WEB PHISHING DETECTION

TEAM ID	**: PNT2022TMID50791**
TEAM LEADER	**: ESTHER J (953219106009)**
TEAM MEMBERS	**: GAYATHRI PRIYADHARSHINI K (953219106010)**
	: MERLIN C(953219106019)
	: SAKTHI ESWARI K (953219106028)
	: UMA S (953219106041)
DEPARTMENT	**: ELECTRONICS & COMMUNICATION ENGINEERING**
COLLEGE NAME	**: UNIVERSITY VOC COLLEGE OF ENGINEERING THOOTHUKUDI.**

LITERATURE SURVEY

1. WEB ADDRESS BASED EVALUATION

1.1. LIST BASED DETECTION TECHNIQUES

A database of URL called list ismaintained. It generally holds URLs, internet protocol (IP) addresses, andkeywords. Some researchers maintain a whitelist, which is a collection of legitimate URLs. Most of the researchers suggest maintaining a blacklist, which is a collection of malicious URLs.list-based detection method acts as a filteringmechanism to sweep away suspicious webpages before entering into the detection process

.

SI. NO	TITLE OF PAPER	YEAR OF PUBLICATION	AUTHOR NAME	DESCRIPTION
1	Anti-phishing based on automated individual white-list	2008	Cao Y. Han W. Le Y.	He proposed an automated individual whitelist (AIWL)-based approach that maintains a local listof user's familiar login user interface (LUI) of websites to alert the user whenever he tries to access an unfamiliar website with LUI. AIWL uses a naïve Bayesian classifier to maintain the list by adding the unknown website. However, this approach cannot stand up against the local machine trojan horse and viruses.
2	A novel approach to protect against phishing attacks at client side using auto-updated white-list	2016	Jain A.K. Gupta B.B	It combined the whitelist approach with heuristics and ML to propose the auto-updated whitelist. Blacklists and whitelists are used as a filtering module in many web phishing detection approaches to reduce the processing time wasted on pre-processing, feature extraction, and so on.

1.2. HEURISTICRULE BASED DETECTION TECHNIQUE

Heuristic rule-based techniques can identify thezero-dayattacks. Therefore, it has a high-detection rate than list-based phishing detectionschemes.The performance and accuracy of the technique wholly depend onthe heuristics applied.

SI. NO	TITLE OF PAPER	YEAR OF PUBLICATION	AUTHOR NAME	DESCRIPTION
1	Machine learning based phishing detection from URLs	2019	Sahingoz O.K. Buber E. Demir O. et al	Applies heuristics to extract natural language processing (NLP) featuresfrom the URL to detect the URL-based web phishing attacks. The heuristics are derived based onparameters such as raw word count, short word length, Alexa ranking, similar brand name count,etc.
2	‘A stacking model using URL and HTML features for phishing webpage detection	2019	Li Y. Yang Z. Chen X. et al	Applies some heuristics on the URL to verify abnormalities such as suspicious symbols (e.g. @, _), https, URL length information, number of dots in a domainname, sensitive vocabulary, and top-level domain.
3.	Intelligent phishing URL detection using association rule mining	2016	Jeeva S.C. Rajsingh E.B.	Computes 14heuristics: length of the host URL, number of slashes, dots in the host name, number of terms in the host name, special characters, IP address, unicode in URL, transport layer security, subdomain, certain keyword, top-level domain, number of dots in the path of the URL, hyphen in the host name and URL length. The extracted features are then fed into associative rule mining algorithms.

4	A phish detector using light weight search features	2016	Varshney G.Misra M..Atrey P.K	Proposed a lightweight phish detector, which extracts the domain name of the URL and title of the webpagewhenever a user accessing a website. The extracted URL domain name and the title page aresearched using a search engine to determine the legitimacy

1.3. LEARNING BASED DETECTION TECHNIQUE

Learning algorithms such as ML and deep learning are used to detect the attacks based on the features extracted from the URL. In learning-based web phishing detection, the statistical features and NLP features of the URLs are extracted and fed into ML algorithms such as supportvector machine (SVM), decision tree, naïve Bayes algorithm, random forest etc. The classifier creates a model based on the inference extracted from the Training samples. The suspicious URL is evaluated based on the model built by the classifier.

SI. NO	TITLE OF PAPER	YEAR OF PUBLICATION	AUTHOR NAME	DESCRIPTION
1	Machine learning based phishing detection from URLs	2019	Sahingoz O.K. Buber E. Demir O. et al.	Practices seven different ML algorithms such as naive Bayes, random forest ,k-nearest Neighbour(KNN),Adaboost,kstar, ,sequentialminimal optimization,anddecision tree on the extracted features from the URL and analysed the best performance among them.
2	‘A stacking model using URL and HTML features for phishing webpage detection	2019	Li Y. Yang Z. Chen X. et al.:	Proposed a deep learning approach to extract the features naturally from the URLs and to detect the web phishing attack. Convolutional neural network(CNN) is used to extract the correlation features and long short term memory (LSTM) network isused to learn sequentialdependency.

3	Phishing website detection based on multidimensional features driven by deep learning	2019	Yang P. Zhao G. Zeng P: ‘.	Proposed aweb phishing detection approach using a neural network. In this work, feature validity value(FVV) is introduced to examine the effect of optimal features. By using the FVV index, the optimal feature selection algorithm is designed to choose the optimal features and is used to mitigate the over fitting problem of neural networks.

ML algorithms can detect zero-day attacks and have a shorter detection time. Howeverthistechnique is feature sensitive and the performance varies based on the characteristics of the ML algorithm applied

2. WEBPAGE CONTENT\ SIMILARITY BASED EVALUATION

2.1. HEURISTIC RULE BASED WEBPAGE SIMILARITYEVALUATION

In heuristic-based webpage similarity calculation, keywords and features are extracted from thesuspicious webpage and verified against the targeted webpage using search methods to enable asecured environment against phishing scams.

SI. NO	TITLE OF PAPER	YEAR OF PUBLICATION	AUTHOR NAME	DESCRIPTION
1	PhishWHO: phishing webpage detection via identity keywords extraction and target domain name finder	2016	Tan C.L. Chiew K.L. Wong K.	Proposes a phishing webpage detection approach four modules- identity keywords extraction, search engine lookup, target domain name finder, and three-tier identity matching.The target domain name and actual domain name are passed as inputs to the three-tier identity matching systemto analyse the status of the query webpage.

2	Phishing-alarm: robust and efficient phishing detection via page component similarity	2017	Mao J. Tian W. Li P. et al.:	Proposed a phishing alarm by extracting the CSS features from the underlyingarchitecture of the web page. Page similarity calculations are applied to the extracted features toclassify the web pages
3	Off-the-hook: an efficient and usable client-side phishing prevention application	2017	Marchal S. Armano G. Gröndahl T. et al.	Designed a client-side phishing detection tool that offers better privacy, real-time protection, effective warnings, and resilience to dynamic phish. This approach uses a phish detector and target identifier mechanisms to detect the Phishing webpages.

2.2. ML-BASED WEBPAGE SIMILARITY EVALUATION

In this technique, HTML, extensible mark-up language (XML), JavaScript (JS), and CSS featuresare extracted from the source code of the webpage and are fed into ML algorithms for furtherclassification.

SI. NO	TITLE OF PAPER	YEAR OF PUBLICATION	AUTHOR NAME	DESCRIPTION
1	Cantina+ a feature-rich machine learning framework for detecting phishing web sites	2011	Xiang G. Hong J. Rose C.P. et al.: ‘	Proposed a content-based approach to detect web phishing by extracting URLfeatures, HTML-based features, and web-based features.The proposed approach is evaluated with two methods that are randomised evaluation and time-based evaluation using the Bayesian network.

2	Detecting phishing websites via aggregation analysis of page layouts	2018	Mao J. Bian J. Tian W. et al.:	Proposed a learning-based layout similarity detection using ML algorithms. SVM and decision trees are used to classify the similarity of the webpages.
3	A new hybrid ensemble feature selection framework for machine learning-based phishing detection system	2019	Chiew K.L. Tan C.L. Wong K. et al.	Proposed a new feature selection framework for ML-based phishing detectionsystem. A novel cumulative distribution function gradient algorithm is designed as an automaticfeature cut-off rank identifier to produce the compact set of primary features and then dataperturbation, and function perturbation techniques are applied on these primary features to derive the hybrid ensemble features.
4	A machine learning based approach for phishing detection using hyperlinks information	2019	Jain A.K. Gupta B.B.	Proposed a novel web phishing detection approach by extracting hyperlinks of the web pages. The proposedapproach has extracted 12 specific hyperlink feature. The extracted features are then fed into ML algorithms such as naïve Bayes, random forest, SVM, Adaboost, neural network, C4.5, and logisticregression. The performance of all the ML algorithms was measured and reported.

3.HYBRID APPROACHES

Hybrid web phishing detection techniques were proposed by combining the existing web phishing detection schemes.

SI. NO	TITLE OF PAPER	YEAR OF PUBLICATION	AUTHOR NAME	DESCRIPTION
1	A comprehensive and efficacious architecture for detecting Phishing webpages	2014	Gowtham R. Krishnamurthi I.:	Proposed a web phishing detection approach using a preapproved site identifier, login form finder, and ML algorithms. The websites which are resulted as suspicious from the modules are further processed by the SVM ML algorithm.
2	A stacking model using URL and HTML features for phishing webpage detection	2019	Li Y. Yang Z. Chen X. et al.	Combined URL features, HTML source code features, and HTML string embedding to detect theweb phishing scam. A stacking model of gradient boost decision tree, Xtreme Gradient Boost(XGBOOST), and LightGBM is used to improve the performance of the system.
3	Phishing website detection based on multidimensional features driven by deep learning	2019	Yang P. Zhao G. Zeng P	Presented a hybrid approach to attain multi-dimensional features to increase the detection rate and to reduce the detection time. URL evaluation, web page similarity approach, and contentbased approach are combined in that work. Both ML (i.e XGBOOST) and deep learning (i.e.CNN-LSTM) algorithms are applied to classify the attack.

4	Two level filtering mechanism to detect phishing sites using lightweight visual similarity approach	2019	Rao R.S. Pais A.R.	Proposed a two level filtering mechanism to detect the web phishing attack. At the first level, a lightweight visual similarity-based blacklist is applied to detect near-duplicate phishing sites. At the secondlevel, heuristic filtering is performed on the bypassed phishing sites from the blacklists.
5	An approach for phishing validation and detection	2017	Li J.H. Wang S.D.:	Proposed a PhishBox approach forphish validation and detection. This approach has a two-stage model. In the first stage, theensemble model is designed to evaluate the phish data, and active learning is applied to reduce the cost of manual labelling. In the second stage, the validated phishing data is used to train the detection model.