Phishing URL Detection Using Machine Learning

DR. S. K SINGH[1], POONAM JAIN[2], RITESH MOURYA[3], AHMAD KHAN[4]
[1] *HOD, Department of IT, Thakur College of Science and Commerce, Thakur Village, Kandivali (East), Mumbai, Maharashtra, India*
[2] *Assistant Professor, Department of IT, Thakur College of Science and Commerce, Thakur Village, Kandivali (East), Mumbai, Maharashtra, India*
[3, 4] *PG Student, Department of IT, Thakur College of Science and Commerce, Thakur Village, Kandivali (East), Mumbai, Maharashtra, India*

Abstract— Phishing is a cybercrime tactic used by malicious actors to deceive individuals or organizations into revealing sensitive information, such as usernames, passwords, credit card numbers, or other personal and financial data. This is done by an attacker by creating a replica of an existing website. This replica is an exact look-alike of famous websites that online users may look for. The term "phishing" is a play on the word "fishing" because it involves luring victims in a similar way to how a fisherman lures fish with bait. Phishing is a prevalent and persistent threat in the digital age, so it's essential to remain cautious and informed to protect your personal and financial information from falling into the wrong hands. In this research, we proposed a method to classify the Uniform Resource Locator (URL) into phishing, suspicious, and non-phishing URLs. This research aims to find the best method for finding a phishing URL when the dataset is in huge numbers. There are many challenges people face when detecting phishing URLs using machine learning algorithms. Protecting users from phishing attacks is vital to maintaining trust and confidence in online services and platforms. Compliance with data protection regulations and industry standards requires effective phishing URL detection to ensure the security of user information.

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

I. INTRODUCTION

In the digital era, phishing attacks loom as a pervasive and perilous threat. Cybercriminals employ cunning tactics to impersonate trusted entities and coax individuals into revealing sensitive information like login details and financial data. One favored method is the creation of phishing URLs web addresses that cunningly mimic genuine websites. Detecting these phishing URLs is a paramount task in cybersecurity, and machine-learning methods have emerged as potent allies for automating this endeavor. Phishing URLs are deceitful web links that masquerade as authentic but lead to malicious sites. They often mirror legitimate websites, making it arduous for users to differentiate between the two. These URLs proliferate through various online channels, with the sole aim of duping users into divulging sensitive data. To safeguard users effectively, detection systems must work in real-time, promptly spotting and thwarting malicious URLs as they surface. Enhancing accuracy and resilience in identifying phishing URLs involves amalgamating multiple detection algorithms or models [9]. However, the task is compounded by imbalanced datasets, where malicious URLs are scarce compared to legitimate ones. To rectify this, techniques like under-sampling and over-sampling are employed. Machine learning is the linchpin of phishing URL detection [5]. Algorithms are primed to discern patterns and traits linked to phishing URLs, enabling automated identification and interception. Feature extraction delves into URL components such as domain names, sub-domains, path segments, and query parameters. Feature engineering further heightens detection accuracy by flagging anomalies and suspicious markers. In the ongoing battle against phishing threats, machine learning stands as a formidable ally.

II. LITERATURE REVIEW

In a paper, authored by Maher Aburrous, M.A. Hossain, Keshav Dahal, and Fadi Thabtah. This paper introduces an innovative approach that combines fuzzy logic and data mining to address the challenges of assessing e-banking phishing websites. The model places particular emphasis on criteria such as URL and Domain Identity. By integrating these techniques, the study contributes significantly to the fight against e-banking phishing websites and underscores the potential of fuzzy data mining in cybersecurity [1]. In another paper, authored by Radha Damodaram & Dr. M.L. Valarmathi. This research focuses on the detection of fake e-banking phishing websites. It employs Association and Classification Data Mining algorithms optimized with Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO). The study highlights the importance of criteria like URL, domain identity, security, and encryption in identifying phishing sites, ultimately demonstrating the effectiveness of the combined Associative Classification and PSO [2]. In one paper, authored by Surbhi Gupta, Abhishek Singhal. This paper explores the use of Artificial Neural Networks (ANN) improved through training with Particle Swarm Optimization (PSO). The proposed PSO-ANN model outperforms traditional Back Propagation Neural Networks (BPNN) in terms of accuracy and RMSE. This research offers a promising approach to enhance the detection of phishing URLs, thereby bolstering internet security [3]. In a paper, authored by E. Konda Reddy, Dr. Rajamani, Dr. M. V. Vijaya Saradhi. The authors present an end-host-based anti-phishing algorithm called Link Guard, which focuses on detecting phishing emails by analyzing phishing hyperlinks' characteristics. Link Guard categorizes URLs, maintains blacklists and whitelists, and works in real-time to classify emails. The algorithm effectively detects up to 96% of unknown phishing attacks, making it a valuable tool for enhancing online security and countering e-banking phishing threats [4].

III. FEATURES EXTRACTION

1. Using the IP Address: If an IP address is used as an alternative of the domain name in the URL, such as "http://125.98.3.123/fake.html", users can be sure that someone is trying to steal their personal information. Sometimes, the IP address is even transformed into hexadecimal code as shown in the following link "http://0x58.0xCC.0xCA.0x62/2/paypal.ca/index.html". Rule: If The Domain Part has an IP Address → Phishing Otherwise→ Legitimate

2. Long URL to Hide the Suspicious Part: Phishers can use long URL to hide the doubtful part in the address bar. For example:
http://federmacedoadv.com.br/3f/aze/ab51e2e319e51502f416dbe46b773a5e/?cmd=_home&dispatch=11004d58f5b74f8dc1e7c2e8dd4105e811004d58f5b74f8dc1e7c2e8dd4105e8@phishing.website.html
Rule: If URL length<54 → feature=Legitimate else if URL length≥54 and ≤75 → feature=Suspicious otherwise→ feature=Phishing

3. Using URL Shortening Services "TinyURL": URL shortening is a method on the "World Wide Web" in which a URL may be made considerably smaller in length and still lead to the required webpage. This is accomplished by means of an "HTTP Redirect" on a domain name that is short, which links to the webpage that has a long URL. For example, the URL "http://portal.hud.ac.uk/" can be shortened to "bit.ly/19DXSk4" Rule: If TinyURL → Phishing Otherwise→ Legitimate

4. URL's having "@" Symbol: Using "@" symbol in the URL leads the browser to ignore everything preceding the "@" symbol and the real address often follows the "@" symbol. Rule: If Url Having @ Symbol→ Phishing Otherwise→ Legitimate

5. HTTPS (Hyper Text Transfer Protocol with Secure Sockets Layer): The existence of HTTPS is very important in giving the impression of website legitimacy, but this is clearly not enough. Checking the certificate assigned with HTTPS including the extent of the trust certificate issuer, and the certificate age. Certificate Authorities that are consistently listed among the top trustworthy names. Rule: If Use https and Issuer Is Trusted and Age of Certificate≥ 1 Years → Legitimate Using https and Issuer Is Not Trusted → Suspicious Otherwise→ Phishing

6. The Existence of "HTTPS" Token in the Domain Part of the URL: The phishers may add the "HTTPS" token to the domain part of a URL in order to trick users. For example, http://https-www-paypal-it-webapps-mpp-home.soft-hair.com/. Rule: If Using HTTP Token in Domain Part of The URL→ Phishing Otherwise→ Legitimate

7. Abnormal URL: This feature can be extracted from WHOIS database. For a legitimate website, identity is typically part of its URL. Rule: If The Host Name Is Not Included In URL → Phishing Otherwise→ Legitimate

IV. METHODOLOGY

1. Data Collection and Preparation: The data is collected from three different sources namely PhishTank, PhishStorm, and Kaggle. In total, we have used 650000 URLs. After cleaning the dataset and handling duplicated and null values we have a total of 620000 unique URLs. We have used 80 % of the data for training and 20% of the data for testing [13] [14] [15].

2. Feature Extraction and Selection: Extract relevant features from the URLs that can be used as input for the machine learning model. Some common features include:

- URL length
- Domain and subdomain analysis
- The Presence of special characters or suspicious keywords
- URL structure (e.g., number of subdirectories)
- IP address information
- Domain reputation and WHOIS data
- HTTP response status codes
- Redirects and URL shortening services

3. Algorithm Implementation and Training

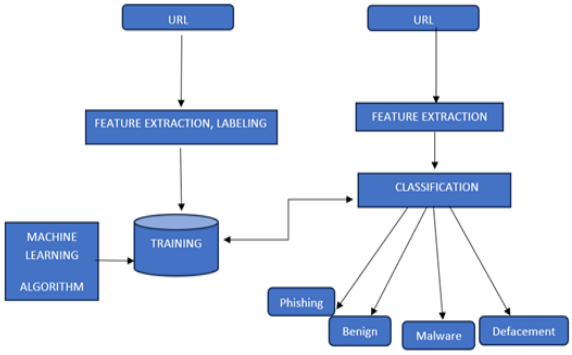

Fig 1: URL Extraction and Splitting

Fig 1 here represents the splitting of URL, feature extraction, labelling of the data.

3. Model Evaluation and Optimization: The trained models will be evaluated using metrics such as accuracy, sensitivity, specificity, and F1 score. Comparative analysis will be conducted to assess the performance of ExtraTree and AdaBoost against traditional methods.

4. 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 [8].

V. FUTURE WORK

Creating features that dissect URL structures and composition, including path depth and character-level anomalies. Incorporating user behavior patterns like mouse movements and click behavior for more accurate detection [5]. Examining web page and email content for suspicious language, images, and other phishing indicators. Investigate how attackers try to bypass detection systems. Develop robust models and countermeasures to thwart these attempts [7]. Implement adaptive systems that swiftly respond to emerging threats. Techniques like online learning and transfer learning can be useful. Improve transparency by making detection models more explainable. Identify and highlight significant features in URLs for user trust [10].

VI. RESULTS

The results for detecting phishing URLs using machine learning models like AdaBoost, ExtraTreesClassifier, 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 ExtraTreesClassifier with an accuracy score of 80.67%. AdaBoostClassifier provides an accuracy of 78.51%. The performance of Logistic regression was not that good since the accuracy was 68.61%.

Model	Accuracy
ExtraTree Classifier	80.67%
AdaBoost Classifier	78.51%
Logistic Regression	68.61%

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] 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.

[2] 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.

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

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

[5] 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/SMART-TECH49988.2020.00026.

[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] 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.

[8] 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

[9] 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.

[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).

[13] Dataset, Marchal, S. (Creator) (2014). PhishStorm - phishing / legitimate URL dataset. Aalto University. urlset(v.zip). 10.24342/f49465b2-c68a-4182-9171-075f0ed797d5

[14] Dataset, http://data.phishtank.com/data/online-valid.xml

[15] Dataset, https://www.kaggle.com/