IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 12, Issue 13, April 2023

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Volume 12, Issue 13, April 2023 ||

International Conference on Recent Development in Engineering and Technology – ICRDET'23

Organized by

Dhaanish Ahmed Institute of Technology, KG chavadi, Coimbatore, India

An Effective and Accurate Phishing Websites Detection using Multiple Features

Gowrishankar.R [1], Mohamed Hashim [2], Mohammed Sulpikar.M [3]

Assistant Professor, Department of Computer Science & Engineering Dhaanish Ahmed Institute Of Technology, CoimbatoreTamil Nādu, India [1]

U.G. Scholar, Department of Computer Science & Engineering, Dhaanish Ahmed Institute Of Technology, Coimbatore, TamilNādu, India[2, 3]

ABSTRACT: Phishing attacks cost Internet users billions of dollars each year and are an ever-growing threat in cyberspace. It is illegal to collect sensitive information from consumers through a variety of social engineering techniques. Emails, instant messaging, pop-up messages, web pages, and other forms of communication can be used to detect phishing tactics. This work provides a model for determining whether a URL link is genuine or fraudulent. The dataset used for categorization comes from the University of New Brunswick database, which contains a collection of benign, malicious, and phishing URLs that includes phishing URLs in many forms such as CSV, JSON, etc., spam, malware, phishing, and defacement URLs. Phishing URLs are identified using a combination of deep neural network methods and more than six (6) machine learning models. The goal of this project is to develop web application softwarethat can identify phishing URLs from a database of more than 5,000 URLs that have been randomly selected, divided into 80,000 training samples and 20,000 test samples, and then divided again into equal proportions of phishing and legal URLs. To distinguish between legal and phishing URLs, the URL dataset is trained and evaluated using features such as address bar-based features, domain-based features, HTML-based features and JavaScript-based features

KEYWORDS: Feature extraction, classification, algorithm, and machine learning.

I. INTRODUCTION

We now rely heavily on the Internet to obtain and disseminate information, notably through social media. Pamela (2021) claims that the Internet is a network of computers that houses important data. As a result, several security measures are in place to safeguard that data, yet there is a weak point: the human. Security measures struggle far harderto secure a user's data and equipment when they easily give up access to their computers or other devices. Therefore, Imperia (2021) describes social engineering as a form of attack that is one of the most widespread social engineering assaults. This type of attack is used to acquire user data, including login passwords and credit card details. The assault occurs when an assailant dupes the victim into acting as though a text, instant message, or email were coming from a reliable source before opening it. When the receiver hits the link, they mistakenly think they've gotten a present and unknowingly click a harmful link, which leads to the installation of malware, the freezing of the machine during a ransom ware assault, or the release of private data. Due to the fast adoption of technological advancements, there has been a significant growth in computer security risks in recent years, which has also increased the susceptibility of human exploitation. Users should be informed of the methods used by phishes as well as ways to assist against falling victim to phishing.

II. RELATED WORK

Yuxiang Guan et. al. (2018) Analyse phishing website features and offers two types of web phishing detection features: domain and name-based features. In order to achieve high accuracy and efficiency, many sorts of characteristics have been extracted. The model has been strengthened by the extraction of several features. Samuel

Marchal et al. (2015) Used the attributes collected from words that constitute a URL based on query data from Google and Yahoo search engines, one idea of intra-URL relatedness was evaluated. These characteristics are then

|| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Volume 12, Issue 13, April 2023 ||

International Conference on Recent Development in Engineering and Technology – ICRDET'23

Organized by

Dhaanish Ahmed Institute of Technology, KG chavadi, Coimbatore, India

employed in a machine-learning-based categorization to identify phishing URLs in a real-world data collection. He has presented a phishing detection machine learning technique that uses solely lexical and domain features.

Vahid Shahrivari et al. (2019) On the phishing website dataset, which contains 6157 legal websites and 4898 phishing websites, multiple classifiers have been implemented and assessed. Logistic Regression, Decision Tree, Support Vector Machine, Ada Boost, and Random Forest are the classifiers that were tested. Different types of classifiers were used, and the results were compared. Random forest emerged as the most efficient classifier after the successful comparison. To evaluate the accuracy, a large data set was used.

Rishikesh Mahajan (2018) uses machine learning technologies to improve the detection process for phishing websites. Using the random forest technique, he achieved 94.14 percent detection accuracy with the lowest false positive rate. In addition, the results demonstrate that classifiers perform better when more data is utilised as training data

M. Kawabata and T Mustafa (2018) proposed feature selection techniques to reduce dataset components for higher order execution It also compared the outcomes of different data mining classification techniques. The phishing website dataset was obtained from the UCI machine learning library. He also proposes the use of a c4.5 decision tree approach to detect phishing URL websites. This method calculates heuristic values by extracting features from the sites. The c4.5 decision tree algorithm was given these values to determine if the site was phishing or not .

.

Gold Phish (2010)extracts page information from a displayed page using optical character recognition. It then employs search engines to determine whether the content of the page is consistent with its domain, so identifying phishing sites. He combines URLs, HTML DOM (Document object model), third-party services, and search engines to create the 15features. To detect phishing assaults, it trains these attributes using SVM (support vector machine).

III. METHODOLOGY

- **Data collection**

The dataset used for the classification was sourced from was gotten from multiple sources listed in the earlier stated methodology. The dataset used for classifying the dataset into phishing and legitimate URLs was sourced from open source websites.

- **Feature extraction on the datasets**

The features extraction used on the dataset are categorized into

i. Address bar based features
ii. Domain-based features
iii. Html & java-script based features

- **Data Analysis & Visualization**

The distribution plot of how legitimate and phishing datasets are distributed base on the features selected and how they are related to each other. the plot of a correlation heat-map of the dataset. The plot shows correlation between different variables in the dataset. the feature importance in the model for Decision tree classifier and Random forest classifier respectively.

- **Data pre-processing**

The datasets were first cleaned to remove empty entries and fill some entries by applying data pre-processing techniques and transform the data to use in the models. The summary of the dataset number of missing values in the dataset which all appear to be zero.

|| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Volume 12, Issue 13, April 2023 ||

International Conference on Recent Development in Engineering and Technology – ICRDET'23

Organized by

Dhaanish Ahmed Institute of Technology, KG chavadi, Coimbatore, India

- **Performance analysis**

The based methodology stated that the proposed system utilizes machine learning models and deep neural networks. These models consist of Decision Tree, Support Vector Machine, XGBooster, Multilayer Perceptions, Auto Encoder Neural Network, and Random Forest. The models determine whether a website URL is phishing or legitimate. The models help give a 2-class prediction (legitimate (0) and phishing (1). In the model development process, over six (6) machine learning models and deep neural network algorithms all together were used to detect phishing URLs.

IV. EXPERIMENTAL RESULTS

This takes a string URL as input and returns a probability value (0-100) of URL being malicious. We declare a URL malicious if it crosses a probability value of 70%. To determine if a URL is malicious or legitimate we use a Neural Network trained on 600,000 URLs.

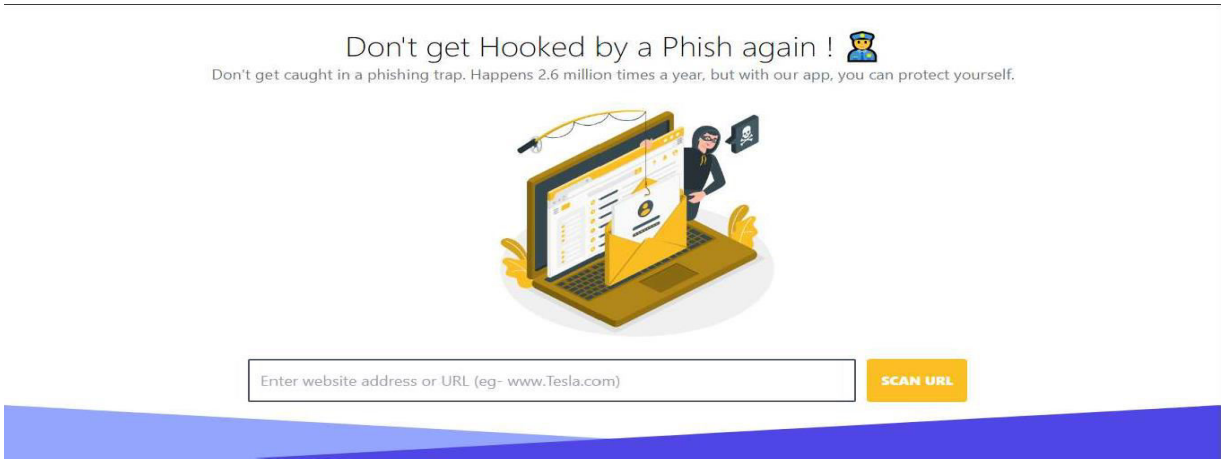

Figure 1

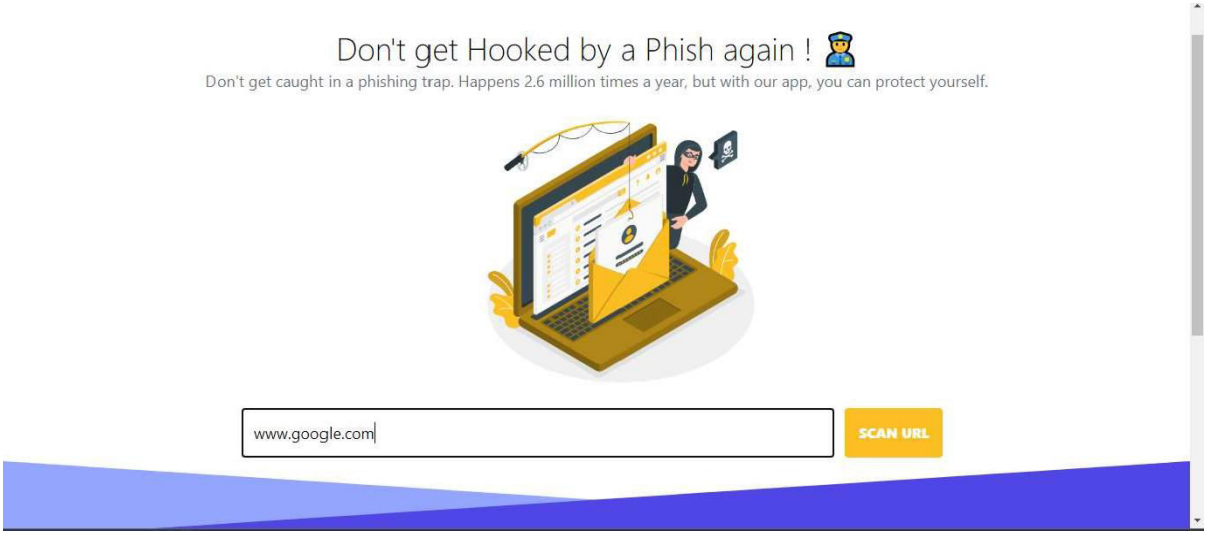

Figure 2

|| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Volume 12, Issue 13, April 2023 ||

International Conference on Recent Development in Engineering and Technology – ICRDET'23

Organized by

Dhaanish Ahmed Institute of Technology, KG chavadi, Coimbatore, India

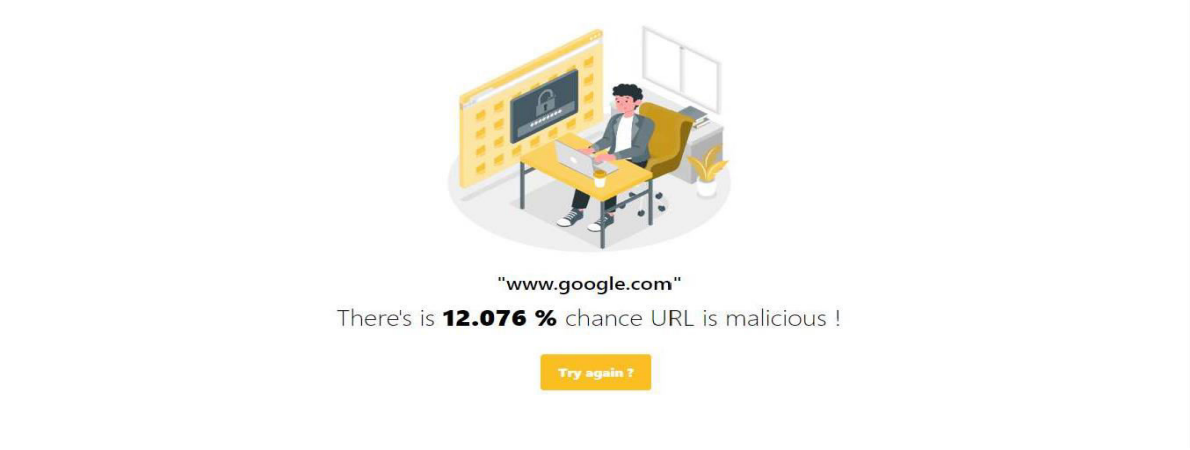

Figure 3

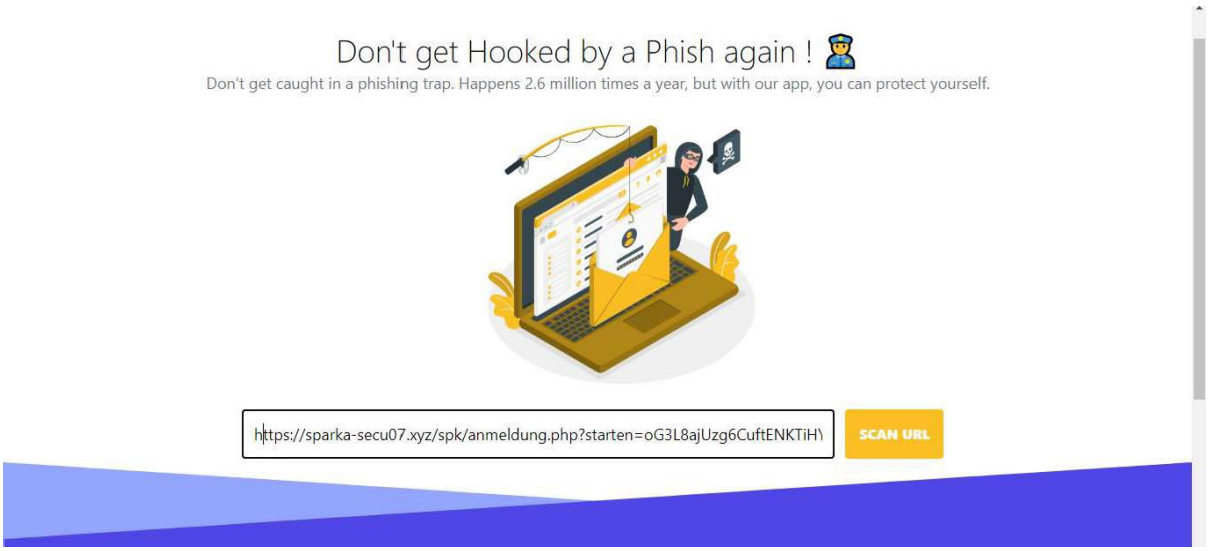

Figure 4

Figure 5

|| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Volume 12, Issue 13, April 2023 ||

International Conference on Recent Development in Engineering and Technology – ICRDET'23

Organized by

Dhaanish Ahmed Institute of Technology, KG chavadi, Coimbatore, India

V. CONCLUSION

This project was able to categorize and identify how phishers perform phishing attacks and the different ways researchers have used to detect phishers. Therefore, the system proposed in this project worked with different features and machine learning and deep neural networks such as decision tree, support vector machine, Xgbooster, multilayer perceptions, auto- encoder neural network and random forest to identify patterns that can easily detect URl links. The highest accuracy model based on the feature extraction algorithm used to detect phishing URLs and legitimate URL links was integrated into a web application where users can enter URL links from websites to identify whether they are legitimate or phishing links.

REFERENCES

1. Abu-Nimeh, S., Nappa, D., Wang, X. and Nair, S. 2007. A comparison of machine learning techniques for phishing detection. Anti-Phishing Working Groups Ecrime Researchers Summit, pp. 60–69.
2. Almseidin, M., Zuraiq. A.A., Al-kasassbeh, M. and Alnidami, N. 2019. Phishing detection based on machine learning and feature selection methods. International journal of interactive mobile technology, 13 : 171–183.
3. Cao, J.C., Qiang, L., Yuede, Ji., Yukun, He. and Dong, Guo. 2014. Detection of Forwarding-Based Malicious URLs in Online Social Networks. International Journal of Parallel Programming, pp. 1–18
4. Chou, N., Ledesma, R., Teraguchi, Y., Boneh, D. and Mitchell, J.C. 2004. Client-side defense against web-based identity theft.Proceedings of the 11th Annual Network and Distributed System Security Symposium (NDSS),pp. 76-82.
5. Dunlop, M., Groat, S. and Shelly, D. May 2010. Goldphish: Using images for content-based phishing analysis.Internet Monitoring and Protection (ICIMP), 5th International Conference on. IEEE, Barcelona, pp. 123–128.
6. Fette, I., Sadeh, N. and Tomasic, A. May 2007. Learning to detect phishing emails. Proceedings of the International World Wide WebConference (WWW), pp. 649-656.
7. Harinahalli, G.L. and BoreGowda, G. 2020 Phishing website detection based on effective machine learning approach. Journal of Cyber Security Technology, pp. 1-14.
8. Hassan, Y.A. and Abdelfettah, B. 2017. Using case- based reasoning for phishing detection. Procedia ComputerScience. 109 : 281–288.
9. Jain, A. K. and Gupta, B.B. 2018. PHISH-SAFE: URL Features-Based Phishing Detection System Using Machine Learning. Cyber Security. Advances in Intelligent Systems and Computing. 729 : 147-152.
10. Kumar, J., Santhanavijayan, A., Janet, B., Rajendran, B. and Bindhumadhava, B.S. 2020. Phishing Website Classification and Detection Using Machine Learning. International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, pp. 1–6.
11. Lee, L.H., Lee, K.C., Juan, Y.C., Chen, H.H, and Tseng, Y.H. 2014. Users Behavioral Prediction for Phishing Detection. Proceedings of the 23rd International Conference on World Wide Web. No. 1, pp. 337–338.
12. Mao, J., Tian, W., Li, P., Wei, T. and Liang, Z. 2017. Phishing website detection based on effective css features of web pages. 12th International Conference on Wireless Algorithms, Systems, and Applications, pp. 804–815.
13. Medvet, E., Kirda, E. and Kruegel, C. September 2008. Visual-similarity-based phishing detection. Proceedings of Secure Comm ACM. 129 : pp.224-230
14. Nourian, A., Ishtiaq, S. and Maheswaran, M. 2009. Castle: A social framework for collaborative anti-phishing databases. ACM Transactions on Internet Technology, pp.Pan,
15. Y. and Ding, X. 2006. Anomaly based web phishing page detection. Computer Security Applications Conference, pp. 381–392.
16. R, Mahajan. 2018. Phishing Website Detection using Machine Learning Algorithms. International Journal of Computer Applications. 123: pp. 45-47.
17. Srinivasa, R.R, and Pais, A. R. 2017. Detecting phishing websites using automation of human behavior.Proceedings of the 3rd ACM workshop on cyber-physical system security, pp 33–42.
18. Tan, C.L., Chiew, K.L, and Wong, K. 2016. Phish WHO: phishing webpage detection via identity keywords extraction and target domain name finder. Decision Support Systems. 88 : 18–27.
19. Wu, C.Y., Kuo, C.C, and Yang, C.S. 2019. A phishing detection system based on machine learning. International Conference on Intelligent Computing and its Emerging Applications (ICEA), pp 28–32

INNO SPACE
SJIF Scientific Journal Impact Factor

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details