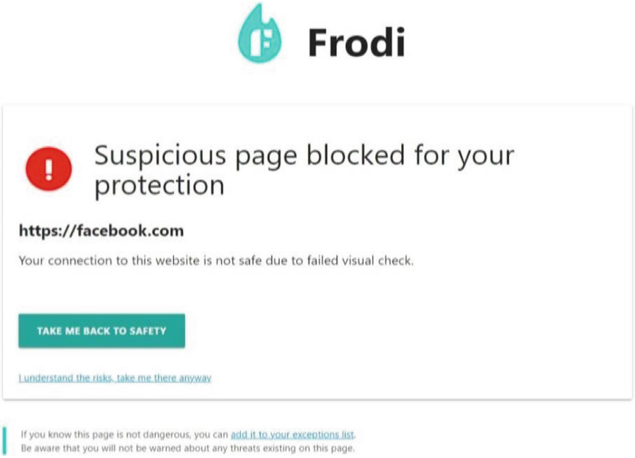
Your connection to this website is not safe due to IP mismatch. Potential MITM Attack!

I understand the risks, take me there anyway

If you know this page is not dangerous, you can add it to your exceptions list.
Be aware that you will not be warned about any threats existing on this page.

Fig. 4. DNS and Who is check.

Method- 4: Visual Similarity
Phishing websites generally create a clone of the UI of the original website. Our model will check for similarity between snapshot of the website visited and the original website and will flag accordingly. This check is the highest efficient check which flags or detects the malicious website when a user enters. The result of visual similarity is shown in Fig. 5.



Suspicious page blocked for your protection

https://facebook.com

Your connection to this website is not safe due to failed visual check.

I understand the risks, take me there anyway

If you know this page is not dangerous, you can add it to your exceptions list.
Be aware that you will not be warned about any threats existing on this page.

Fig. 5. Visual similarity check

5 Conclusion

Phishing is a severe cyberattack that affects tens of lakhs of people every day. It has risen over time as more individuals turn to the internet. We need a dependable strategy to stop these cyber crooks from robbing people of their money. So, in this proposed system we developed system proposed four different methods, and they are Statistic Database check, DNS and who is check, Verifying Ascii content of the URL and Visual similarity for effective detection by the help of Visual Studio and Golang. For the result & experimentation purpose, we have created a fake website to check if our system is working correctly and we have also checked for the real time websites & these strategies have yielded impressive outcomes for us.

References

1. Gururaj, H.L., BoreGowda, G.: Phishing website detection based on effective machine learning approach. J. Cyber Secur. Technol. **5**. 1–14 (2020). https://doi.org/10.1080/23742917.2020.1813396
2. Subasi, A., Molah, E., Almkallawi, F., Chaudhery, T.: Intelligent phishing website detection using random forest classifier. In: 2017 International Conference on Electrical and Computing Technologies and Applications (ICECTA), pp. 1–5 (2017)

3. Tan, C.L., Chiew, K.L., Sze, S.: Phishing website detection using URL-assisted brand name weighting system. In: 2014,International Symposium on Intelligent Signal Processing and Communication Systems, ISPACS 2014, pp. 54–59 (2015)https://doi.org/10.1109/ISPACS.2014.7024424
4. Yang, P., Zhao, G., Zeng, P.: Phishing website detection based on multidimensional features driven by deep learning. IEEE Access **7**, 15196–15209 (2019). https://doi.org/10.1109/ACCESS.2019.2892066
5. Aburrous, M., Hossain, M.A., Thabatah, F., Dahal, K.: Intelligent phishing website detection system using fuzzy techniques. In: 2008 3rd International Conference on Information and Communication Technologies: from Theory to Applications, pp. 1–6 (2008). https://doi.org/10.1109/ICTTA.2008.4530019
6. Ahmed, A., Abdullah, N.A.: Real time detection of phishing websites. In: 2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), pp. 1–6, (2016).https://doi.org/10.1109/IEMCON.2016.7746247
7. Chang, E.H., Chiew, K.L., Sze, S.N., Tiong, W.K.: Phishing detection via identification of website identity. In: 2013 International Conference on IT Convergence and Security (ICITCS), pp. 1–4 (2013).https://doi.org/10.1109/ICITCS.2013.6717870
8. Efe-Odenema, O., Jaiswal, J.: (2020). Issue 6 www.jetir.org (ISSN- 2349–5162)
9. Alkhozae, M.G., Batarfi, O.A.: Phishing websites detection based on phishing characteristics in the webpage source code. Int. J. Inf. Commun. Technol. Res. **1**(6) (2011)
10. Fu, A.Y., Wenyin, L., Deng, X.: Detecting phishing web pages with visual similarity assessment based on earth mover's distance (EMD). IEEE Trans. Dependable Secure Comput. **3**(4), 301–311 (2006). https://doi.org/10.1109/TDSC.2006.50
11. Huh, J.H., Kim, H.: Phishing detection with popular search engines: simple an effective. In: Proceedings of the 4th Canada- France MITACS Conference on Foundations and Practice of Security, ser. FPS 2011. Berlin, Heidelberg: Springer-Verlag, pp. 194–207 (2012). https://doi.org/10.1007/978-3-642-27901-0 15
12. Ramesh, G., Krishnamurthi, I., Kumar, K.S.S.: An efficacious method for detecting phishing webpages through target domain identification. Decision Support Syst. **61**, 12–22 (2014)
13. Pan, Y., Ding, X.: Anomaly based web phishing page detection. In: 2006 22nd Annual Computer Security Applications Conference (ACSAC 2006), pp. 381–392 (2006)
14. Zhang, Y., Hong, J.I., Cranor, L.F.: Cantina: a content-based approach to detecting phishing web sites. In: Proceedings of the 16th international conference on World Wide Web (WWW 2007). Association for Computing Machinery, New York, NY, USA, pp. 639–648 (2007).https://doi.org/10.1145/1242572.1242659
15. Dunlop, M., Groat, S., Shelly, D.: Goldphish: using images for content-based phishing analysis. In: 2010 Fifth International Conference on Internet Monitoring and Protection, (ICIMP), pp. 123–128. IEEE (2010)
16. Garera, S., Provos, N., Chew, M., Rubin, A.D.: A framework for detection and measurement of phishing attacks. In: Proceedings of the 2007 ACM workshop on Recurring malcode (WORM 2007). Association for Computing Machinery, New York, NY, USA, pp. 1–8 (2007). https://doi.org/10.1145/1314389.1314391
17. Zhang, Y., Egelman, S., Cranor, L., Hong, J.: Phinding phish: Evaluating anti-phishing tools (2007)
18. S.Al-Sharafat, W.: Development of genetic-based machine learning for network intrusion detection (GBML-NID)". world academy of science, engineering and technology, open science index 31. Int. J. Comput. Inf. Eng. **3**(7), 1677–1681 (2009)
19. Anti-Phishing Working Group Phishing, Anti-Phishing Working Group Phishing Trends Report (2014)