IDEATION PHASE

WEBPHISING DETECTION

-A LITERATURE SURVEY

DOMAIN: APPLIED DATA SCIENCE

TEAM_ID: PNT2022TMID29648

BATCH No. : B9-3A5E

PAPER TITLE	APPLYING MACHINE LEARNING TECHNIQUES TO DETECT AND ANALYZE WEB PHISHING ATTACKS
AUTHOR	Cuzzocrea, Alfredo and Martinelli, Fabio and Mercaldo, Francesco
YEAR	2018
JOURNAL	the 20th International Conference on Information Integration and Web-based Applications \& Services
TECHNIQUE	URL detection, Support Vector Machine, Phishing detection system, Machine Learning, Natural Language Processing
FINDINGS/ CONCLUSIONS	• In most of the phishing website, the attackers use a malicious URL which will display to the user like an authorized URL. Different algorithms like Naive Bayes, Random Forest, K nearest neighbor are performed in detection of the URL, by using algorithm their accuracy level will be different. • This paper adopt the best classification machine learning algorithm with SVM (Support Vector Machine), this predicts the phishing or non-phishing status of the given URL and it is the best algorithm in classification (based on the features of given data) and regression (is the continuous prediction of uniform data) from which we have to improve our accuracy level. • The solution is powerful to catch phishing URLs and used as a plug-in in the browser to filter the phishing site

PAPER TITLE	WEB PHISHING DETECTION TECHNIQUES: A SURVEY ON THE STATE-OF-THE-ART, TAXONOMY AND FUTURE DIRECTIONS
AUTHOR	Vijayalakshmi, M and Mercy Shalinie, S and Yang, Ming Hour and U, Raja Meenakshi
YEAR	2020
JOURNAL	Iet Networks
TECHNIQUE	List-based detection, Heuristic rule-based detection, Learning-based detection techniques, deep learning
FINDINGS/ CONCLUIONS	• Blacklist- and whitelist-based approaches alone cannot effectively detect the emanating phishing attacks as these lists are having issues with update mechanism. Visual similarity techniques also have computational and space complexity issues • Lightweight web phishing detection techniques using hybrid approaches would be the better choice for handling current phishing scams. Applying deep learning in web phishing detection is also considered. • Webpage content or similarity-based phishing detection approach using deep learning methodologies is recognised as an open call for future work in web phishing detection.

PAPER TITLE	PHISHZOO: AN AUTOMATED WEB PHISHING DETECTION APPROACH BASED ON PROFILING AND FUZZY MATCHING
AUTHOR	Afroz, Sadia and Greenstadt, Rachel
YEAR	2009
JOURNAL	5th IEEE Int. Conf. Semantic Comput.(ICSC)
TECHNIQUE	fuzzy hashing techniques
FINDINGS/ CONCLUSION	• a phishing detection approach—PhishZoo—that uses profiles of trusted websites' appearances built with fuzzy hashing techniques to detect phishing. Evaluate this approach on over 600 phishing sites imitating 20 real sites and show that it provides similar accuracy to blacklisting approaches, with the advantage that it can classify new attacks and targeted attacks against smaller sites (such as corporate intranets). • It has a beneficial impact on the phishing "arms race"by reducing the effectiveness of sites that look too much like the real sites and thus giving users a chance to detect sites that "look phishy." • This method is also most accurate against sites that look most like the real sites (those hardest for end users to detect)

PAPER TITLE	URL PHISHING DETECTION USING MACHINE LEARNING TECHNIQUES BASED ON URLS LEXICAL ANALYSIS
AUTHOR	Mohammed Abutaha, Mohammad Ababneh, Khaled Mahmoud ,Sherenaz Al-Haj Badda
YEAR	October 2021
JOURNAL	2021 12th International Conference on Information and Communication Systems (ICICS)
TECHNIQUE	Neural Network, Random Forest, SVM, GBC, Machine Learning, URL Analysis.
FINDINGS/ CONCLUSIONS	• The experiments were carried out on a dataset that originally contained 1056937 labeled URLs (phishing and legitimate). • This dataset was processed to generate 22 different features that were reduced further to a smaller set using different features reduction techniques. • Random Forest, Gradient Boosting, Neural network and Support Vector Machine (SVM) classifiers were all evaluated, and results show the superiority of SVMs, which achieved the highest accuracy in detecting the analyzed URLs with a rate of 99.89%. • This approach can be incorporated within add-on/middleware features in Internet browsers for alerting online users whenever they try to access a phishing website using only its URL.

PAPER TITLE	INTELLIGENT WEB-PHISHING DETECTION AND PROTECTION SCHEME USING INTEGRATED FEATURES OF IMAGES, FRAMES AND TEXT
AUTHOR	Adebowale, Moruf A and Lwin, Khin T and Sanchez, Erika and Hossain, M Alamgir
YEAR	2019
JOURNAL	Expert Systems with Applications
TECHNIQUE	Adaptive Neuro-Fuzzy Inference System (ANFIS)
FINDINGS/ CONCLUSION	• An Adaptive Neuro-Fuzzy Inference System (ANFIS) based robust scheme using the integrated features of the text, images and frames for web-phishing detection and protection. • An efficient ANFIS algorithm was developed, tested and verified for phishing website detection and protection based on the schemes proposed in Aburrous et al. (2010) and Barraclough et al. (2015). • A set of experiments was performed using 13,000 available datasets. The approach showed an accuracy of 98.3%, which so far, is the best-integrated solutions for web-phishing detection and protection.

PAPER TITLE	MACHINE LEARNING BASED PHISHING DETECTION FROM URLS
AUTHOR	Ozgur Koray Sahingoz , Ebubekir Buber, Onder Demir , Banu Diri.
YEAR	September 2018
JOURNAL	Experts Systems with Application
TECHNIQUE	Cyber security Phishing attack Machine learning Classification algorithms Cyber attack detection
FINDINGS	• A real-time anti-phishing system, which uses seven different classification algorithms and natural language processing (NLP) based features, is proposed. • The system has the following distinguishing properties from other studies in the literature: language independence, use of a huge size of phishing and legitimate data, real-time execution, detection of new websites, independence from third-party services and use of feature-rich classifiers. • For measuring the performance of the system, a new dataset is constructed, and the experimental results are tested on it. • According to the experimental and comparative results from the implemented classification algorithms, Random Forest algorithm with only NLP based features gives the best performance with the 97.98% accuracy rate for detection of phishing URLs%

PAPER TITLE	WEB PHISHING DETECTION USING A DEEP LEARNING FRAMEWORK
AUTHOR	Yi, Ping and Guan, Yuxiang and Zou, Futai and Yao, Yao and Wang, Wei and Zhu, Ting
YEAR	2018
JOURNAL	Hindawi
TECHNIQUE	deep learning framework
FINDINGS	• Applying a deep learning framework to detect phishing websites. • This paper first designs two types of features for web phishing: original features and interaction features. A detection model based on Deep Belief Networks (DBN) is then presented. • The test using real IP flows from ISP (Internet Service Provider) shows that the detecting model based on DBN can achieve an approximately 90% true positive rate and 0.6% false positive rate.

PAPER TITLE	PHISHING WEBSITE DETECTION: AN IMPROVED ACCURACY THROUGH FEATURE SELECTION AND ENSEMBLE LEARNING
AUTHOR	Alyssa Anne Ubing, Syukrina Kamilia Binti Jasmi, Azween Abdullah, NZ Jhanjhi , Mahadevan Subramaniam
YEAR	January 2019
JOURNAL	International Journal of Advanced Computer Science and Applications ·
TECHNIQUE	Random Forest, Logistic regression
FINDINGS	• A feature selection algorithm is employed and integrated with an ensemble learning methodology, which is based on majority voting, and compared with different classification models including Random Forest, Logistic Regression, Prediction model etc. • It demonstrates that current phishing detection technologies have an accuracy rate between 70% and 92.52%. • The experimental results prove that the accuracy rate yield up to 95%,

PAPER TITLE	DETECTION OF PHISHING WEBSITES USING MACHINE LEARNING APPROACH
AUTHOR	Kaksha, Sameena Naaz
YEAR	February 2019
JOURNAL	International Conference on Sustainable Computing in Science, Technology & Management (SUSCOM-2019)
TECHNIQUE	Random Forest, Decision Tree, Linear Model, Neural Model
FINDINGS	• Data mining tools can be applied in this regard as the technique is very easy and can mine millions of information within seconds and deliver accurate results. • With the help of machine learning algorithms like, Random Forest, Decision Tree, Neural network and Linear model we can classify data into phishing, suspicious and legitimate. • This can be done based on unique features of phishing websites and user does not need to check individual websites. • Rather we can identify and predict phishing, suspicious and legitimate websites by extracting some unique features.

PAPER TITLE	ANOMALY BASED WEB PHISHING PAGE DETECTION
AUTHOR	Pan, Ying and Ding, Xuhua
YEAR	December 2006
JOURNAL	2006 22nd Annual Computer Security Applications Conference (ACSAC'06)
TECHNIQUE	DOM object anomalies detection
FINDINGS/ CONCLUSIONS	• To examine the anomalies in web pages, in particular, the discrepancy between a web site's identity and its structural features and HTTP transactions • It demands neither user expertise nor prior knowledge of the website. • The evasion of this phishing detection entails high cost to the adversary • Identity extraction is critical to a successful detection. However, it remains as an open problem how to extract the identity from web pages with an overwhelming success probability • A possible improvement is to explore information from logo images in web pages

PAPER TITLE	WEB PHISHING DETECTION BASED ON PAGE SPATIAL LAYOUT SIMILARITY
AUTHOR	Zhang, Weifeng and Lu, Hua and Xu, Baowen and Yang, Hongji
YEAR	2013
JOURNAL	Informatica
TECHNIQUE	Spatial layout features, DOM tree based spatial layout features, Image segmentation based spatial layout features
FINDINGS	• a spatial layout similarity based approach for phishing web detection. This novel approach takes into account important spatial layouts of web pages. • Invented two meaningful methods that extract the spatial layout features from web pages. After obtaining such spatial layout features, they define a similarity function to quantize how visually similar two web pages are. • Such a similarity measurement indicates how a suspicious page is a phishing one in relation to a legitimate web page. • In order to speed up searching the legitimate feature library, we further design an R-tree index for all spatial layout features in the library and develop search algorithms accordingly. • Evaluated the proposed approach through a series of experiments. The results demonstrate the effectiveness and efficiency

PAPER TITLE	DETECTION OF PHISHING WEBSITES BY USING MACHINE LEARNING-BASED URL ANALYSIS
AUTHOR	Mehmet Korkmaz
YEAR	July 1-3, 2020
JOURNAL	IEEE
TECHNIQUE	cybersecurity, phishing, machine learning, website classification
FINDINGS/ CONCLUSIONS	• A machine learning-based phishing detection system by using eight different algorithms to analyze the URLs, and three different datasets to compare the results with other works