IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, May 2022
Impact Factor: 6.252

Detection of Phishing Website Using Machine Learning

P. Y. Pawar[1], Faheem Shaikh[2], Pooja Garg[3], Kunal Rokde[4], Omkar Shelke[4]
Assistant Professor, Department of Information Technology[1]
UG Scholar, Department of Information Technology[2,3,4,5]
Sinhgad Academy of Engineering, Pune, Maharashtra, India
Savitribai Phule Pune University, Pune, Maharashtra, India

Abstract: *Phishing is a common method of tricking people int disclosing their entire personal information by using fake websites. Phishing records process tool URLs are used to steal personal information such as customer names, passwords, and online banking activities. Phishers (assailants) employ websites with rectangular diplomas that are visually and semantically similar to the real ones. As the century progressed, phishing strategies advanced swiftly, and this might be avoided by employing anti-phishing technologies to detect phishing. A strong gadget that is frequently utilized in the direction of phishing attacks is machine planning to apprehend. The capabilities used for detection and detection strategies by using Machine Learning have also been investigated in the suggested system.*

Keywords: Phishing, Phishing Websites, Detection, Machine Learning

I. INTRODUCTION

Phishing is the practice of imitating the characteristics and alternatives of emails in order to make them appear identical to the genuine. It appears to be similar to the authentic supply. This e-mail appears to have come from a legitimate employer or organization, according to the consumer. This forces the user to visit the phishing website by clicking on the links provided in the phishing email. These phishing websites were made to imitate the seams of a clever website. Phishers compel people to keep track of their personal information by sending them baleful messages or validating account messages, for example, so that they can keep track of the information they want to misuse.

They design things so that the user doesn't always have a choice but to visit their fake website. In the cyberspace, phishing is one of the most dangerous illegal physical acts. Since the majority of users log on to gain access to services provided by government and financial institutions, there has been a considerable increase in phishing attacks in recent years. Phishers began to earn money and attempted to turn this into a lucrative company. Phishing may be illegal, but the reason why phishers do this crime is that it is incredibly trustworthy, it doesn't cost anything, and it is effective.

Phishing will actually gain access to someone's e-mail identity. It's incredibly earnest to look for out someone's email identity every day, and you'll send an email to everyone who is freely available throughout the world. These assailants have a much less money and power to get critical information quickly and effectively. Malware infections, data loss, fraud, and other issues are all affected by phishing scams. The important data of a user, such as the password, OTP, sensitive know-how associated with company, medical understanding, secret information, and so on, is something that those cyber crooks are interested in at some point. Frequently, these fraudsters obtain data that allows them to have direct access to their social media accounts and emails. For phishing detection, a variety of software/methods and algorithms are employed. These are utilised in both academic and industrial settings. A phishing address and also the parallel online page have numerous characteristics that are distinct from the address. For example, to cover the initial domain selection, the phishing assaulter will sense a terribly long and complex domain name. This is frequently and glaringly obvious.

II. OBJECTIVES

- To compare various categorization data mining algorithmic methodologies, as well as different feature selection scenarios.
- To create a multi-classifier integration model that combines clustering and several classification techniques to improve phishing website detection and protection.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, May 2022
Impact Factor: 6.252

- To find the best phishing detection classification algorithm.

III. Literature Survey

A. Lakshmanarao and P. Surya Prabhakara Rao, "Phishing website detection using novel machine learning fusion approach", IEEE 2021

For phishing detection, different machine learning methods such as logistic regression, decision tree classifier, random forest classifier, AdaBoost, and gradient boosting classifier were used.

Jitendra Kumar and A. Santhanavijayan, "Phishing Website Classification and Detection Using Machine Learning", International Conference on Computer Communication and Informatics, 2020

Logistic regression, Gaussian Nave Bayes, and Random Forest were proposed in this study.

Mehmet Korkmaz and Ozgur Koray Sahingoz, "Detection of Phishing Websites by Using Machine Learning-Based URL Analysis", IEEE 2020

Using eight distinct algorithms to analyse URLs and three different datasets to compare the results with other research, a machine learning-based phishing detection system was proposed.

Charu Singh, "Phishing Website Detection Based on Machine Learning: A Survey",IEEE 2020

A review was completed. With the large amount of phishing emails or messages received every day, companies or people are unable to detect all of them, where many reviews for phishing assault detection using machine learning have been presented.

Vaibhav Patil and Pritesh Thakkar, "Detection and Prevention of Phishing Websites using Machine Learning Approach", IEEE 2018

Three methods for identifying phishing websites were proposed. The first strategy examines various elements of the URL, the second examines the authenticity of the website by learning where it is housed and who manages it, and the third way examines the website's visual look.

T. Peng, I. Harris, and Y. Sawa, "Detecting Phishing Attacks Using Natural Language Processing and Machine Learning," Proc. - 12th IEEE Int. Conf. Semant. Comput. ICSC 2018, vol. 2018–Janua, pp. 300–301, 2018

Using linguistic communication analysis and machine learning, this paper proposes a method for detecting phishing email attempts. To detect malicious intent, it is customary to search the text's syntax. To decode each sentence and identify the semantic jobs of words inside the sentence, a natural language processing (NLP) technique is used in conjunction with a predicate. The blacklist of harmful pairs is created via computer supervised learning

IV. Implementation Details Of Module

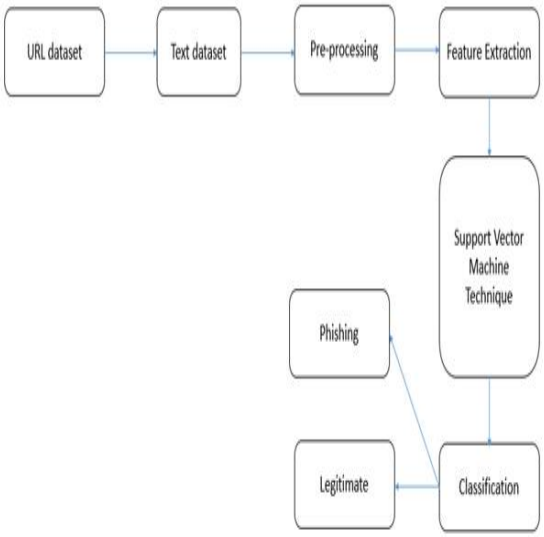

Figure: System Architecture

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, May 2022
Impact Factor: 6.252

The application receives a dataset of phishing and legal URLs, which is then pre-processed so that the facts are in a usable manner for analysis. The functions revolve around numerous phishing website characteristics that have been utilized to distinguish them from authentic websites. Each category is characterized by its own set of phishing properties and values. For each URL, the specified attributes are extracted, and valid stages of inputs are found. Each phishing internet site risk is then assigned one of these values. The phishing properties esteems are addressed with a double no 0 and 1, indicating whether or not the characteristic is present.

4.1 Support Vector Machine

- Bring the data into the programme.
- Investigate the data to see how they appear.
- Pre-process the information
- Organize the information into attributes and labels.
- Separate the data into two groups: training and testing.
- The SVM algorithm should be trained.
- Make some forecasts.
- Analyze the algorithm's results.

S. L. Bangare et al. [7-14] have worked in the brain tumor detection. N. Shelke et al [15] given LRA-DNN method. Suneet Gupta et al [16] worked for end user system. Gururaj Awate et al. [17] worked on Alzheimers Disease. P. S. Bangare et al [18] worked on the object detection. Kalpana Thakare et al [19-24] have worked on various machine learning algorithms. M. L. Bangare et al. [25] worked on the cloud platform.

V. CONCLUSION

Education awareness is the most significant strategy to protect users from phishing attacks. Internet users should be aware of all security recommendations made by professionals. Every user should also be taught not to follow links to websites that require them to enter sensitive information on a whim. It is critical to double-check the URL before proceeding to the website. It has escalated into a severe network security issue, resulting in billions of dollars in losses for both customers and e-commerce enterprises. Phishing has made e-commerce distrusted and appealing to ordinary consumers, and it may do so even more in the future. Education awareness is the most significant strategy to protect users from phishing attacks. Internet users should be aware of all security recommendations made by professionals. Every user should also be taught not to follow links to websites that require them to enter sensitive information on a whim. Before adding a website into the suggested system, it is critical to double-check the URL.

REFERENCES

[1]. A. Lakshmanarao and P.Surya Prabhakara Rao, “Phishing website detection using novel machine learning fusion approach ”, IEEE 2021

[2]. Jitendra Kumar and A. Santhanavijayan , “Phishing Website Classification and Detection Using Machine Learning ”, International Conference on Computer Communication and Informatics, 2020

[3]. Mehmet Korkmaz and Ozgur Koray Sahingoz, “Detection of Phishing Websites by Using Machine Learning-Based URL Analysis”, IEEE 2020

[4]. Charu Singh , “Phishing Website Detection Based on Machine Learning: A Survey”,IEEE 2020

[5]. Vaibhav Patil and Pritesh Thakkar , “Detection and Prevention of Phishing Websites using Machine Learning Approach”, IEEE 2018

[6]. T. Peng, I. Harris, and Y. Sawa, “Detecting Phishing Attacks Using Natural Language Processing and Machine Learning,” Proc. - 12th IEEE Int. Conf. Semant. Comput. ICSC 2018, vol. 2018–Janua, pp. 300–301, 2018.

[7]. S. L. Bangare, “Classification of optimal brain tissue using dynamic region growing and fuzzy min-max neural network in brain magnetic resonance images”, Neuroscience Informatics, Volume 2, Issue 3, September 2022, 100019, ISSN 2772-5286, https://doi.org/10.1016/j.neuri.2021.100019 .

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, May 2022
Impact Factor: 6.252

[8]. S. L. Bangare, G. Pradeepini, S. T. Patil, "Implementation for brain tumor detection and three dimensional visualization model development for reconstruction", ARPN Journal of Engineering and Applied Sciences (ARPN JEAS), Vol.13, Issue.2, ISSN 1819-6608, pp.467-473. 20/1/2018 http://www.arpnjournals.org/jeas/research_papers/rp_2018/jeas_0118_6691.pdf

[9]. S. L. Bangare, S. T. Patil et al, "Reviewing Otsu's Method for Image Thresholding." International Journal of Applied Engineering Research, ISSN 0973-4562, Volume 10, Number 9 (2015) pp. 21777-21783, © Research India Publications https://dx.doi.org/10.37622/IJAER/10.9.2015.21777-21783

[10]. S. L. Bangare, G. Pradeepini, S. T. Patil, "Regenerative pixel mode and tumor locus algorithm development for brain tumor analysis: a new computational technique for precise medical imaging", International Journal of Biomedical Engineering and Technology, Inderscience, 2018, Vol.27 No.1/2. https://www.inderscienceonline.com/doi/pdf/10.1504/IJBET.2018.093087

[11]. S. L. Bangare, A. R. Khare, P. S. Bangare, "Quality measurement of modularized object oriented software using metrics", ICWET '11: Proceedings of the International Conference & Workshop on Emerging Trends in Technology, February 2011, pp. 771–774. https://doi.org/10.1145/1980022.1980190.1.

[12]. S. L. Bangare, G. Pradeepini and S. T. Patil, "Brain tumor classification using mixed method approach," 2017 International Conference on Information Communication and Embedded Systems (ICICES), 2017, pp. 1-4, doi: 10.1109/ICICES.2017.8070748.

[13]. S. L. Bangare, S. Prakash, K. Gulati, B. Veeru, G. Dhiman and S. Jaiswal, "The Architecture, Classification, and Unsolved Research Issues of Big Data extraction as well as decomposing the Internet of Vehicles (IoV)," 2021 6th International Conference on Signal Processing, Computing and Control (ISPCC), 2021, pp. 566-571, doi: 10.1109/ISPCC53510.2021.9609451.

[14]. S. L. Bangare, G. Pradeepini, S. T. Patil et al, "Neuroendoscopy Adapter Module Development for Better Brain Tumor Image Visualization", International Journal of Electrical and Computer Engineering (IJECE) Vol. 7, No. 6, December 2017, pp. 3643~3654. http://ijece.iaescore.com/index.php/IJECE/article/view/8733/7392

[15]. N. Shelke, S. Chaudhury, S. Chakrabarti, S. L. Bangare et al. "An efficient way of text-based emotion analysis from social media using LRA-DNN", Neuroscience Informatics, Volume 2, Issue 3, September 2022, 100048, ISSN 2772-5286, https://doi.org/10.1016/j.neuri.2022.100048 .

[16]. Suneet Gupta, Sumit Kumar, Sunil L. Bangare, Shibili Nuhmani, Arnold C. Alguno, Issah Abubakari Samori, "Homogeneous Decision Community Extraction Based on End-User Mental Behavior on Social Media", Computational Intelligence and Neuroscience, vol. 2022, Article ID 3490860, 9 pages, 2022. https://doi.org/10.1155/2022/3490860.

[17]. Gururaj Awate, S. L. Bangare, G. Pradeepini and S. T. Patil, "Detection of Alzheimers Disease from MRI using Convolutional Neural Network with Tensorflow",arXiv, https://doi.org/10.48550/arXiv.1806.10170

[18]. P. S. Bangare, S. L. Bangare, R. U. Yawle and S. T. Patil, "Detection of human feature in abandoned object with modern security alert system using Android Application," 2017 International Conference on Emerging Trends & Innovation in ICT (ICEI), 2017, pp. 139-144, doi: 10.1109/ETIICT.2017.7977025

[19]. Kalpana S. Thakare, Viraj Varale, "Prediction of Heart Disease using Machine Learning Algorithm", Bioscience Biotechnology Research Communications (Special issue) Volume 13, Issue 12, 2020 (Dec 2020 issue).

[20]. Kalpana S. Thakare, A. M. Rajurkar, "Shot Boundary Detection of MPEG Video using Biorthogonal Wavelet Transform", International Journal of Pure and Applied Mathematics, Volume 118, No. 7, pp. 405-413, ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version), url: http://www.ijpam.eu

[21]. Kalpana S. Thakare, A. M. Rajurkar, R. R. Manthalkar, "Video Partitioning and Secured Key frame Extraction of MPEG Video", Proceedia Computer Science Journal, Volume 78, pp 790-798, Elsevier, 2016. Scopus DOI: htpp://10.1016/j.procs.2016.02.058, www.sciencedirect.com/science/article/pii/S1877050916000600

[22]. Kalpana S. Thakare, A. M. Rajurkar and R. R. Manthalkar, "Content based Video Retrieval using Latent Semantic Indexing and Color, Motion and Edge Features", International Journal of Computer Applications 54(12):42-48, September 2012, Published by Foundation of Computer Science, New York, USA. DOI: 10.5120/8621-2486

[23]. Kalpana S. Thakare, Archana M. Rajurkar, R. R. Manthalkar, "A Comprehensive System Based on Spatiotemporal Features Such as motion, Quantized Color and Edge Features", International Journal of Wireless

and Microwave Technologies (IJWMT) ISSN 1449 (Print), ISSN: 2076-9539 (Online), Vol.1, No.3, June. 2011, DOI: 10.5815 /ijwmt

[24]. Kalpana S. Thakare, Archana M. Rajurkar, Dr. R. R. Manthalkar, "An effective CBVR system based on Motion, Quantized color and edge density features", International Journal of Computer Science & Information Technology (IJCSIT), ISSN 0975 – 3826, Vol 3, No 2, April 2011 DOI: 10.5121/ijcsit.2011.3206 78.

[25]. M. L Bangare, Sarang A Joshi, "Kernel interpolation-based technique for privacy protection of pluggable data in cloud computing", International Journal of Cloud Computing, Volume 9, Issue 2-3, pp.355-374, Publisher Inderscience Publishers (IEL)