Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 16, Issue. 1: 2025
ISSN : **1906-9685**

Journal of Nonlinear Analysis and Optimization : Theory & Applications
ISSN : 1906-9685
Editors-in-Chief :
Sompong Dhompongsa
Somyot Plubtieng
[illegible]

A STUDY ON ADVERSARIAL SAMPLE RESISTANCE AND DEFENSE MECHANISM FOR MULTIMODAL LEARNING

B. AMARNATH REDDY[1], **P.CHANDRASEKHAR**[2]

[1]Assistant Professor, Dept. of MCA, QIS College of Engineering and Technology, Ongole, Andhra Pradesh.

[2]PG Scholar, Dept. of MCA, QIS College of Engineering and Technology, Ongole, Andhra Pradesh.

ABSTRACT— Recent advancements in Artificial Intelligence (AI) have greatly impacted cybersecurity, particularly in detecting phishing websites. Traditional methods struggle to address evolving vulnerabilities, but research shows that Machine Learning (ML), Ensemble Learning (EL), and Deep Learning (DL) are effective in developing defenses. However, these methods face challenges with adversarial examples (AEs). The multimodal model (MM) is a promising solution, yet there is a significant lack of research using multimodal techniques specifically for phishing website detection (PWD) against adversarial websites. To tackle this challenge, this paper assesses 15 learning-based models, particularly multimodal ones, for phishing and adversarial detection, aiming to enhance their defense capabilities. Due to the scarcity of adversarial websites, training and testing models are limited. Therefore, this study proposes an innovative attack framework, AWG- Adversarial Website Generation that employs Generative Adversarial Networks (GAN) and transfer-based black box attacks to create AEs. This framework closely mirrors real-world attack scenarios, ensuring high effectiveness and realism. Finally, we present defense strategies with straightforward implementation and high effectiveness to enhance the resistance of models. The models underwent training and testing on a dataset collected from reputable sources such as OpenPhish, PhishTank, Phishing

Database, and Alexa. This approach was chosen to ensure the dataset's diversity and relevance to reflect real-world conditions.

Index Terms – Multimodal, deep learning, machine learning, ensemble learning, phishing website detection

I. INTRODUCTION

Phishing is a typical kind of cybercrime by employs both social engineering and technical subterfuge to steal the consumer's personal identity data and financial account credentials. Phishing attackers typically create fraudulent websites with domains, URLs, and appearances that closely resemble legitimate websites to deceive users and obtain their sensitive data. According to the latest report of the Anti Phishing Working Group (APWG) in Q4 2023, APWG has logged more than 5 million phishing attacks. Since the beginning of 2019, the number of phishing attacks has grown by more than 150% per year. The rapid increase in the number of phishing websites is facilitated by the emergence of phishing toolkits, which enable attackers to quickly create phishing websites while ensuring necessary functionalities and increasing website quality. This poses challenges for differentiation based on human vision. In the realm of cybersecurity, various techniques have emerged to detect phishing attacks. Typically, the blocklist-based method is the easiest to implement and has gained widespread adoption. However, these blocklists are updated periodically, which may result in delayed detection of unknown malicious phishing web sites. In light of this limitation, ML techniques have been introduced to detect phishing websites effectively. Based on previous studies there are various approaches to applying AI in phishing website detection, these include designing feature extraction frameworks suitable for web applications for ML models so that the models can operate stably, efficiently, and resource-light building DL model architectures to leverage automatic feature extraction or employing EL to leverage the individual strengths of different model groups. Among these numerous studies, the majority focus on URL-based approaches for detecting phishing websites while others rely on domain-based approaches. Some studies also concentrate on the external appearance of websites or rely on third-party sources to gather information. All studies aim to provide solutions that meet the demands of real world deployment, are highly efficient and resource-saving, require minimal effort for preparation and deployment, and effectively utilize comprehensive datasets.

However, despite their high reliability, it is widely acknowledged that ML/DL methods are susceptible to AEs. Adding AEs with perturbations can cause ML-based detectors to make incorrect decisions. For instance, the ML/DL methods focusing on early URL appearance for phishing website detection continue to be extensively researched and applied due to their effectiveness. However, despite these shining points, URL-based models are susceptible to evasion techniques. Sabir et al. research has highlighted the weakness and vulnerability of URL-focused ML models against AEs. In the studies, SpacePhish and Multi-SpacePhish by the authors Apruzzese, it was demonstrated that AEs could heavily impact state-of-the-art ML-PWD models. This was shown through the implementation of 12 different adversarial attack types. This highlights the challenge faced by AI research in identifying phishing websites. 137806 To address adversarial attacks, MMs present a promising solution. They focus on various entities to extract features. For example, the study by Revue which concentrates on URLs along with JavaScript source and website content, or the Shark-eyes model which focuses on HTML tag groups and domains for feature extraction. Yuan et al. have proposed a multimodal approach that combines image feature extraction and structural feature extraction representation to detect malicious URLs. These studies have achieved positive results and show promising potential for robustness against phishing websites. However, there is a lack of research on the resilience of MMs against adversarial attacks in the context of Pwd. Forinstance, although SpacePhish and Multi-SpacePhish represent the first statistically validated assessments of state-of-the-art ML-PWD against 12 evasion attacks, they still lack validated assessments involving modern DL, EL, and MM models. Furthermore, there is a limited number of publicly available AEs, especially in the context of phishing websites.

II. LITERATURE SURVEY

A. On the top threats to cyber systems

The technological innovation of cyber systems and increase dependence of individuals, societies and nations on them has brought new, real and everchanging threat landscapes. In fact, the threats evolving faster than they can be assessed. The technological innovation that brought ease and efficiency to our lives, has been met by similar innovation to take advantage of cyber systems for other gains. More threat actors are noted to be sponsored by nation-states and the skills and capabilities of organizations to defend against these attacks are lagging. This warrants an increase in automation of threat analysis and response as well as increased adoption of security measures by at-risk organizations. Thus, to properly prepare defenses and mitigations to the threats introduced by cyber, it is necessary to understand these threats. Accordingly, this paper provides an overview of top cyber security threats in together with current and emerging trends. The analyses include general trends in the complexity of attacks, actors, and the maturity of skills and capabilities of organizations to defend against attacks. Top threats are discussed with regard to instances of attacks and strategies for mitigation within the kill chain. A brief discussion of threat agents and attack vectors adds context to the threats.

B. Safer: Social capital-based friend recommendation to defend against phishing attacks

The tremendous growth of social media has been accom-panied by highly advanced online social network (OSN)technologies. Such advanced technologies have been heav-ily utilized by perpetrators as convenient tools for deceiv-ing people in online worlds. Social capital has been dis-cussed as a powerful mechanism to leverage interpersonalrelationships in social networks in order for an individ-ual to achieve his/her goal. The beauty of social capitalis the ability to materialize non-monetary, less costly, andnon-economic resources into tools to solve social problems.In this paper, we aim to leveragesocial capital(SC) tominimize online users' vulnerabilities to online deception.

In particular, we propose a Social cApital-based FriEndRecommendation scheme, called SAFER, that can protectOSN users from phishing attacks. We quantify three dimen-sions of social capital, namely, structural, cognitive, and re-lational, based on user features obtained from real datasetsand model a user's friending behavior based on their social capital. In addition, to model a user's behavior upon being attacked by a phishing attacker, we developed the so-called SER-SEIR (Susceptible, Exposed, Recovered-Susceptible ,Exposed, Infected, and Recovered) model as a variant of theSEIR model. Via extensive simulation experiments based on two real datasets considering bot-based and human-based attackers performing phishing attacks, we demonstrate the per-formance of four SC-based friend recommendation schemes with three non-SC-based comparable counterparts in terms of the ratio of detecting attackers and the fraction of users in the states of S, E, I, and R. Based on the performance comparison, we analyze the overall trends of their performance in terms of the extent of resistance against phishing attacks bybot or human attackers.

C. *Phishing or not phishing? A survey on the detection of phishing websites*

Phishing is a security threat with serious effects on individuals as well as on the targeted brands. Although this threat has been around for quite a long time, it is still very active and successful. In fact, the tactics used by attackers have been evolving continuously in the years to make the attacks more convincing and effective. In this context, phishing detection is of primary importance. The literature offers many diverse solutions that cope with this issue and in particular with the detection of phishing websites. This paper provides a broad and comprehensive review of the state of the art in this field by discussing the main challenges and findings.

III. PROPOSED SYSTEM

The overview of our proposed system is shown in the below figure.

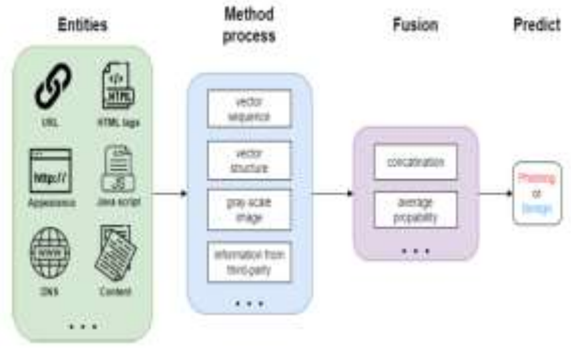

Fig. 1: System Overview

Implementation Modules

Load Dataset

- ✓ For this project we'll need bunch of legitimate and phishing url's, each categorized by (0) and (1).
- ✓ It contains 450k domain url's out of which 345k are legitimate and 104k are malicious. The Imbalanced dataset is oversampled using the SMOTE technique, which increases the total number of samples to around 600k

Preprocess

- ✓ In this module, we preprocess the url data in which we follow the steps like, data cleaning, data transform and data normalization using python library numpy

Feature Extraction

- ✓ The dataset till now consist of only legit and malicious urls, in this stage we extract some useful features from these urls and further improve our dataset to make it more suitable for training ML models.
- ✓ The below mentioned category of features are extracted from the URL data:
- ✓ Length based Features (5 features extracted)
- ✓ Count based Features (11 features extracted)
- ✓ Binary Features (2 features extracted)
- ✓ All together 18 features are extracted from each url of the dataset.

IV. RESULTS

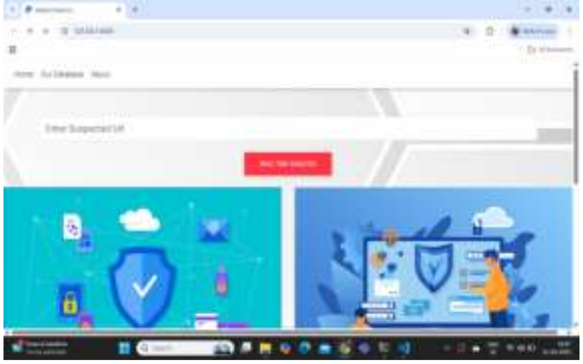

Fig.2: Home Page

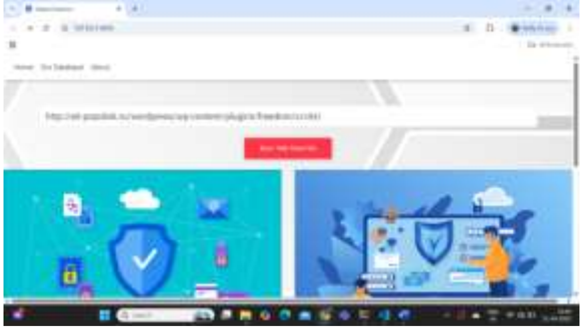

Fig.3: Enter URL

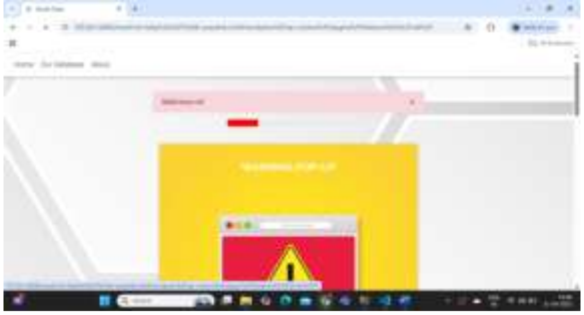

Fig.4: Malicious URL

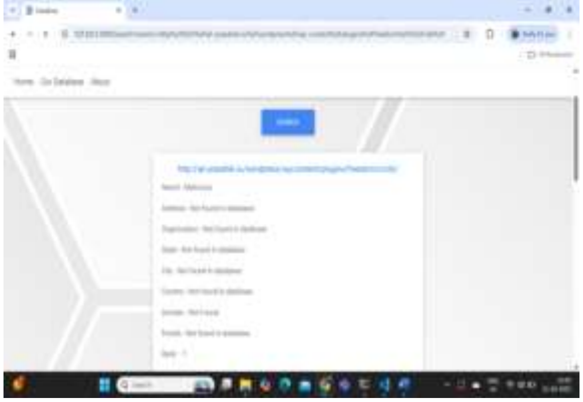

Fig.5: URL Details

Fig.6: Safe URL

V. CONCLUSION

In this study, we introduce an AWG model for generating adversarial samples by leveraging the concept of transfer attacks against phishing website detection models. These samples not only replicate the original structure but also exhibit high evasion capabilities. Specifically, the generated AEs, in conjunction with our attack phase, can effectively target most state-of-the-art PWD models, including ML, DL, EL and MM. This poses a significant challenge for scam detection systems based on blocklist datasets or AI models. Furthermore, AEs generated from our attack phase can serve as valuable resources for training and helping AI models adapt to new cyber threats. Besides, Besides, MMs demonstrate superior resistance to adversarial attacks, with the highest detection rate compared to others. Lastly, our defense strategy is acknowledged as a straightforward, easy-to-implement, highly effective approach. Nonetheless, one adversarial website bypassing the detector in phishing attacks can have significant consequences. Therefore, in the future, we aim to refine our attack and defense methods further to bolster the robustness of AI models. Ultimately, our study seeks to pioneer advancements on building more robust learning-based PWD in the realm of adversarial phishing detection research.

REFERENCES

[1] H. Kettani and P. Wainwright, ‘‘On the top threats to cyber systems,’’ in Proc. IEEE 2nd Int. Conf. Inf. Comput. Technol. (ICICT), Mar. 2019, pp. 175–179.

[2] Anti-Phishing Working Group. Phishing Attack Trends Report—4Q 2023. Accessed: Mar. 21, 2024. [Online]. Available: https://docs.apwg.org/reports/apwg_trends_report_q4_2023.pdf

[3] M. A. Shaik, G. Rakshitha, K. Saipriya, T. Thrisha, M. Varshini and J. G. Sai, "Machine Learning for Detecting the Phishing Threats," 2025 6th International Conference on Mobile Computing and Sustainable Informatics

(ICMCSI), Goathgaun, Nepal, 2025, pp. 1221-1226, doi: 10.1109/ICMCSI64620.2025.10883227.

[4] R. Zieni, L. Massari, and M. C. Calzarossa, ‘‘Phishing or not phishing? A survey on the detection of phishing websites,’’ IEEE Access, vol. 11, pp. 18499–18519, 2023.

[5] Z.Alkhalil, C.Hewage, L.Nawaf, and I.Khan, ‘‘Phishing attacks: Arecent comprehensive study and a new anatomy,’’ Frontiers Comput. Sci., vol. 3, Mar. 2021, Art. no. 563060.

[6] M. A. Shaik, V. S. Rani, A. Fatima, M. Parveen, J. Juwairiyyah and N. Fatima, "Secure Data Exchange in Cloud Computing: Enhancing Confidentiality, Integrity, and Availability Through Data Partitioning and Encryption," 2024 International Conference on Smart Technologies for Sustainable Development Goals (ICSTSDG), Chennai - 600077, Tamil Nadu, India, 2024, pp. 1-6, doi: 10.1109/ICSTSDG61998.2024.11026651

[7] W. Li, S. Manickam, S. U. A. Laghari, and Y.-W. Chong, ‘‘Uncovering the cloak: A systematic review of techniques used to conceal phishing websites,’’ IEEE Access, vol. 11, pp. 71925–71939, 2023.

[8] M. A. Shaik, A. Fatima, M. Parveen, A. Soumya Rani, A. Mohammad and A. Rahim, "Dual-Model Approach for Lung Disease Classification Using Convolutional Neural Networks and Support Vector Machines," 2024 International Conference on Integrated Intelligence and Communication Systems (ICIICS), Kalaburagi, India, 2024, pp. 1-6, doi: 10.1109/ICIICS63763.2024.10860090

[9] P. Prakash, M. Kumar, R. R. Kompella, and M. Gupta, ‘‘PhishNet: Predictive blacklisting to detect phishing attacks,’’ in Proc. IEEE INFOCOM, Mar. 2010, pp. 1–5.

AUTHORS Profile

Mr. B. Amarnath Reddy is an Assistant Professor in the Department of Master of Computer Applications at QIS College of Engineering and Technology, Ongole, Andhra Pradesh. He earned his M.Tech from Vellore Institute of Technology(VIT), Vellore. His research interests include Machine Learning,Programming Languages. He is committed to advancing research and fostering innovation while mentoring students to excel in both academic and professional pursuits.

Mr. P. Chandrasekhar has revived has received her B.sc (computers)And Degree From ANU 2022 Pursuing MCA QIS College of Engineering And Technology Affiliated to JNTUK 2023-2025.