IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

"Detection of Phishing WebsitesUsing Machine Learning"

[1] Sugat Ingle, [2]Prince Gupta, [3] Shreyas Bhole, [4] Kamlesh Janawale, Prof. A.N. Kalal,

Department of Information Technology at Anantrao Pawar College of Engineering & Research, Pune-411009,India.

Abstract: The detection of phishing websites and online content yields various indicators. One prevalent form of successful cybercrime involves phishing sites that lure users to deceptive websites mimicking legitimate ones, aiming to illicitly obtain personal and sensitive information. The suggested Extreme Learning Machine (ELM)-based version has proven effective in identifying phishing websites. Internet page types exhibit diverse characteristics, requiring the utilization of a set of web page features for protection against phishing attacks. To counter these threats, a machine learning strategy is implemented. The phishing dataset, including authentic URLs from the database and collected data, undergoes pre-processing. Four groups of URL characteristics—domain-based, address-based, anomalous-based, and HTML or JavaScript features are employed for phishing detection. The analysed data is utilized to extract URL characteristics and generate corresponding attribute values. Machine learning approaches are applied to analyse URLs, establishing threshold and range values for URL properties. The project aims to develop an ELM categorization for various database characteristics and identify potential phishing sites.

Keywords: Extreme Learning Machine (ELM), Support Vector Machine (SVM), Random Forest Algorithm, URL Phishing Websites, Browser add-ons.

Introduction: Phishing constitutes a fraudulent activity wherein attackers impersonate trustworthy entities in electronic communication to illicitly acquire sensitive information like usernames, passwords, and credit card details for malicious purposes. Security researchers are deeply concerned about phishing attacks, primarily because attackers can effortlessly craft fake websites that closely mimic legitimate ones. Although experts can usually identify fake websites, many users, lacking such expertise, become unsuspecting victims. The primary objective of attackers is to steal banking credentials, leading to substantial financial losses for businesses. The conventional method for detecting phishing websites involves utilizing blacklisted URLs and IP addresses present in antivirus databases, known as the "blacklist" method. However, attackers can avoid these blacklists by employing tactics like complication and algorithmic generation of new URLs. Social engineering tactics, such as using deceptive email addresses and messages, also contribute significantly to the effectiveness of phishing attacks.

To create an illusion of legitimacy, attackers often employ social engineering tactics to deceive users. For instance, they might craft emails that appear to originate from a trusted source, urging the user to click on a link leading to a fraudulent website. These emails often contain urgent or

compelling language to coerce users into acting hastily without careful consideration. Another tactic employed by attackers is spear phishing, a targeted form of phishing that involves thorough research on the victim. Attackers create personalized messages that seem to be from someone the victim knows or trusts, significantly increasing the likelihood of the victim fallingfor the scam.

Literature Survey:

SK Hasane Ahammad , Sunil D. Kate , Gopal D. Upaddhye , Sandeep Dwarkanath Pnade el al , ELSEVIER Jan 1 2022 . [1] a machine learning model can be created with the help of all the algorithms discussed above, and for testing and training, the machine learing model and 80% of the dataset were used for training and 20% for testing. [2] Learning dataset gives Acuuracy measure's by decision tree , randomforest and other algorithms.

Mohammed Hazim Alkawaz, Stephanie Joanne Steven, AsifIqbal Hajamydeen, Detecting Phishing Websites Using Machine Learning, 2020 16th IEEE International Colloquium on Signal Processing & its Applications, 28-29 Feb. 2020 . [1] After reviewing and researching for appropriate monitoring tools, proposed system has been identified and chosen to address the complexity of monitoring requirement for current situation. This software is designed to show awareness of the extensive level of its functionality, features that can be displayed in the monitoring era . [2] In conclusion, this system is designed for resources are used as intended, prevents from valuable information from leaks out, produce better control mechanism and alerts the user to keep their private information safe.

Meenu, Sunila Godara, Phishing Detection using Machine Learning Techniques, IJEAT, ISSN: 2249 – 8958, 2, December,2019. [1] This investigation proposes framework that utilization machine learning systems to beat the spam issue. A model of the framework has been produced on the Azure stage and the conduct of email servers has been examined. Develop a phishing detection model by using various data mining techniques to enhance the phishing detection accuracy and a feature selection method
[2] Finally, the comparison various machine learning techniques like two class logistic regression technique and two class boosted decision tree (DT) ,two class neural network(NN) and two class support vector machine (SVM) and improved logistic regression is proposed to detect spam .

Ankit Kumar Jain and B.B. Gupta EURASIP Journal On Information Security (2016)2016:9.
[1] In this paper, we leraned a novel approach to protect against phishing attack using auto- updated white-list of legitimate sites accessed by the individual user. Furthermore, there approach is able to check the legitimacy of a webpage using hyperlink features. There experimental results showed that the proposed approach is very effective in protecting against phishing attacks as it has 86.02 % true positive rate with a very less false positive rate of1.48 %.

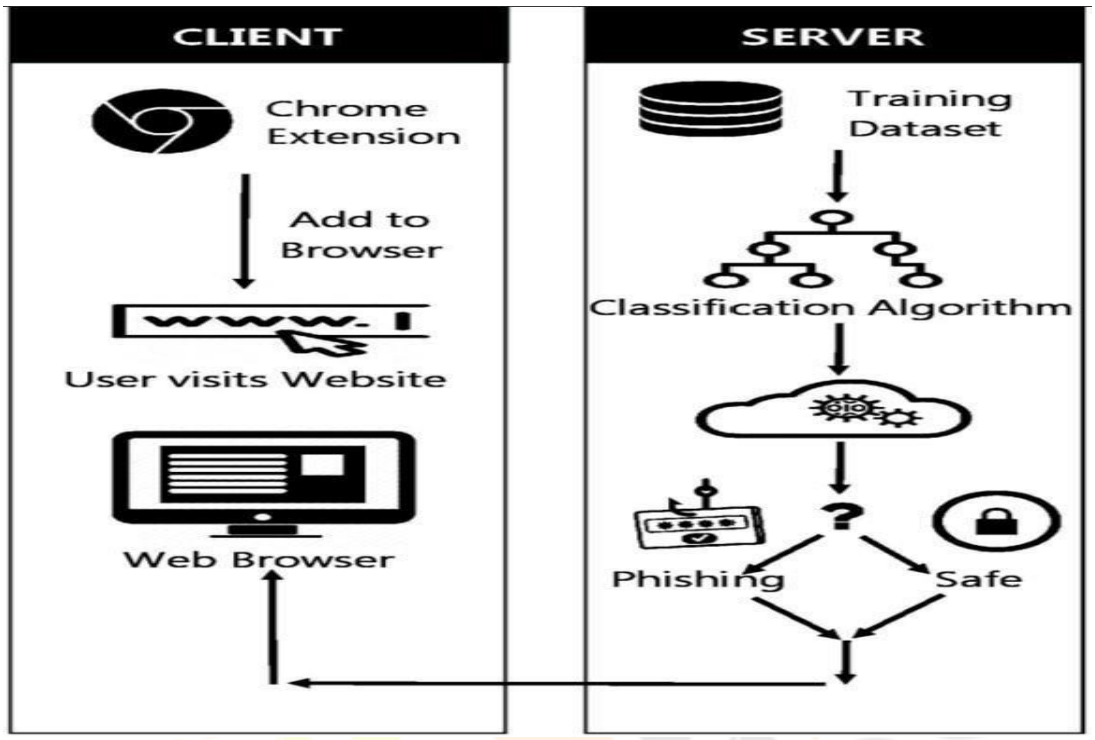
Proposed System:

The outlined approach involves importing sets of phishing information along with valid computer addresses from records, and subsequently preprocessing the imported data. Phishing webpage detection relies on the training of URL features, encompassing domain-based, address-based, abnormal-based, and HTML/JavaScript features. These URL characteristics areextracted from processed data, and values are generated for each feature.

For every URL attribute, values are computed through a machine learning method, determining both the range value and the threshold value for the URL attributes. Subsequently, the URL undergoes analysis, and based on

the calculated values, it is classified as either a phishing or an official URL. The feature values are derived from traits extracted from phishing websites, playing a crucial role in determining the range and threshold values.

System Architecture:



1. Feature Extraction:

Feature extraction involves the conversion of raw data into numerical features, ensuring the preservation of information from the original dataset. This method surpasses the direct application of machine learning to raw data, leading to more effective outcomes.

2. Trained Dataset:

Training data refers to an extensive dataset employed to educate machine learning models. Its purpose is to instruct prediction models, utilizing machine learning algorithms, in extracting pertinent features aligned with specific business objectives. In the context of supervised machine learning models, the training data is annotated or labelled. Conversely, for unsupervised machine learning models, the training data lacks labelling.

Machine Learning Algorithms:

Extreme Learning Machine (ELM): The Extreme Learning Machine (ELM) represents a one-hidden-layer Artificial Neural Network (ANN) model. To ensure advanced learning, the threshold value, weight, and activation parameters of an ANN must be appropriately set for the data system being modelled. Unlike gradient-based learning systems where these parameters are repeatedly adjusted, ELM adopts a different approach.

Random Forest Algorithm: The random forest (RF) algorithm is a regression and classification method designed to address issues related to data classification. It utilizes decision trees for making predictions. During the training phase, a specified number of decision trees are constructed and then employed for class prediction.

This involves considering the graded classes of individual trees, with the output being the class with the highest grade.

Support Vector Machine (SVM): SVM is widely applied in the medical industry for tasks such as disease detection, text content recognition, and image classification. It utilizes a fixed rule, quadratic equation, and statistical methods to segregate data into classes. For binary classification, a separating hyperplane is employed, reducing the margin space based on kernel characteristics. Although SVM does not handle large amounts of data, it is effective in identifying optimal solutions to specific problems.

Conclusion

Websites serve as versatile platforms facilitating tasks ranging from data entry to scientific applications. Processing statistical inputs allows for the generation of valuable information. Their pervasive presence is evident across diverse fields such as medicine, technology, business, education, and economics. However, this ubiquity also makes them susceptible to malicious activities by hackers who may exploit these platforms for nefarious purposes. A prime example is the deployment of apparent phishing attacks through rogue elements. Numerous research contributions have introduced innovative methodologies and approaches for identifying phishing URLs, underscoring the need for robust defences against cyber threats. The objective is to categorize phishing as a distinct attack type within the spectrum of cyber threats. The proposed system proactively alerts users to potential phishing URLs,

recommending safe alternatives even before encountering such websites, thereby mitigating the risk of falling victim to a phishing attack. The implementation involves leveraging machine learning tools, utilizing a dataset sourced from UCI for the purpose of this study. Future work will delve into refining and expanding these methodologies to enhance the overall effectivenessof cybersecurity measures.

References

1. SK Hasane Ahammad , Sunil D. Kate , Gopal D. Upaddhye , Sandeep Dwarkanath Pnade el al , ELSEVIER Jan 1 2022
2. Mohammed Hazim Alkawaz, Stephanie Joanne Steven, AsifIqbalHajamydeen, Detecting Phishing Websites Using Machine Learning, 2020 16th IEEE International Colloquium on Signal Processing & its Applications, 28-29 Feb. 2020.
3. Meenu, Sunila Godara, Phishing Detection using Machine Learning Techniques, IJEAT, ISSN: 2249 – 8958, 2, December,2019.
4. Ankit Kumar Jain and B.B. Gupta EURASIP Journal On Information Security (2016)2016:9.
5. Joby James, Sandhya L, Ciza Thomas, Detection of Phishing URLs Using Machine Learning Techniques, 2013 International Conference on Control Communication and Computing (ICCC), December 2013.
6. Suleiman Y. Yerima, Mohammed K. Alzaylaee, High Accuracy Phishing Detection Based on Convolutional Neural Networks, IEEEXplore.
7. Megha N, KR Ramesh Babu, Elizabeth Sherly, An Intelligent System for Phishing Attack Detection and Prevention, IEEE Xplore ISBN: 978-1-7281-1261-9, 2019 IEEE.

8. AmaniAlswailem, BashayrAlabdullah, Norah Alrumayh, Dr. Aram Alsedrani, Detecting Phishing Websites Using Machine Learning 978-1-7281-0108- 8/19/ 2019 IEEE.

9. https://www.hindawi.com/journals/jam/2014/425731/(randomforest)