TAGORE ENGINEERING COLLEGE

LITERATURE SURVEY

APPLIED DATA SCIENCE BASED
WEB PHISHING DETECTION USING MACHINE LEARNING

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE AND ENGINEERING

RAKSHANA .G (412719104025)

PREETHI.D (412719104022)

MONISHA.N (412719104018)

JAYASHREE.K (412719104014)

ABSTRACT

Phishing URL is a widely used and common technique for cyber security attacks. Phishing is a cybercrime that tries to trick the targeted users into exposing their private and sensitive information to the attacker. The motive of the attacker is to gain access to personal information such as usernames, login credentials, passwords, financial account details, social networking data, and personal addresses. These private credentials are then often used for malicious activities such as identity theft, notoriety, financial gain, reputation damage, and many more illegal activities. This paper aims to provide a comprehensive and comparative study of various existing free service systems and research based systems used for phishing website detection. The systems in this survey range from different detection techniques and tools used by many researchers. The approach included in these researched papers ranges from Blacklist and Heuristic features to visual and content-based features. The studies presented here use advanced machine learning and deep learning algorithms to achieve better precision and higher accuracy while categorizing websites as phishing or benign. This article would provide a better understanding of the current trends and existing systems in the phishing detection domain.

TABLE

Analysis of heuristics and machine learning – based techniques

Authors	Novelty	Book / Journal	Dataset / accuracy	Drawbacks
Tuan et al	The novelty is in the identification of sex minimal features claimed to provide a high accuracy	Phishing detection from URL by using Neural Network (2018)	Dataset : Phishtank database, 11660 phishing sites; Acc = 97.16%	Complex as is employs various heuristics, making its deployment difficult as a real – time client side tool; heavily dependent on third parties for its operations.
Mohammad et al.	The accuracy of various data mining techniques has been studied for phishing detection, and CBA has been identified as the best performing one.	Intelligent phishing website detection using random forest classifier (2017)	Lowest error rate of 4.5% was obtained using CBA.	A practical implantation as an anti phishing tool and / or its effectiveness and cost benefit analysis is missing
Kausar et al.	Experimental validation and identification of the best combination of phases from the two approaches (62,63) taken to improve detection	Phishing detection using hybrid Ensemble Model JERT(2019)	Dataset: 89 phishing websites and 71 legitimate websites Acc = 87.5%	Adds little in terms of proposing a new scheme for phishing detection. Work is more towards study of combining the phases from Tuan (62) and Gu (63)

	accuracy			approach.
Barraclough et al	Unlike others (57,61.65) the novelty is the addition of neuro fuzzy approach and the inclusion of user behavioural profile of website interaction as input for phishing detection	Intelligent Phishing Website detection using Random Forest classifier (IEEE explore) 2017	Acc= 98.5% claims that it provides better results than Netcraft (72) CANTINA (73) and Spool Guard (74)	Complex and highly dependent of inputs; creating user profile of interaction with a set of websites is tedious, time consuming and resource intensive.
Birhanu	Concept of combining static and dynamic analysis approaches with machine learning. unlike others (64,65) used evolutionary search and optimization algorithm for better accuracy	A survey and classification web phishing detection schemes (2016)	Exists as a proposal	Not enough details of testing and validation are available to comment on the effectiveness of the proposed work and its benefits.