**MACHINE LEARNING BASED PHISHINING URL CYBERCRIME DETECTION**

DR .Sangeetha Kalyaniraman\*1,Dommaraju Poojitha\*2,Keerthika.R\*3,Poona Sowmya Reddy \*4

2,3,4Student, Department of Computer Science & Engineering, Panimalar Engineering College, Chennai, Tamil Nadu, India

1Assistant Professor, Department of Computer Science & Engineering, Panimalar Engineering College, Chennai, Tamil Nadu, India

**ABSTRACT:**

A phishing attack is the easiest way to get sensitive information from innocent users. Phishers aim to obtain important information such as username, password and bank account details. Cybersecurity professionals are now looking for reliable and stable detection techniques to detect phishing websites. This article discusses machine learning technology to detect phishing URLs by extracting and analyzing various features of legitimate and phishing URLs. Logistic Regression, Random Forest and Support vector machine algorithm ,…..etc. are used to detect phishing sites. The goal of the article is to detect phishing URLs and narrow down to the best machine learning algorithm by comparing each algorithm's accuracy rate, false positive rate, and false negative rate.

# INTRODUCTION:

In recent years, the use of web services has skyrocketed due to the current digital transformation. Companies motivate change by providing their services online such as e-banking, e-commerce, or SaaS (Software as a Service) [1]. Currently, due to the COVID-19 pandemic, restrictions have extended work -a model from home that includes millions more workers, students and teachers developing their activities at a distance [2], which leads to substantial additional load for services such as email, student platforms, VPN or company portals. Therefore, there are even more potential targets exposed to phishing attacks where phishers try to impersonate legitimate websites to steal user credentials or payment information [3], [4]. Recent studies [5], [6] concluded that phishing is one of most significant social engineering attacks during the COVID-19pandemic, along with spam emails and websites to carry out these attacks.

Identifying phishing sites through their HTTP protocol is no longer a valid rule. in10in Q3 2017 [7] APWG reported that less than 25% of phishing websites were hosted under the HTTPS protocol, while this amount increased to 83% in Q1 2021 [8]. This website provides a secure end-to-end communication that, when created, conveys a false impression of security to the user online transactions [9]. Furthermore, the anti-phishing working group(APWG) [10] reported a significant increase in phishing attacks, i.e. from 165;772 to 611; 877 websites, just between the \_first quarter of 2020 and 2021respectively. The reason for this increase may be that people have resorted (a still are) to online services during the COVID-19 pandemic.Phishing is a social engineering cyber attack where criminals trick usersobtain their credentials through a login form that sends the data to the malicious entity

server. In this article, we compare machine learning and deep learning techniques present a method capable of detecting phishing sites using URL analysis. in state-of-the-art phishing detection solutions, the legitimate class consists of homepages without login forms. On the contrary, in both classes we use the URLs from the login page because we think it's correct they are much more representative of a real case scenario and we demonstrate that it exists techniques get a high rate of false positives when testing with URLs from legitimate login page. In addition, we use datasets from different years to show how models decrease their accuracy over time by training the base model with old datasets and test it with recent URLs. We also perform frequency analysis of the current phishing domains to identify the various techniques carried out by phishers in their campaigns. To prove these claims, we created a new dataset called Phishing Index Login URL (PILU-90K) which consists of 60K legitimate URLs, including index and login sites and 30,000 phishing URLs. Finally presentment a logistic regression model which, in combination with term frequency -Inverse Document Frequency Feature Extraction (TF-IDF), gets 96:50% accuracy on an established dataset of login URLs.

# II. EXISTINGSYSTEM

In existing system URL against phishing database. Some examples of this solution are Google SafeBrowsing, PhishTank, OpenPhish or SmartScreen. If the requested URL matches any record, the request will be blocked and the user will be warned before visiting the website. However, despite the possibilities of a list-based approach, it would fail if the phishing URL was not reported earlier and would require constant efforts to update the database with more recent phishing data. We found that many phishing URLs were removed from Phishtank after the first day, while OpenPhish removed all URLs from its reporting after seven days. This issue allows attackers to reuse the same URL when it is removed from different lists .

**III . PROPOSED APPROACH :**

The proposed system consists of 2 steps. This article presents a dataset of phishing URLs using legitimate login sites to obtain URLs from such sites. We then evaluate machine and deep learning techniques to recommend a method with higher accuracy. Next, we show how models trained with legitimate homepages struggle to classify legitimate login URLs, demonstrating our hypothesis on detecting phishing and legitimate login URLs. Next, we show how accuracy decreases over time on models trained with datasets from 2016 and evaluated on data collected in 2021. Finally, we provide an overview of current phishing encounters, explaining attack tricks and approaches.The main contributions of the post can be summarized as follows:

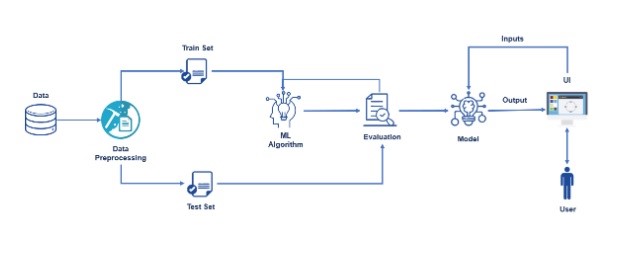
• Expanded our previous PILU-60K (Phishing Index Login URL) dataset from 60,000 to 90,000 URLs evenly divided into three classes: Phishing, Legitimate Homepage, and Legitimate Login. We are making this extended data set, PILU-90K, available to the public for research purposes

• Using PILU-90K, we implemented and evaluated three channels for detecting phishing URLs: (i) encountered. Depending on their input data, these approaches

• They can be divided into two categories: URL-based and content-based.

# 

**IV. ARCHITECTURAL DIAGRAM**



* respective classes and hence reduces the chance of malicious attachments.
* machine learning methods should be trained with recent URLs to prevent substantial ageing from the date of its release.
* learning models using handcrafted URL features with increased performance,

**V.MODULES:**

**• Dataset Collection**

**• Algorithm**

**• Detection**

**VII. CONCLUSION**

* This project aims to enhance detection methods to detect phishing websites using machine learning technology.
* The final takeaway from this project is to explore various machine learning models, perform Exploratory Data Analysis on phishing datasets and understand their features.
* Gradient Boosting Classifier currently classify URL up to 97.4%.

**VIII. REFERNCE**

[1] J. Spaulding, S. Upadhyaya, and A. Mohaisen, ‘‘The landscape of domain

name typosquatting: Techniques and countermeasures,’’ in Proc. 11th Int.

Conf. Availability, Rel. Secur. (ARES), Aug. 2016, pp. 284–289.

[2] W. H. E. Day and H. Edelsbrunner, ‘‘Efficient algorithms for agglomerative

hierarchical clustering methods,’’ J. Classification, vol. 1, no. 1, pp. 7–24,

1984.

[3] C. Peersman, C. Schulze, A. Rashid, M. Brennan, and C. Fischer, ‘‘ICOP:

Live forensics to reveal previously unknown criminal media on P2P

networks,’’ Digit. Invest., vol. 18, pp. 50–64, Sep. 2016.

[4] L. Li, E. Berki, M. Helenius, and S. Ovaska, ‘‘Towards a contingency approach with whitelist- and blacklist-based anti-phishing applications:What do usability tests indicate?’’ Behaviour Inf. Technol., vol. 33, no. 11,pp. 1136–1147, Nov. 2014.

[5] N. Samarasinghe and M. Mannan, ‘‘On cloacking behaviours of malicious websites,’’ Comput. Secur., vol. 101, pp. 102–114, Feb. 2021.

[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. 51–59.

[7] E. Buber, B. Diri, and O. K. Sahingoz, ‘‘NLP based phishing attackdetection from URLs,’’ in Proc. Int. Conf. Intell. Syst. Design Appl.,vol. 736, 2018, pp. 608–618.

[8] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, ‘‘Machine learning based

phishing detection from URLs,’’ Expert Syst. Appl., vol. 117, pp. 345–357,

Mar. 2019.

[9] G. Xiang, J. Hong, C. P. Rose, and L. Cranor, ‘‘CANTINA+: A feature-rich machine learning framework for detecting phishing web sites,’’ ACMTrans. Inf. Syst. Secur., vol. 14, no. 2, pp. 1–28, Sep. 2011.

[10] M. Moghimi and A. Y. Varjani, ‘‘New rule-based phishing detection

method,’’ Expert Syst. Appl., vol. 53, pp. 231–242, Jul. 2016.

[11] R. S. Rao and A. R. Pais, ‘‘Detection of phishing websites using an

efficient feature-based machine learning framework,’’ Neural Comput.Appl., vol. 31, no. 8, pp. 3851–3873, Aug. 2019.

[12] E. Buber, B. Diri, and O. K. Sahingoz, ‘‘NLP based phishing attackdetection from URLs,’’ in Proc. Int. Conf. Intell. Syst. Design Appl.,vol. 736, 2018, pp. 608–618.

[13] R. De’, N. Pandey, and A. Pal, ‘‘Impact of digital surge during COVID-

19 pandemic: A viewpoint on research and practice,’’ Int. J. Inf. Manage.,vol. 55, Dec. 2020, Art. no. 102171.

[14] Phishing Activity Trends Report 3Q, Anti-Phishing Working Group,

International, 2017. Accessed: Sep. 12, 2021.

[15] Phishing Activity Trends Report 4Q, Anti-Phishing Working Group,

International, 2020. Accessed: Sep. 12, 2021.

[16] M. A. Adebowale, K. T. Lwin, E. Sánchez, and M. A. Hossain, ‘‘Intelligent

web-phishing detection and protection scheme using integrated features of images, frames and text,’’ Expert Syst. Appl., vol. 115, pp. 300–313,

Jan. 2019.

[17] M. Rami, M. Lee, and T. Fadi, ‘‘UCI machine learning repository,’’ Univ.

Huddersfield, Huddersfield, U.K., Tech. Rep., 2015.

[18] L. Yang, J. Zhang, X. Wang, Z. Li, Z. Li, and Y. He, ‘‘An improved ELM-

based and data preprocessing integrated approach for phishing detection considering comprehensive features,’’ Expert Syst. Appl., vol. 165,

Mar. 2021, Art. no. 113863.

[19] F. Sadique, R. Kaul, S. Badsha, and S. Sengupta, ‘‘An automated framework for real-time phishing URL detection,’’ in Proc. 10th Annu. Comput. Commun. Workshop Conf. (CCWC), Jan. 2020, pp. 335–341.

[20] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, ‘‘Distributed

representations of words and phrases and their compositionality,’’ in Proc.

Adv. Neural Inf. Process. Syst., vol. 26, no. 4, Dec. 2013, pp. 3111–3119.

[21] M. Somesha, A. R. Pais, R. S. Rao, and V. S. Rathour, ‘‘Efficient deep

learning techniques for the detection of phishing websites,’’ Sadhan ̄ a ̄,

vol. 45, no. 1, pp. 1–18, Dec. 2020.

[22] A. Aljofey, Q. Jiang, Q. Qu, M. Huang, and J.-P. Niyigena, ‘‘An effective

phishing detection model based on character level convolutional neural

network from URL,’’ Electronics, vol. 9, no. 9, pp. 1–24, 2020.

[23] A. Al-Alyan and S. Al-Ahmadi, ‘‘Robust url phishing detection based

on deep learning,’’ KSII Trans. Internet Inf. Syst., vol. 14, no. 7,

pp. 2752–2768, 2020.

[24] J. Zhao, N. Wang, Q. Ma, and Z. Cheng, ‘‘Classifying malicious URLs

using gated recurrent neural networks,’’ in Proc. Int. Conf. Innov. Mobile

Internet Services Ubiquitous Comput., vol. 773, 2019, pp. 385–394.

[25] X. Zhang, J. Zhao, and Y. Lecun, ‘‘Character-level convolutional networks

for text classification,’’ in Proc. Adv. Neural Inf. Process. Syst., Jan. 2015,

pp. 649–657.

26] Y. Kim, ‘‘Convolutional neural networks for sentence classification,’’ in

Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), 2014,

pp. 1746–1751.

[27] K. L. Chiew, C. L. Tan, K. Wong, K. S. C. Yong, and W. K. Tiong, ‘‘A new

hybrid ensemble feature selection framework for machine learning-based

phishing detection system,’’ Inf. Sci., vol. 484, pp. 153–166, May 2019.

[28] R. S. Rao, T. Vaishnavi, and A. R. Pais, ‘‘CatchPhish: Detection of phishing

websites by inspecting URLs,’’ J. Ambient Intell. Humanized Comput.,

vol. 11, no. 2, pp. 813–825, Feb. 2020.

[29] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu,

‘‘LightGBM: A highly efficient gradient boosting decision tree,’’ in Proc.

Adv. Neural Inf. Process. Syst., Dec. 2017, pp. 3147–3155.

[30] T. Chen and C. Guestrin, ‘‘XGBoost: A scalable tree boosting system,’’

in Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., vols. 13–17,

2016, pp. 785–794.

[31] A. Aizawa, ‘‘An information-theoretic perspective of tf–idf measures,’’ Inf.

Process. Manage., vol. 39, no. 1, pp. 45–65, 2003.

[32] M. W. Al-Nabki, E. Fidalgo, E. Alegre, and R. Aláiz-Rodríguez, ‘‘File

name classification approach to identify child sexual abuse,’’ in Proc. 9th

Int. Conf. Pattern Recognit. Appl. Methods, 2020, pp. 228–234.

[31] Zhang, J. Zhao, and Y. Lecun, ‘‘Character-level convolutional networks

for text classification,’’ in Proc. Adv. Neural Inf. Process. Syst., Jan. 2015,

pp. 649–657.