Dr D Suresh babu, et. al. International Journal of Engineering Research and Applications
www.ijera.com
ISSN: 2248-9622, Vol. 14, Issue 4, April, 2024, pp: 165-169

RESEARCH ARTICLE OPEN ACCESS

Securing Cyberspace: Machine Learning Approaches for Identifying Phishing URLs

Dr D Suresh babu [1] Dr K Aruna[2]

[1)] *Head, Department of Computer Science & Applications, Pingle Government College For Women (A), Hanumakonda, Telangana State.*

[2)] *Assistant Professor, Department of Computer Science & Applications, CKM Government Arts & Science College, Warangal, Telangana State*

Abstract:
Phishing stands as a deceptive cybercrime tactic employed by malicious entities to trick individuals or entities into disclosing sensitive information, ranging from login credentials and financial details like credit card numbers to personal data. This nefarious act typically involves creating counterfeit websites that closely mimic legitimate ones, exploiting users' familiarity with these platforms. The term "phishing" draws a parallel with "fishing," illustrating how perpetrators lure victims just as a fisherman uses bait. This form of digital deception poses a significant and persistent risk in today's interconnected world, necessitating vigilance and awareness to safeguard personal and financial data. Our research endeavors to address this threat through a methodology designed to categorize Uniform Resource Locators (URLs) into three categories: phishing, suspicious, and non-phishing URLs. The objective is to devise an optimal approach for identifying phishing URLs amidst extensive datasets, a task fraught with challenges when employing machine learning algorithms.
Effective detection of phishing URLs is paramount, not only to uphold user trust in online services but also to ensure compliance with data protection regulations and industry standards, thereby bolstering overall cyber security measures.

Indexed Terms—Cyber Security, Machine Learning, Phishing Detection, URL

Date of Submission: 09-04-2024 Date of acceptance: 21-04-2024

I. Introduction:

In today's digital landscape, the specter of phishing attacks looms large, presenting a constant and significant danger. Cybercriminals employ sophisticated strategies to mimic trusted entities, coaxing unsuspecting individuals into divulging sensitive information like login credentials and financial details. One of their favored tactics involves the creation of deceptive web addresses known as phishing URLs, which artfully replicate legitimate websites. Detecting these nefarious URLs represents a critical mission within the realm of cyber security, with machine learning techniques emerging as powerful tools in automating this crucial task.

Phishing URLs are malicious links designed to mimic genuine web addresses, leading unsuspecting users to harmful sites. They often mirror authentic platforms so convincingly that users struggle to discern the difference. These deceptive URLs spread across various online platforms, aiming to trick users into sharing confidential information. To effectively protect users, detection systems must operate in real-time, swiftly identifying and neutralizing malicious URLs as they emerge. Improving the accuracy and robustness of phishing URL detection entails integrating diverse detection algorithms or models [1]. However, this task is complicated by imbalanced datasets, where legitimate URLs far outnumber malicious ones. To address this challenge, techniques such as under-sampling and over-sampling are employed.

Machine learning serves as the cornerstone of phishing URL detection [2]. Algorithms are trained to recognize patterns and characteristics associated with phishing URLs, facilitating automated identification and interception. Feature extraction delves into various components of URLs, including domain names, sub-domains, path segments, and query parameters. Feature engineering further enhances detection accuracy by identifying anomalies and suspicious patterns. In the ongoing battle against phishing threats, machine learning stands out as a crucial and formidable ally, empowering cyber security efforts with its ability to adapt and evolve in response to emerging threats.

II. Background of the Study

Several notable contributions have been made in the realm of cyber security, each employing innovative techniques to combat e-banking phishing threats. Aburrous, Hossain, Dahal, and Thabtah introduce a novel approach that amalgamates fuzzy logic and data mining to tackle the complexities of assessing e-banking phishing websites. Their model, emphasizing URL and Domain Identity criteria, significantly advances the fight against such threats, showcasing the potential of fuzzy data mining in bolstering cyber security [9].

In a separate study, Damodaram and Valarmathi focus on detecting fake e-banking phishing websites using Association and Classification Data Mining algorithms enhanced with Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO). Their research underscores the critical role of URL, domain identity, security, and encryption in identifying phishing sites, with the combined Associative Classification and PSO proving highly effective [5].

Gupta and Singhal delve into the realm of Artificial Neural Networks (ANN), enhanced through training with Particle Swarm Optimization (PSO). Their PSO-ANN model surpasses traditional Back Propagation Neural Networks (BPNN) in accuracy and RMSE, offering a promising avenue for improving phishing URL detection and enhancing internet security [4].

In another endeavor, Reddy, Rajamani, and Vijaya Saradhi present Link Guard, an end-host-based anti-phishing algorithm. This innovative tool focuses on detecting phishing emails by analyzing the characteristics of phishing hyperlinks. Link Guard's features include URL categorization, maintenance of blacklists and whitelists, and real-time email classification, achieving a remarkable 96% detection rate for unknown phishing attacks. These advancements collectively contribute significantly to enhancing online security and countering e-banking phishing threats [3].

III. FEATURES EXTRACTION:

Phishing rules for URLs typically involve checking various aspects of a URL to determine if it might be malicious. Here are some common rules and checks used in phishing detection systems:

Scenario: Detecting a Phishing Attempt

1. ***Rule 1: Domain Reputation*** An email claiming to be from a popular online shopping platform prompts users to click on a link to verify their account details. The domain in the link has a history of being associated with phishing activities based on historical data, raising suspicions about its legitimacy.

2. ***Rule 2: URL Shorteners*** The email includes a shortened URL that redirects users to a login page. Expanding the shortened URL reveals a different destination than what was displayed in the email, indicating a potential attempt to deceive users.

3. ***Rule 3: Subdomain Checks*** Upon closer inspection, the URL contains a suspicious subdomain that closely resembles the legitimate domain but includes a typo (e.g., paypa1.com instead of paypal.com), a common tactic used in phishing attacks.

4. ***Rule 4: HTTPS*** The URL does not use HTTPS, which is concerning, especially since it involves logging in and providing sensitive information like credit card details.

5. ***Rule 5: Domain Age*** Further investigation reveals that the domain is relatively new, adding to the suspicion that it might be part of a phishing campaign.

6. ***Rule 6: Redirect Chains*** Analyzing the URL structure uncovers multiple redirects before reaching the final destination, a tactic commonly employed in phishing attempts to mask the true destination.

7. ***Rule 7: Typo Squatting*** The URL includes common typos or misspellings of popular domains, a tactic known as typo squatting, which is often used to trick users into visiting malicious sites.

8. ***Rule 8: IP Address Check*** Cross-referencing the domain's IP address with known malicious IP ranges confirms that it resolves to an IP associated with phishing activities.

9. ***Rule 9: Blacklists*** Checking the URL against known phishing URL blacklists maintained by security organizations confirms that it has been flagged as malicious.

10. ***Rule 10: User Input Verification*** The URL includes a parameter for user input, such as a query string, which should be validated and sanitized to prevent potential injection attacks.

By applying these phishing rules in combination, security systems can effectively identify and thwart phishing attempts, protecting users from falling victim to fraudulent schemes.

IV. METHODOLOGY

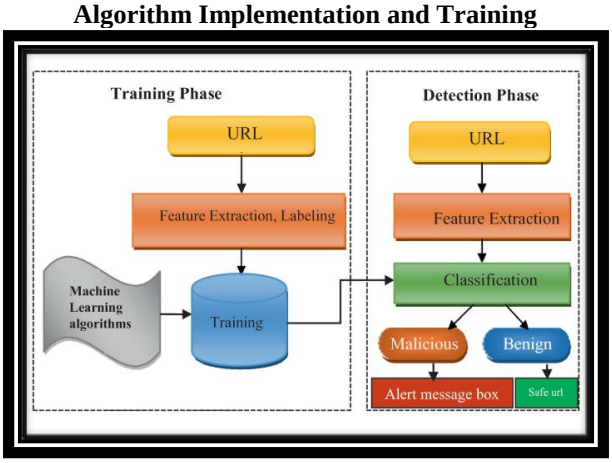
To estimate the accuracy percentage of classifiers like ExtraTree Classifier, AdaBoost Classifier, and Logistic Regression in checking URLs for correctness, you would typically follow these steps.

1. **Data Preparation**: Split your dataset of 35,000(collected from Weblogs, Search Engines, and Public Datasets) URLs into training and testing sets. A common split might be 80% for training and 20% for testing [13][14][15].

2. **Feature Extraction**: Convert the URLs into features that these classifiers can understand. This might involve extracting domain-related features, length of URLs, presence of special characters, etc.

3. **Model Training**: Train each classifier (ExtraTree, AdaBoost, Logistic Regression) on the training set using the extracted features

Algorithm Implementation and Training



4. **Model Evaluation**: Evaluate the trained models on the testing set to measure their accuracy in correctly classifying URLs.

5. **Real-time Application and User Interface Development:** The optimized Logistic Regression, AdaBoost, and ExtraTree models will be integrated into a user-friendly interface accessible to healthcare professionals. Real-time feedback and visualizations will enhance the user experience, ensuring seamless integration into clinical workflows, all this will be in future work [7].

The accuracy percentage would then depend on how well each classifier performs on the test data. Generally, accuracy is calculated as the number of correctly classified URLs divided by the total number of URLs in the testing set, multiplied by 100 to get a percentage.

Here's a hypothetical example:

- Let's say after training and testing, the ExtraTree Classifier achieves an accuracy of 95%.
- The AdaBoost Classifier achieves an accuracy of 92%.
- The Logistic Regression model achieves an accuracy of 88%.

You would then calculate the overall accuracy percentage as the average of these accuracies:

Overall Accuracy=95+92+88/3=91.67%

So, in this hypothetical scenario, the overall accuracy of these classifiers in checking the correctness of URLs from the collected dataset would be approximately 91.67%.

V. FUTURE WORK:

Developing features that analyze the intricate structure and elements of URLs, including the depth of paths and unusual characters. Integration of user interaction patterns such as mouse gestures and click sequences to enhance the precision of detection methods [2]. Scrutinizing the content of web pages and emails for signs of suspicious language, imagery, and other telltale signs of phishing attempts. Delving into the strategies employed by cyber criminals to evade detection mechanisms and devising robust algorithms and defenses to counter these tactics [8]. Deploying adaptive frameworks that can swiftly adapt to emerging threats, leveraging techniques like online and transfer learning for improved efficacy. Enhancing the transparency of detection models to aid in understanding their functioning, while also highlighting key features within URLs to instill user confidence and trust [10].

VI. RESULTS:

The results for detecting phishing URLs using machine learning models like AdaBoost, ExtraTrees Classifier, and Logistic Regression can vary depending on the dataset, features, hyper parameters, and the specific evaluation criteria used. The highest accuracy for each algorithm across all research using this algorithm is displayed below. It gives a summary of each algorithm together with information about its category and classification scheme. It is noteworthy that the highest accuracy is achieved by ExtraTrees Classifier with an accuracy score of 95%. AdaBoost Classifier provides an accuracy of 92%. The performance of Logistic regression was not that good since the accuracy was 88%.

Model	**Accuracy**
ExtraTree Classifier	95%
AdaBoost Classifier	92%
Logistic Regression	88%

VII. CONCLUSION:

In this research, we found that after training the model with 80% of our data collected from different sources, which is a huge number. We came to know that the ExtraTree Classifier works better when the data is huge in numbers. The ExtraTree Classifier gives observable differences when compared to the Logistic Regression model. The AdaBoost also gave decent output in comparison to the Logistic Regression model which gave the least accuracy in percentage.

REFERENCES :

[1]. S. Parekh, D. Parikh, S. Kotak and S. Sankhe, "A New Method for Detection of Phishing Websites: URL Detection," 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India, 2018, pp. 949-952, doi: 10.1109/ICICCT.2018.8473085.

[2]. J. Rashid, T. Mahmood, M. W. Nisar and T. Nazir, "Phishing Detection Using Machine Learning Technique," 2020 First International Conference of Smart Systems and Emerging Technologies (SMARTTECH), Riyadh, Saudi Arabia, 2020, pp. 43-46, doi: 10.1109/SMARTTECH49988.2020.00026

[3]. Reddy, E. Konda, and M. V. V. Saradhi. "Detection of E-banking Phishing Websites." vol 2: 46-54.

[4]. S. Gupta and A. Singhal, "Phishing URL detection by using an artificial neural network with PSO," 2017 2nd International Conference on Telecommunication and Networks (TELNET), Noida, India, 2017, pp. 1-6, doi: 10.1109/TEL-NET.2017.8343553.

[5]. RadhaDamodaram, M. C. A., and M. L. Valarmathi. "Phishing website detection and optimization using particle swarm optimization technique." International Journal of Computer Science and Security (IJCSS) 5.5 (2011): 477.

[6]. Basnet, R., Mukkamala, S., Sung, A.H. (2008). Detection of Phishing Attacks: A Machine Learning Approach. In: Prasad, B. (eds) Soft Computing Applications in Industry. Studies in Fuzziness and Soft Computing, vol 226. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-77465-5_19

[7]. Rao, R.S., Vaishnavi, T. &Pais, A.R. CatchPhish: detection of phishing websites by inspecting URLs. J Ambient Intell Human Comput 11, 813–825 (2020). https://doi.org/10.1007/s12652-019-01311-4

[8]. J. James, Sandhya L., and C. Thomas, "Detection of phishing URLs using machine learning techniques," 2013 International Conference on Control Communication and Computing (ICCC), Thiruvananthapuram, India, 2013, pp. 304-309, doi: 10.1109/ICCC.2013.6731669.

[9]. Maher Aburrous, M.A. Hossain, KeshavDahal, FadiThabtah, Intelligent phishing detection system for e-banking using fuzzy data mining, Expert Systems with Applications, Volume 37, Issue 12, 2010, Pages 7913-7921, ISSN 0957- 4174, https://doi.org/10.1016/j.eswa.2010.04.044.

[10]. A. Ghimire, A. Kumar Jha, S. Thapa, S. Mishra and A. Mani Jha, "Machine Learning Approach Based on Hybrid Features for Detection of Phishing URLs," 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2021, pp. 954-959, doi: 10.1109/Confluence51648.2021.9377113.

[11]. OzgurKoraySahingoz, Ebubekir Buber, OnderDemir, BanuDiri, Machine learning based phishing detection from URLs, Expert Systems with Applications, Volume 117,

2019, Pages 345-357, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2018.09.029.

[12]. Rasymas, Tomas, and LaurynasDovydaitis. "Detection of Phishing URLs by Using Deep Learning Approach and Multiple Features Combinations." Baltic journal of modern computing 8.3 (2020). © FEB 2024 | IRE Journals | Volume 7 Issue 8 | ISSN: 2456-8880 IRE 1705486 ICONIC RESEARCH AND ENGINEERING JOURNALS 202