Heuristic Phishing Detection and URL Checking MethodologyBased on Scraping and Web Crawling

Harshal Kanifnath More[1] Chaitanya Gangadhar Umbarkar[2] Suyash Sunil Zalte[3] Radheshyam Rahulbhai Suryavanshi[4] Prof. Rahul M. Raut[5]

[1,2,3,4]BE Student [5]Project Guide

[1,2,3,4,5]Department of Information Technology

[1,2,3,4,5]Sandip Institute of Technology & Research Centre, Nashik, India

Abstract—
The Project with the improvement of techniques used by the attackers, the detection and prevention of threats such as phishing and malware can represent a problem and computational challenge. In the past, various research studies have tried to identify and classify the factors contributing towards the detection of phishing websites. Recent research has found that phishing and malicious code infection are the main threats triggered by social engineering. In this work, the attack vectors that cause these threats are analyzed, proposing a method of checking specific strings in URLs and e-mail messages, which can be used in conjunction with proxies and Anti-Spam filters. The method was implemented in an experimental scenario and is capable of detecting the presence of the main elements that have direct contact with the user, such as:

form fields, redirection of links and downloadable files. Furthermore, the proposed method was able to detect phishing URLs with accuracy valuesbetween 73.3% and 97.66% with anaverage time of 30 seconds**.**

Keywords— Uniform resource locators, Whitelists, Phishing, Unsolicited e-mail, Web pages, Tools, Malware

I. Introduction

URL phishing attacks can use various means to trick a user into clicking on the malicious link. For example, a phishing email may claim to be from a legitimate company asking the user to reset their password due to a potential security incident. Alternatively, the malicious email that the user needs to verify their identity for some reason by clicking on the malicious link. Once the link has been clicked, the user is directed to the malicious phishing page. This page may be designed to harvest a user's credentials or other sensitive information under the guise of updating a password or verifying a user's identity. Alternatively, the site may serve a "software update" for the user to download and execute that is actually malware. Phishing is a type of social engineering attack often used to steal user data, including login credentials and credit card numbers. It occurs when an attacker, masquerading as a trusted entity, dupes a victim into opening an email, instant message, or text message. The recipient is then tricked into clicking a malicious link, which can lead to the installation of malware, the freezing of the system as part of a ransom ware attack or the revealing of sensitive information. An attack can have devastating results. For individuals, this includes unauthorized purchases, the stealing of funds, or identify theft. Moreover, phishing is often used to gain a foothold in corporate or governmental networks as a part of a larger attack, such as an advanced persistent threat (APT) event.

In this latter scenario, employees are compromised in order to bypass security perimeters, distribute malware inside a closed environment, or gain privileged access to secured data.

II. RELATED WORK / LITERATURE REVIEW

1) Nutjahan, Farhana Nizam, Shudarshon Chaki, Shamim Al Mamun, M. Shamim Kaiser," Attack Detection and Prevention in the Cyber Physical System". [2016] [1] In this paper proposes Cyber Physical System cyber-attack detection and prevention To detect distributed denial of service and false data injection attacks, the Chi square detector and Fuzzy logic-based attack classifier (FLAC) were utilized. Activity profiling, average packet rate, change point detection algorithm, cusum algorithm, unexpired user sessions, injected incomplete information, and reuse of session key are some of the fuzzy features used to choose the attacks described. An example scenario has been created using OpNET Simulator. Simulation results depict that the use of Chi-square detector and FLAC are able to detect the mentioned cyber physical attacks with high accuracy. Compared to existing Fuzzy logic-based attack detector, the proposed model outperforms the traditional distributed denial of service and false data detector.

2) Yong Fang, Cheng Huang, Yijia Xu and Yang Li, "RLXSS: Optimizing XSS Detection Model to Defend Against Adversarial Attacks Based on Reinforcement Learning". [2019] [2]. In this research, we introduce RLXSS, a reinforcement learning-based strategy for optimising the XSS detection model to defend against adversarial attacks. First, the adversarial samples of the detection model are mined by the adversarial attack model based on reinforcement learning. Secondly, the detection model and the adversarial model are alternately trained. After each round, the newly excavated adversarial samples are marked as a malicious sample and are used to retrain the detection model. The proposed RLXSS model successfully mines adversarial samples that avoid black-box and white-box detection while retaining aggressive features, according to experimental data. Furthermore, by alternating training the detection model and the confronting assault model, the detection model's escape rate is continuously reduced, indicating that the model can increase the detection model's ability to defend against attacks.

3) Rishikesh Mahajan, Irfan Siddavatam, "Phishing Website Detection using Machine Learning Algorithms". [2018] [3] Phishing is the most basic method of obtaining sensitive information from unsuspecting consumers. The goal of phishers is to obtain sensitive information such as usernames, passwords, and bank account information. Cyber security professionals are now looking for dependable and consistent

detection solutions for phishing websites. The purpose of this work is to discuss machine learning technology for detecting phishing URLs by extracting and analysing various aspects of authentic and phishing URLs. To detect phishing websites, the Decision Tree, Random Forest, and Support Vector Machine algorithms are used. The goal of this study is to detect phishing URLs as well as to narrow down the best machine learning method by analysing each algorithm's accuracy rate, false positive and false negative rate.

4) Vishnu. B. A, Ms. Jevitha. K. P., "Prediction of Cross-Site Scripting Attack Using Machine Learning Algorithms". [2018] [4] Cross-site scripting (XSS) is one of the most frequently occurring types of attacks on web applications, hence is of importance in information security. XSS occurs when an attacker injects malicious code, usually JavaScript, into a web application such that it can be executed in the user's browser. Detecting malicious scripts is an important aspect of an online application's defence. This study studies the use of SVM, k-NN, and Random Forests to detect and limit known and undiscovered assaults on JavaScript code by developing classifiers. It shown that using an interesting feature set that combines language syntax and behavioural information resulted in classifiers that provide excellent accuracy and precision on huge real-world data sets without focusing solely on obfuscation.

5) Zohre Nasiri Zarandi, Iman Sharif, "Detection and Identification of Cyber-Attacks in Cyber-Physical Systems Based on Machine Learning Methods". [2020] [5] The CPS is modelled in this study as a network of agents that move in unison with one another, with one agent acting as a leader and the other agents being ordered by the leader. In this study, the proposed strategy is to employ the structure of deep neural networks for the detection phase, which should tell the system of the existence of the attack in the early stages of the attack. In the leader-follower mechanism, the employment of robust control algorithms in the network to isolate the misbehaving agent has been examined. In the presented control method, after the attack detection phase with the use of a deep neural network, the control system uses the reputation algorithm to isolate the misbehave agent. Experiments reveal that deep learning algorithms outperform traditional approaches in detecting assaults, making cyber security simpler, more proactive, less expensive, and considerably more successful.

6)

III. MOTIVATION

Cyber and Network attacks are an important problem in today's communication environments. The network traffic must be monitored and analyzed to detect malicious activities and attacks to ensure reliable functionality of the networks and security of users' information.
Recently, machine learning techniques have been applied toward the detection of network attacks. Machine learning models are able to extract similarities and patterns in the network traffic.
Applying machine learning algorithms can automatically build predictive models for thedetection of network attacks.
Problem Definition and Objectives ..

IV. OBJECTIVES

1) To overcome these shortcomings, there is a need to collect representative intrusion detection data to develop and analyze detection mechanisms for computer networkattacks.
2) In addition to a representative normal data, it should also contain a proper phishingattack of attacks.
3) To detect network attacks by applying machine learning methods.
4) To reduced operational time.
5) To increased accuracy and reliability.
6) To increased operational efficiency.
7) To provide data security

V. Problem Statement

- To address the aforementioned challenges, we proposed a novel algorithm and build an web based application for detection of different four types of attacks which is, SQL Injection, Cross-Site Scripting (XSS), Phishing Attacks, and Normal Intrusion Detection Attack (IDS).

- In Proposed studies shows that the problem definition gets more specific for any attack type and includes an expanded definition of the attack and its behavior. Further, we confirmed that the performance of neural network increases with increase in accuracy and performance of algorithms.

VI. PROPOSED APPROACH /METHODOLOGY

System Structure: -

1) **Planning and Setup Phase:** The attackers begin by determining who or what their target is, be it a company, an ordinary person, or an entire nation. After then, it's up to them to learn more about the company and the people that work there. It's possible to achieve this by physically travelling to the location or by watching the network traffic coming in and out [6]. The next stage is to set up the assaults by utilizing a viable method, such as a website or email with malicious links that might lead the target to a fraudulent web page.

2) **Phishing Phase:** The next phase is to utilize the gathered email addresses to deliver such spoof emails, e.g., pretending to be some reputable financial institution to the target, asking the user to modify certain information immediately by tapping on some harmful link. Emails can be sent to a group of people or to a particular person inside a company.

3) **Break-in Phase:** When a target clicks on a fraudulent link, whether malware is placed upon that system, allowing the attacker to get access and modify the system's settings, or access privileges are altered to reflect this. Sometimes, it might take you to a bogus website that requests login credentials.

4) **Information Gathering Phase:** Upon gaining access to the victim's machine, the attacker will be able to retrieve the necessary data. If the victim provides his/her login credentials to the attacker, they will have full access to the victim's account, which may result in financial consequences. Malware assaults now have the potential to provide the attacker remote access to the system, allowing him

to steal any information he desires, or to exploit the hacked devices for DDoS attacks or other malicious purposes. Rootkits are used by phishers to disguise their malicious files.

5) **Break-out Phase:** To hide their tracks, the phisher creates fake accounts on a variety of popular websites and uses that information to trick others into providing their personal information. As a result, it's been discovered that they keep tabs on how successful their attack was in order to make improvements on future ones.

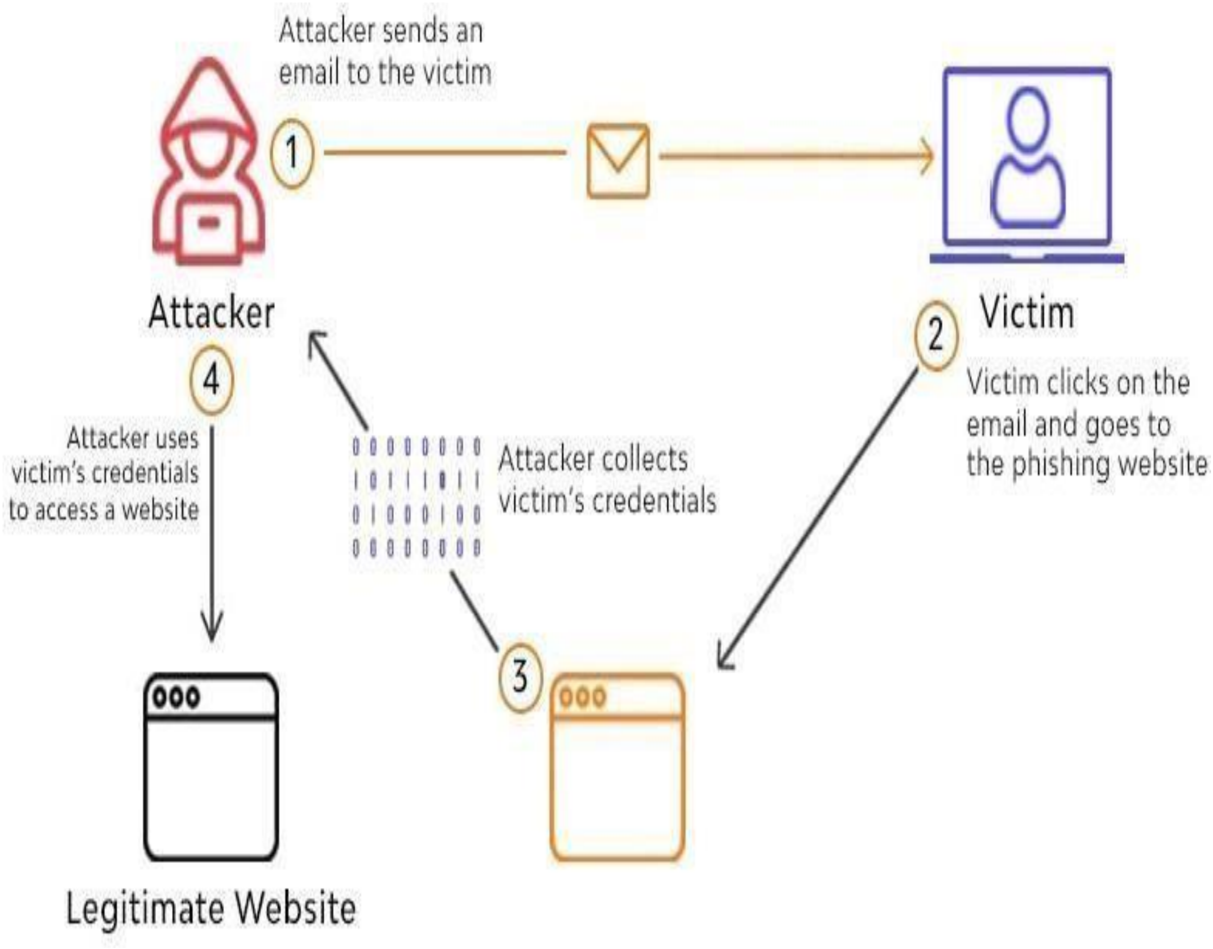

Attacker sends an email to the victim
1
Attacker
Victim
2
Victim clicks on the email and goes to the phishing website
4
Attacker uses victim's credentials to access a website
Attacker collects victim's credentials
3
Legitimate Website

Fig. 1: System Architecture .

VII. HARDWARE REQUIREMENT

- **C.P.U. :** Core 2 Due
- **Motherboard :** Intel chip/ original based Pentium pc
- **RAM :** 2 GB
- **Monitor :** Color (SVGA)
- **Resolution :** 800*600
- **HDD :** 240 GB

VIII. SOFTWARE REQUIREMENT

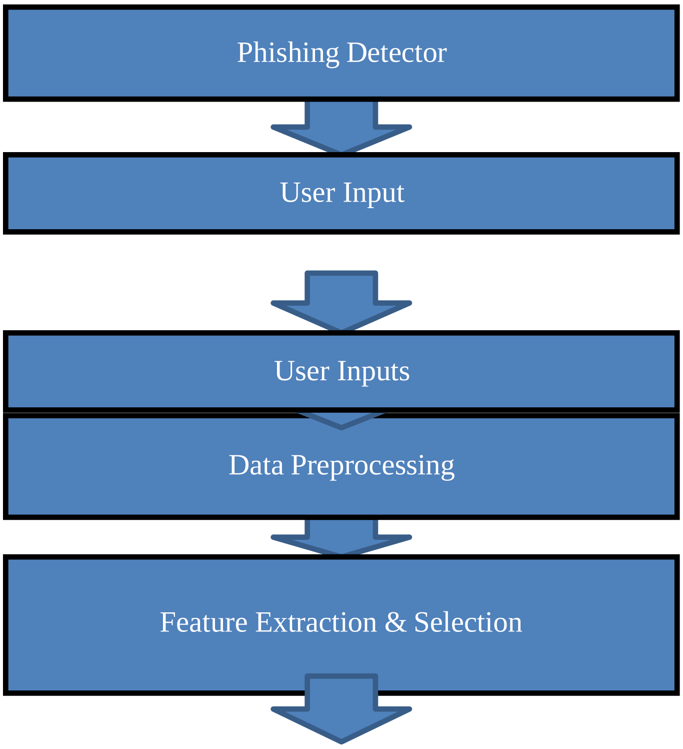
- **Operating system:** Windows 7 onwards
- **IDE :** Python IDEL
- **Design :** Photo Shop
- **Server :** Wampp Server
- **Browser :** Preferred Chrome

IX. DATA FLOW CHART

In this diagram, the user inputs are taken as the inpu to the system. The data preprocessing module is responsible for cleaning and formatting the input data to make it suitable for feature extraction and selection. The feature extraction and selection module extracts relevant features from the input data and selects the most important features.

The model training/testing module uses the selected features to train a machine learning model or apply an already trained model to test the input data. The classification results module outputs the prediction of whether the input data is a phishing attempt or not.

Finally, the alert notification module generates alerts to notify the user if the input data is a phishing attempt.



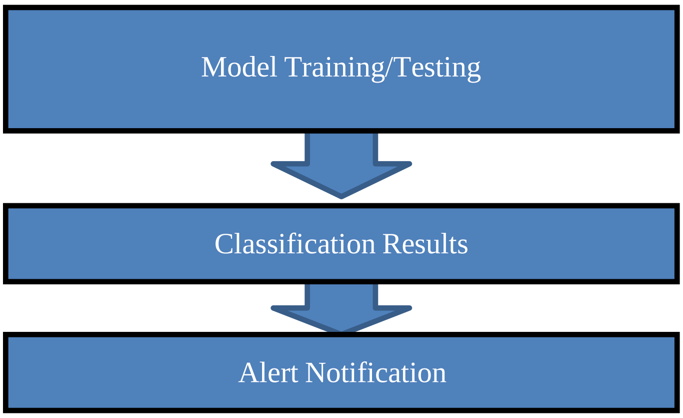


Fig. 2: Data Flow Diagram

X. Conclusions

In this study, an attempt was made to use the resilient control consensus method in complex discrete cyber-physical networks with a number of local attacks off. By applying this control method, it was observed that even in the presence of cyber-attacks, the system can remain stable and isolate the attacked node and the performance of the system is not weakened. Using the neural network used in this study, it was observed that with a deep neural network, with 7 hidden layers, the system shows better performance. Also in a recurrent neural network integrated with a deep neural network, a deep layer network with a linear function performs better. Therefore, it can be said that the system has less complexity. So With deep learning method, systems can analyse patterns and learn from them to help prevent similar attacks and respond to changing behaviour. To summarise, ML has the potential to make cyber security simpler, more proactive, less expensive, and considerably more successful. Using the collections proposed in this work, it was also possible to analyze which malicious URLs are going unnoticed by the phishing reporting platforms, demonstrating cases where fake pages were detected by the methodology proposed in this research and had not yet been assessed as phishing on the PhishTank platform, even being online for more than five days. In some situations, URLs detected as phishing by the proposed algorithm were not recognized as fraud by Virus- Total, especially those that contained suspicious download files and a valid digital certificate. It was also possible to observe which phishing maneuvers were detected as Spam, highlighting the cases in which false messages are intended for infection by malicious code through files attached to the message. It was also evident in the experiments from the period 3/12/2020 to 7/05/2020, that the main malicious maneuvers involve themes associated with COVID-19 and government services. This work has language independence when compared to the related works presented in section II. It has achieved accuracy in detecting phishing characteristics, which can vary between 73.3% and 31 97.66%. The lowest accuracy values were observed in situations where the pages present. The scraping technique proposed in this work has some limitations in extracting information from pages with more complex structures, which use other languages embedded in the layout. Heavier pages took response time over the 30 seconds to return the extracted values, especially regarding the verification of hyperlinks. Also, the structure is dependent on a white list previously configured in the Firewall, containing a list of official URLs, thus avoiding a large number of false positives and checking real pages.

References

[1] Z. N. Zarandi and I. Sharifi, "Detection and Identification of Cyber-Attacks in Cyber- Physical Systems Based on Machine Learning Methods," 2020 11th International Conference on Information and Knowledge Technology (IKT), 2020

[2] Nurjahan, F. Nizam, S. Chaki, S. Al Mamun and M. S. Kaiser, "Attack detection and prevention in the cyber Physical System," 2016 International Conference on Computer Communication and Informatics (ICCCI), 2016, pp. 1-6.

[3] Ding Chen, Qiseng Yan, Chunwang Wu and Jun Zhao, “SQL Injection Attack Detection and Prevention Technique Using Deep Learning,” Journal of Physics: Conference Series, Volume 1757, International Conference on Computer Big Data and Artificial Intelligence (ICCBDAI 2020) 24-25 October 2020, Changsha, China

[4] Ercan NurcanYılmaz, SerkanGönen, “Attack detection/prevention system against cyberattack in industrial control systems,” Computers & Security Volume 77, August 2018, Pages 94-105

[5] Arpitha. B, Sharan. R, Brunda. B. M, Indrakumar. D. M, Ramesh. B. E, “Cyber Attack Detection and notifying system using ML Techniques,” International Journal of Engineering Science and Computing (IJESC), Volume 11**, Issue No.06**

[6] P. Prakash, M. Kumar, R. R. Kompella and M. Gupta, Phishnet: Predictive Blacklisting to Detect Phishing Attacks, In INFOCOM, 2010 Proceedings IEEE, pp. 1 5, March (2010).

[7] Y. Cao, W. Han and Y. Le, Anti-Phishing based on Automated Individual White-List, In Proceedings of the 4th ACM Workshop on Digital Identity Management ACM, pp. 51 60, October (2008).

[8] Y. Joshi, S. Saklikar, D. Das and S. Saha, PhishGuard: A Browser Plug-In for Protection from Phishing, In 2nd International Conference

on Internet Multimedia Services Architecture and Applications, IMSAA 2008, IEEE, pp. 1 6,December (2008).

[9] Y. Zhang, J. I. Hong and L. F. Cranor, Cantina: A Content-Based Approach to Detecting Phishing Web Sites, In Proceedings of the 16th International Conference on World Wide Web, ACM, pp. 639 648, May (2007).

[10] L. Wenyin, G. Huang, L. Xiaoyue, Z. Min and X. Deng, Detection of Phishing Webpages based on Visual Similarity, In Special Interest Tracks and Posters of the 14th International Conference on World Wide Web, ACM, pp. 1060 1061, May(2005).