IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Human-in-the-loop Systems for Machine Learning in Detecting Phishing Attacks in HTML URLs

[1]Vudem Sumanth Reddy, [2]Vishnu Chiluveri, [3]Venkata Vaishnavi Balpunuri, [4]Anjali Choudary Mutyala, [5]Akash Gundapuneni

[1,2,3,4,5]Research Scholars
[1,2,3,4,5] Computer Science and Engineering,
[1,2,3,4,5,]Guru Nanak Institutions Technical Campus, Hyderabad, India.

Abstract : The phishing attacks continue to raise significant threats in cybersecurity as they exploit human vulnerabilities through deceptive techniques employed in HTML URLs. The machine learning algorithms were found promising towards the detection of such attacks, but their effectiveness can be hampered by the dynamic and evolving nature of the phishing campaigns. In this paper, a new suggestion of a framework that incorporates human-in-the-loop systems with machine learning models in enhancing the capability to detect phishing attacks through HTML URLs has been proposed. We propose to present such methods and techniques that are currently being used, underline their scopes, and shortcomings. In this paper, we propose a conceptual framework to integrate human feedback into the phishing detection pipeline by fully exploiting the complementarity of automated algorithms and human intelligence. We show, through experimental validation and case studies, the effectiveness of our approach in improving detection accuracy and reducing false positive rates. Our work contributes toward establishing an even stronger and more flexible cybersecurity setup capable of mitigating even further the skyrocketing threats of phishing attacks in HTML URLs.

Keywords **- Phishing Attacks, HTML URLs, Machine Learning, Human-in-the-loop Systems, Cybersecurity, Detection Methods, Social Engineering, Cyber Threats, Human Feedback, Hybrid Detection Systems.**

I.INTRODUCTION

Phishing attacks have emerged as one of the universal cybersecurity threats that exploit human vulnerabilities to deceive users and get access to some sensitive information [1,2]. Such attacks normally use HTML URLs, where the attackers create a deceitful page or email that mimics a real source and gets the user to disclose private information, such as login details and financial or personal data [3]. Phishing campaigns are dynamic, and the sophistication level of the attackers increases very rapidly; therefore, the traditional methods to find out the evolving landscapes keep themselves in dire need of innovative approaches to tackle it.

A small amount of work has been done on the use of machine learning algorithms for phishing attack detection by checking various features derived from the HTML content and the URLs [4][5]. The limitation of this automated system for attack detection could, however, be its effectiveness due to static features and predefined rules that could often not catch even the most subtle or context-specific indicators of the attack. At this limitation, there has been an increasing interest in how human intelligence could be melded with machine learning models to better phishing detection [6].

The idea of the human-in-the-loop systems represents a radical change from the current cybersecurity paradigms, in which fully automated algorithms are equipped with the contribution of human expertise and feedback for enhanced accuracy and adaptability of detection [7, 8]. The hybrid systems leverage human cognitive abilities to identify the minutest patterns and anomalies, thus yielding much higher detection performance with resilience against sophisticated attacks [9].

In this paper, we present the integration of human-in-the-loop systems with ML in the detection of phishing attacks on HTML URLs. We will cover a wide range of methodologies and approaches in this phishing detection while talking about strengths, weakness, and challenges[10][11]. With this background, we propose a new framework that uses the power of machine learning automated algorithms in tandem with human adaptable mechanisms to build a more robust phishing detection system.

One important part in the integration of human-in-the-loop systems would be the development of effective feedback mechanisms that would foster communication between the humans and the machines. Traditional phishing detection approaches are in most cases based on automated algorithmic systems and thus show less efficiency in cases of new and rapidly changing attack techniques. This approach allows users to add real-time contributions of suspicious URLs and web content that would otherwise not have been detected by systems operating purely based on automated detection.

We illustrate the effectiveness of the proposed framework for improving the detection accuracy and reducing the false-positive rate through experimental evaluation and case studies. This has led to the development of human-in-the-loop systems that improve the resilience of cybersecurity and reduce relative impacts on the corresponding environment [12][13].

This research paper, therefore, adds to this development in cybersecurity by suggesting a human-centered approach to phishing detection in HTML URLs. We believe that by marrying human expertise with the latest in machine learning algorithms, we will be able to come up with a defense system that is more adaptive, efficient, and reliable. To tackle cyber crimes , the potential in integrating ML with human in the loop can further pave a path for machine co-working. Further research in this field will be key to rooting out our cyber issues and protecting our users from the increasing threat of phishing attacks. This paper is used to lay further groundwork toward the exploration and development of Human in the Loop system for cybersecurity applications, which would eventually lead to a much safer and secure environment.

II. LITERATURE REVIEW

Phishing attacks have sharply increased in recent years, posing a serious challenge for professionals, organizations, and law enforcement all over the world. To address this threat in HTML URLs effectively, this paper presents the use of ML algorithms for detection, and the integration of human in the loop systems to manage these threats.

Phishing attacks, a type of social engineering, involve deceiving individuals into disclosing sensitive information[1]. Extensive research on phishing has highlighted the structure of these attacks and under-scored the need for robust countermeasures. In a detailed survey, Alabdan categorized these attacks based on the techniques and vectors used by cybercriminals, which often include malicious URLs, spoofed emails, and deceptive content[3].

Machine learning algorithms are increasingly recognized as valuable tools for detecting phishing because they analyze data from HTML content and URLs. Gupta et al. evaluated modern methods for phishing detection, asserting the potential of machine learning to enhance detection accuracy and efficiency. Aljofey et al. demonstrated the effectiveness of machine learning models in identifing phishing site by using URL and HTML data[5]. Li et al. developed a stacking model that combines URL and HTML information to detect phishing webpages, illustrating the promise of ML based techniques[6].

Human-in-the-loop systems represent a new paradigm in cybersecurity, where automated algorithms and human work together to improve detection accuracy and adaptability[15]. Wu et al. highlighted the potential applicability of human-in-the-loop systems across knowledge over various domains, particularly in cybersecurity[10]. Zanzotto emphasized the cooperative interaction between humans and machines in decision-making processes[11]. Agnisarman et al., examined human-machine cooperation in challenging tasks, demonstrating the effectiveness of human-in-the-loop systems in enhancing cybersecurity operations[12].

Recent research has tended to mix machine learning algorithms with human-in-the-loop methods for the purpose of phishing detection using HTML URLs. Sarker et al. express a summary of the role of human feedback in improving detection accuracy from the perspective of cybersecurity data science in machine learning [13]. Ouyang et al. proposed language modeling training with human-in-the-loop feedback on written instruction completion, proving the possibility of human intelligence with machine learning workflows in practice [24]. Griffith et al. discussed techniques which were policy-shaping that combine human feedback in a reinforcement learning algorithm, pointing to the potential of human-in-the-loop systems in cybersecurity applications [25]. These techniques emphasize the realization of the need to combine human intelligence with machine learning algorithms to combat phishing[25].

In summary, it exposes the various areas of phishing attacks in regard to the role machine learning plays in their detection, emergence in human-in-the-loop systems as regards to machine learning, and integration. Based on these, in this paper, we develop a new architecture using an adaptive and robust phishing detection system by combining automated ML algorithms with human feedback mechanisms. Future research in the area would further the development of more resilient cybersecurity solutions against evolving phishing threats.

III. METHODOLOGY

3.1 Data Collection and Preprocessing

3.1.1 Dataset Collection:

The dataset was collected from numerous sources of cybersecurity repositories, a public set of phishing databases, web crawlers, and from industry partners. The dataset contains URLs that are classified according to the level of security threat such as “0” for “safe URL’s”, “1” for those with “suspected phishing”, and “2” for those confirming the “presence of malware”. The data collection was in a way to represent and reflect diversity from domains, regions, etc., in capturing the diversity of phishing attempts.

3.2 Data Preprocessing:

The raw URL data goes through a thorough pre-processing stage for feature extraction and making it ready for use in machine learning algorithms. The pre-processing stage involves:

3.2.1 Feature Extraction: Features such as length, count of letters, digits, and special characters in a URL, presence of shortened URLs, abnormal URL patterns, secure HTTP protocol, IP address, URL region, root domain, and URL ID are extracted from the raw data.

3.2.2 Missing Values: Using imputation techniques, the data is made data complete and reliable for analysis.

3.2.3 Encoding Categorical Variables: Categorical variables are converted into numeric values that can be used as input in a machine learning model.

3.2.4 Data Scaling: The numeric data is scaled accordingly to avoid bias during features precessing in the training of the model.

3.3 Existing Approach:

3.3.1 J48 Algorithm: The most widely used algorithm for phishing detection is J48 algorithm, an open source java implementation of C45 decision tree algorithm. J48 is derived from the C4.5 algorithm by Ross Quinlan and is best suited for the generation of decision trees over extracted features of the URLs and HTML content. The recursive process employs a fast selection of the most informative features at each node to classify instances into phishing or legitimate categories. It studies attributes like the length of the URL, age of domain, specific keyword presence in the URL at a given node to take an informed decision. The Decision Tree Model is an interpretable and easily understood model by security analysts. Therefore, it is of great importance in valuing phishing detection.

3.4 Proposed Techniques and Algorithms:

3.4.1 Ensemble Methods: This is where ensemble methods, such as Random Forest, apply to combine multiple decision trees for better phishing detection. The most important thing about Random Forest is its great ability to adapt to the treatment of data set features with great diversity and to avoid overfitting and noise contained in data.Also while bringing robustness and generalization to the phishing detection system.

Support Vector Machines (SVMs) is a supervised learning model which has a hyperplane, a plane in the feature space, it is one dimension less than the space in which it is embedded. Support Vector Machines (SVMs) form hyperplanes that separate instances of phishing from legitimate ones in high-dimensional feature spaces.

The system utilizes Support Vector Machines (SVMs) because of their characteristic of learning with complex decision boundaries. This characteristic of theirs has made them form a strong line of defense against the growing and stronger approaches of phishers. On the other hand, SVMs optimize by finding the best hyperplane that maximizes the margin between classes and, hence, separates phishing from legitimate URLs effectively to improve system accuracy.

3.4.2 k-NN (k-Nearest Neighbors):
The k-Nearest Neighbors (k-NN) algorithm classifies by proximity; thus, it is effective in recognizing patterns within HTML and URL structures. k-NN does this by measuring the similarity of instances based on their feature vectors and assigning labels to new instances based on the majority class among the k closest neighbors of the instance. This method makes k-NN very appropriate for the detection of even very minimal changes or patterns in phishing URLs, adding to the total effectiveness of the detection system.

3.4.3 Human-in-the-loop Systems Integration with Machine Learning: Behind human-in-the-loop systems, there is the same idea: introduce machine learning algorithms to better phishing detection in HTML URLs.This covers the development of feedback mechanisms that would facilitate input from the users in real-time of the URLs of suspicious websites and their content.
According to the Review of Existing Phishing Detection Mechanisms, Volume 3, the system is adaptive for identifying phishing using human expertise and feedback. Human feedback mechanisms allow the system to continuously learn and update in the quest to enhance protection against newly evolved tactics.

3.4.4 Training and Evaluation :
The model is trained in way that is robustness and applicability in a practical setup. The data set is divided into two, they are training set and validation set. The cross validation techniques are used to avoid overfitting, improve generalization to make the model perform well on over unseen data.

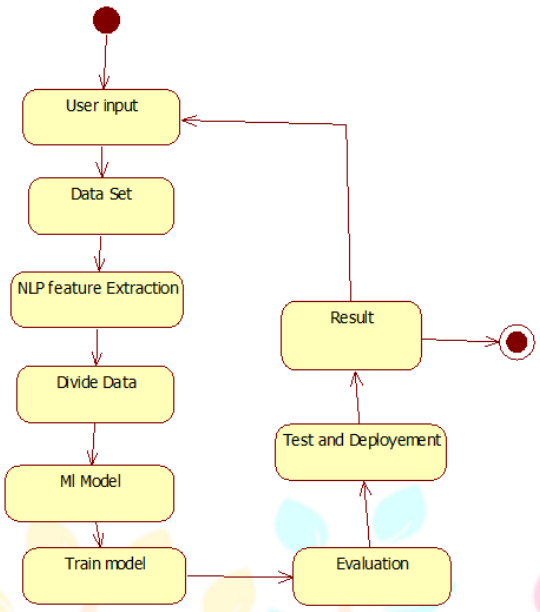
3.4.5 Evaluation Metrics:
The performance graph is used to evaluate a model are accuracy, precision, recall under various conditions. Additional metrics include mutual information and structural similarity indexes, which assess multimodal fusion efficacy and overall system performance. This would give valuable inputs on the strengths and limitations of the detection systems, which could give direction for further refinement and optimization towards the same.

3.4.6 Testing:

The models are broadly tested in different environmental conditions for its stability and performance, especially in the use of temperature data. The model should be tested under real environmental conditions, simulating every situation in which the model is supposed to be applied. Through intensive tests, it makes sure that under all the conditions it has anticipated, its performance is maintained, further consolidating the robustness of the phishing detection system to be more reliable in practice.

Its purpose is to develop adaptive and reliable defense mechanisms against emergent phishing attacks in HTML URLs, which threaten individual cyberspace safety by integrating the machine learning algorithms into human-in-the-loop systems and employing strict training and evaluation methodologies[22].

Figure 3.4.1: Working Model



IV. EXPERIMENT AND RESULT:

The experiment and results of the above tell us about the details of the proposed system's design, implementation, and performance evaluation.There is a detailed overview of the experiment methodology, includes data set, data cleaning, NLP feature extraction , Divide data, ML model, Train model, Evaluation, test and Deployment.

4.1 Experiment Methodology:

4.1.1 Data Set: A dataset is a structured collection of data, usually organized into rows and columns, where each row represents an individual instance or observation, and each column represents a specific attribute or feature of that instance. Datasets can come in various forms, such as spreadsheets, databases, text files, or specialized formats for specific applications

4.1.2 Data Cleaning: Data cleaning is the process of identifying and correcting errors or inconsistencies in a dataset to improve its quality and reliability for analysis. It involves tasks such as handling missing values, removing duplicates, correcting inaccuracies, standardizing formats, and dealing with outliers.

4.1.3 NLP Feature Extraction: Natural Language Processing (NLP) feature extraction involves transforming text data into numerical or categorical features that can be used for machine learning tasks.

4.1.4 ML Model: Machine learning models are computational algorithms that learn patterns and relationships from data to make predictions or decisions. They encompass a variety of techniques such as regression, classification, clustering, and deep learning, and are trained on labeled or unlabeled datasets to generalize patterns and solve specific tasks, enabling automated decision-making in various domains.

4.1.5 Train Model: Training a model using a training dataset involves model learning patterns and relationships in a way to minimize prediction errors and give accurate predictions or classifications.

4.1.6 Test and Deployment:

Testing a model involves assessing its performance on unseen data to evaluate the effectiveness of the trained model predictions or classifications and to ensure it generalizes well and meets desired accuracy thresholds. Deployment involves the trained model which is tested being pushed into the production line for predicting or classifying on the unseen and real-time data.

We are testing the trained ML model using the test data set, to evaluate the performance of the trained model ability to predict whether the given URL is normal or malicious or else if it could be a fake URL being used for Phishing attacks. The performance of the model is represented in the form of a bar graph. The test accuracy of the trained model is 96 percent against the train accuracy of 90 percent.

Thus indicating the model is giving a resilient performance. As for how it is done is shown in the following sample tables and figures given below.

Table 1 Training set used to train the model

URL	TYPE
www.gamespot.com/xbox360/action/alanwake/	phishing
www.eurogamer.net/articles/p_alanwake_nextgen	phishing
www.fileplanet.com/101264/0/0/0/1/section/Movies	phishing
www.gamefaqs.com/xbox360/928006-alan-wake/data	phishing
www.gamespot.com/pc/adventure/aloneinthedarktrilogy/	phishing
http://122.232.53.176:49541/Mozi.m	malware
http://111.42.103.27:36535/Mozi.m	malware
http://111.43.223.112:47199/Mozi.m	malware
http://42.235.63.163:42517/Mozi.m	malware
ideas.repec.org/j/G29.html	benign
mayfairshoppingcentre.com/	benign
answers.com/topic/tousignant	benign
http://torcache.net/torrent/CDB448FA6438C6BC82765547A51D8E553AD7F0B3.torrent?title=[kickass.to] mrs.brown.s.boys.d.movie.2014.720p.brrip.x264.yify	benign
perfectpeople.net/biography/5780/ramona-amiri.htm	benign
stuffaboutnames.com/wilson/index.htm	benign

Table 2 Data for testing of the model trained.

Url length	letters_count	digits_count	special_chars_count	shortened	abnormal_url	secure_http	ip_address	url_region	root_domain	url_id
129	104	6	19	0	1	0	0	32604616	72933545	0
119	73	20	26	0	1	0	0	32604616	2159309	1
19	17	0	2	0	0	0	0	32604616	70075576	2
49	33	6	10	0	0	0	0	32604616	41540124	0
16	14	0	2	0	0	0	0	32604616	64530026	0
28	25	0	3	1	0	0	0	32604616	59876238	0
38	23	10	5	0	0	0	0	32604616	27914812	2
32	28	0	4	0	0	0	0	32604616	44132954	0
257	140	94	23	0	1	0	0	32604616	70511031	2
14	11	0	3	0	0	0	0	86977603	15181556	0
82	64	3	15	0	1	0	0	20751160	123982	1
31	23	2	6	0	0	0	0	81788327	12029793	0
40	34	0	6	0	0	0	0	32604616	43594747	0

Figure 4.1.7: It shows the prediction based on the variables

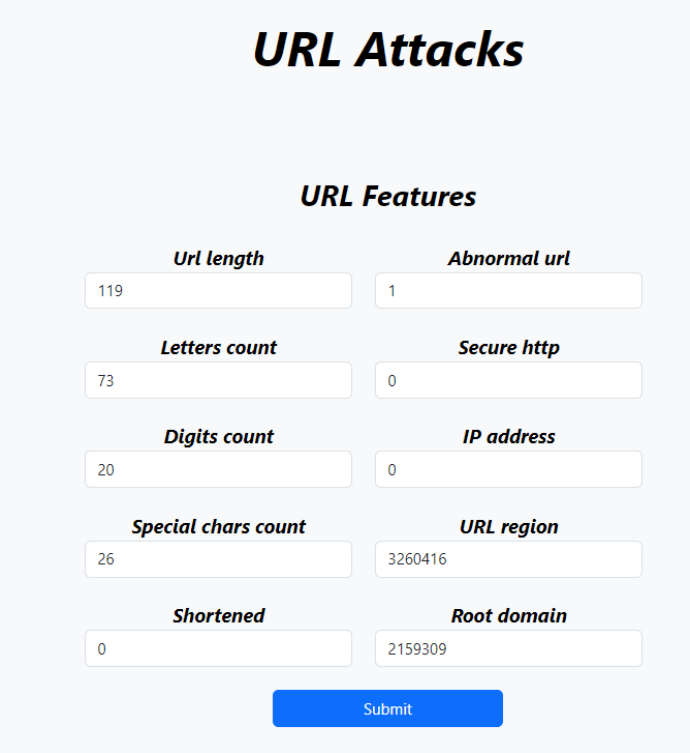


Figure 4.1.8: It shows the result i.e the type of URL(in the above it shows the given URL could be prone to phishing)

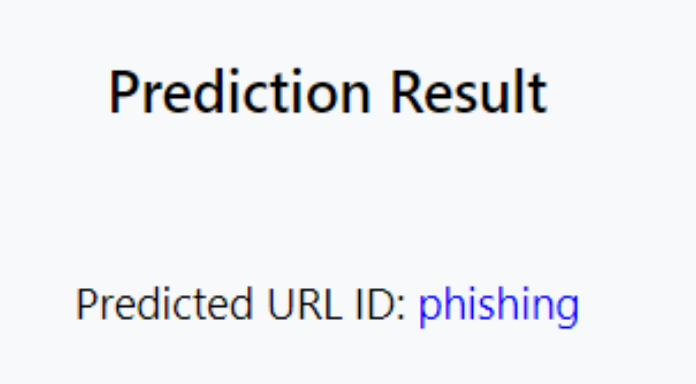


Figure 4.1.9: It shows the result i.e the type of URL(in the above it shows the given URL could be prone to phishing)

Figure 4.1.10: Showing the given URL is of Normal i.e it's safe.

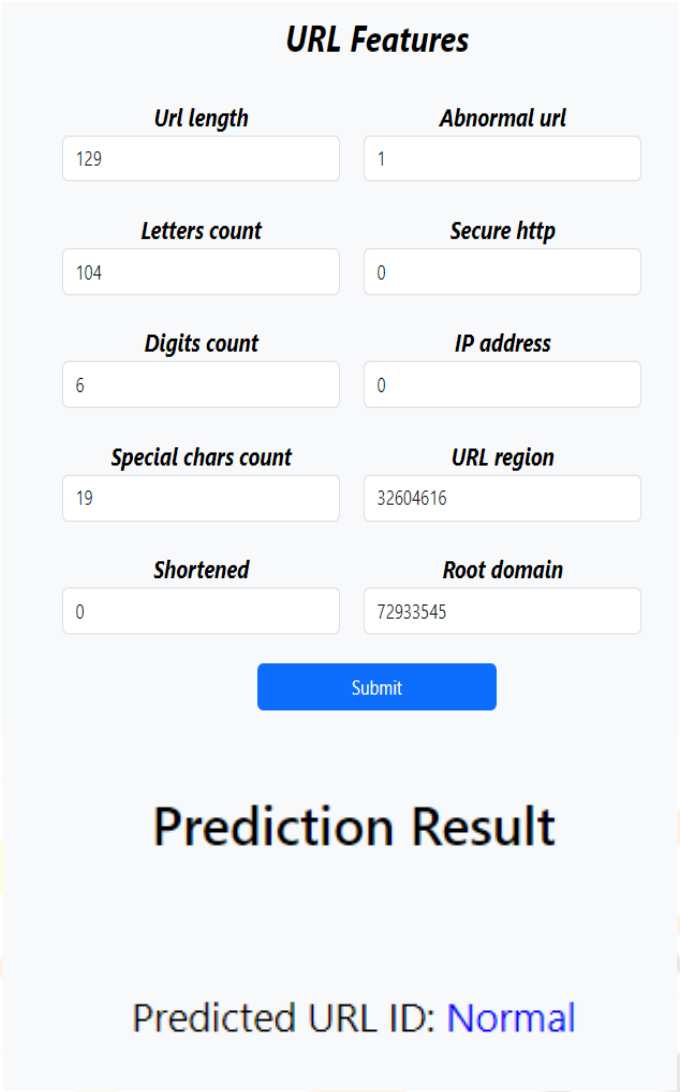
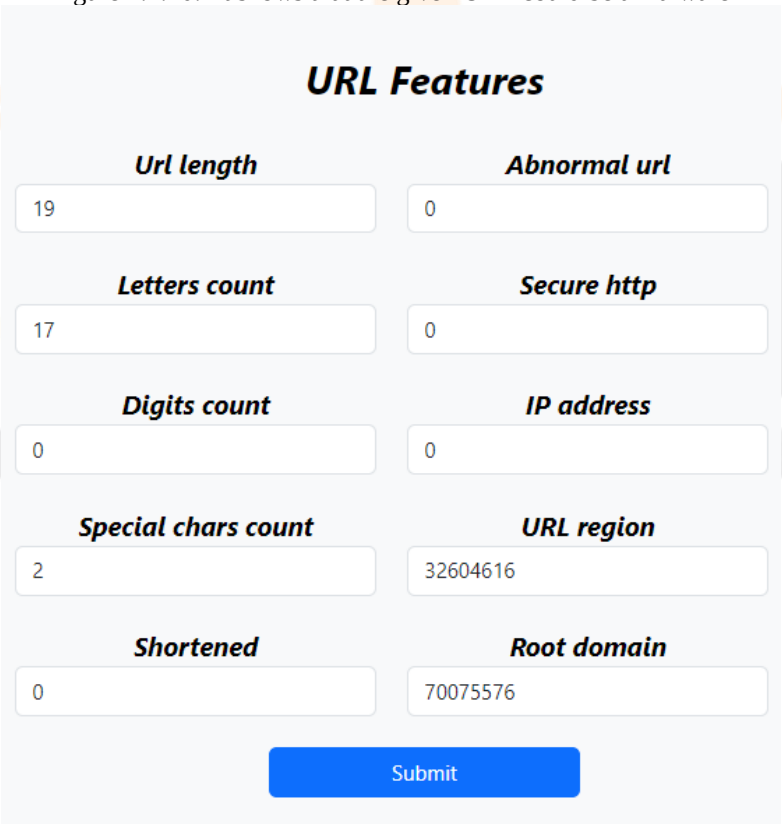
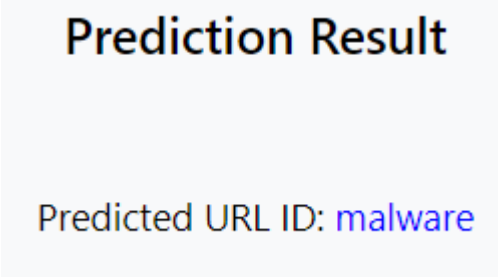


Figure 4.1.10: It shows that the given URL could be a malware





V. DISCUSSION:

In recent years, phishing attacks have emerged as a significant cyber threat, employing social engineering tactics to deceive users and steal sensitive information such as personal identities and financial data [1]. Attackers often masquerade as legitimate sources, leveraging platforms like email and social media to reach unsuspecting victims or in simple terms the attacker acts as a legitimate user to get your data and steal it . The presence of social media platforms has further facilitated these attacks, allowing attackers to target a wide audience with minimal effort .At a worry rate , reports from the Anti-Phishing Working Group (APWG) reveal a sharp rise in phishing attacks, with a staggering 250,000 increase, reported in January 2021 alone [4]. Moreover, the frequency of business compromises rose by 56% between the last quarter of 2020 and the first quarter of 2021, with financial institutions, social media, and web emails emerging as prime targets [5].

To combat this growing spree of attacks , organizations have majorly relied on human expertise to detect phishing attempts . However, the inherent complexity of these attacks, coupled with the evolving tactics employed by perpetrators, has rendered traditional detection methods increasingly ineffective . While human analysts check message attachments such as URLs and email IDs for signs of phishing, attackers continuously devise new strategies to evade detection . For instance, they craft phishing URLs and web pages that closely mimic legitimate ones, which are almost difficult to differentiate.

Researchers have explored various detection solutions, including blacklists, traditional machine learning, and deep learning (DL) approaches . Blacklists entail maintaining lists of known phishing URLs to block suspicious activity, but they are limited by their reactive nature . Traditional machine learning models, while capable of detecting phishing attacks, require manual feature extraction and struggle to keep pace with evolving attack techniques . DL, on the other hand, shows good results in automating feature extraction but faces challenges related to dataset size and model complexity .

Researchers have investigated different types of data and feature extraction methods to enhance phishing detection . URL-based approaches focus solely on URL information, offering the advantage of early detection without exposing users to potential risks . Content-based methods, meanwhile, analyze web page content for indicators of phishing but entail the risk of inadvertently triggering malicious activity. Hybrid approaches seek to leverage the strengths of both URL and content-based features, offering a potential avenue for improving detection accuracy.

Overall,the escalating threat posed by phishing attacks shows the urgent need for robust detection mechanisms . The aim is to address this imperative by conducting a comprehensive survey and analysis of HTML and URL-based phishing attacks, with a specific emphasis on the development and evaluation of machine learning models for automated detection.

VI. CONCLUSION:

In conclusion, it's clear that adopting intelligent detection methods is crucial for effectively combating HTML URL phishing attacks. While our proposed solutions show promising results, future research should focus on exploring key feature representations, refining DL architectures, and integrating human feedback to make phishing detection systems more resilient against evolving threats. Combining machine learning with human insight can significantly strengthen defenses against cyberattacks.

This paper underscores the vital role of machine learning in enhancing cybersecurity, especially in phishing detection. As attackers devise more sophisticated ways to deceive users and steal sensitive information, detection systems must evolve equally or surpass these threats to be effective. Integrating human feedback can improve the adaptability and accuracy of detection models by leveraging the collective expertise of human analysts to spot new phishing tactics.

Collaboration among researchers, industry stakeholders, and cybersecurity professionals is essential to foster innovation and develop robust solutions to address the constantly changing landscape of phishing attacks. By staying vigilant and proactive, we can create a more secure and resilient cyber environment for everyone involved.

VI. FUTURE WORK:

In the future, the following methods and techniques for phishing detection methods could be further improved such as enhancing the efficacy of Human-in-the-loop Systems for Machine Learning in detecting phishing attacks in HTML URLs lies in refining feature extraction methodologies. Feature extraction plays an important role in distinguishing between legitimate and malicious URLs by reducing the dimensionality of input data and extracting relevant features. While traditional machine learning approaches rely on human intervention for feature extraction, deep learning models such as convolutional neural networks (CNNs) have shown promise in automating this process by learning from input data and labels to extract pertinent features [27].

Researchers could explore the application of CNNs for feature extraction in phishing detection systems. Using CNNs to identify subtle patterns and distinguishing characteristics in HTML URLs, detection models could become more adept at identifying phishing attempts with higher accuracy[27]. Additionally, integrating human feedback mechanisms into the training process could further refine feature extraction algorithms, enhancing the adaptability and robustness of detection systems against evolving phishing techniques . Collaboration among researchers, industry stakeholders, and cybersecurity professionals is essential to foster innovation and develop robust solutions to address the constantly changing landscape of phishing attacks.

REFERENCES

[1] Alkhalil, Z., Hewage, C., Nawaf, L., & Khan, I. (2021). Phishing attacks: A recent comprehensive study and a new anatomy. Frontiers in Computer Science, 3, 563060.

[2] Gupta, B. B., Tewari, A., Jain, A. K., & Agrawal, D. P. (2017). Fighting against phishing attacks: state of the art and future challenges. Neural Computing and Applications, 28, 3629-3654.

[3] Alabdan, R. (2020). Phishing attacks survey: Types, vectors, and technical approaches. Future internet, 12(10), 168.

[4] Kabachinski, J. (2002). Surfing the Web: http, URLs, and HTML. Biomedical instrumentation & technology, 36(1), 49-52.

[5] Aljofey, A., Jiang, Q., Rasool, A., Chen, H., Liu, W., Qu, Q., & Wang, Y. (2022). An effective detection approach for phishing websites using URL and HTML features. Scientific Reports, 12(1), 8842.

[6] Li, Y., Yang, Z., Chen, X., Yuan, H., & Liu, W. (2019). A stacking model using URL and HTML features for phishing webpage detection. Future Generation Computer Systems, 94, 27-39.

[7] Alpaydin, E. (2021). Machine learning. MIT press.

[8] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.

[9] Mahesh, B. (2020). Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet], 9(1), 381-386.

[10] Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., & He, L. (2022). A survey of human-in-the-loop for machine learning. Future Generation Computer Systems, 135, 364-381.

[11] Zanzotto, F. M. (2019). Human-in-the-loop artificial intelligence. Journal of Artificial Intelligence Research, 64, 243-252.

[12] Agnisarman, S., Lopes, S., Madathil, K. C., Piratla, K., & Gramopadhye, A. (2019). A survey of automation-enabled human-in-the-loop systems for infrastructure visual inspection. Automation in Construction, 97, 52-76.

[13] Sarker, I. H., Kayes, A. S. M., Badsha, S., Alqahtani, H., Watters, P., & Ng, A. (2020). Cybersecurity data science: an overview from machine learning perspective. Journal of Big data, 7, 1-29.

[14] Kemmerer, R. A. (2003, May). Cybersecurity. In 25th International Conference on Software Engineering, 2003. Proceedings. (pp. 705-715). IEEE.

[15] Jang-Jaccard, J., & Nepal, S. (2014). A survey of emerging threats in cybersecurity. Journal of computer and system sciences, 80(5), 973-993.

[16] Bhuyan, M. H., Bhattacharyya, D. K., & Kalita, J. K. (2013). Network anomaly detection: methods, systems and tools. Ieee communications surveys & tutorials, 16(1), 303-336.

[17] Vinod, P., Jaipur, R., Laxmi, V., & Gaur, M. (2009, March). Survey on malware detection methods. In Proceedings of the 3rd Hackers' Workshop on computer and internet security (IITKHACK'09) (pp. 74-79).

[18] Salahdine, F., & Kaabouch, N. (2019). Social engineering attacks: A survey. Future internet, 11(4), 89.

[19] Syafitri, W., Shukur, Z., Asma'Mokhtar, U., Sulaiman, R., & Ibrahim, M. A. (2022). Social engineering attacks prevention: A systematic literature review. IEEE access, 10, 39325-39343.

[20] Gupta, S., Singhal, A., & Kapoor, A. (2016, April). A literature survey on social engineering attacks: Phishing attack. In 2016 international conference on computing, communication and automation (ICCCA) (pp. 537-540). IEEE.

[21] Bécue, A., Praça, I., & Gama, J. (2021). Artificial intelligence, cyber-threats and Industry 4.0: Challenges and opportunities. Artificial Intelligence Review, 54(5), 3849-3886.

[22] Ghelani, D. (2022). Cyber security, cyber threats, implications and future perspectives: A Review. Authorea Preprints.

[23] Ahsan, M., Nygard, K. E., Gomes, R., Chowdhury, M. M., Rifat, N., & Connolly, J. F. (2022). Cybersecurity threats and their mitigation approaches using Machine Learning—A Review. Journal of Cybersecurity and Privacy, 2(3), 527-555.

[24] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in neural information processing systems, 35, 27730-27744.

[25] Griffith, S., Subramanian, K., Scholz, J., Isbell, C. L., & Thomaz, A. L. (2013). Policy shaping: Integrating human feedback with reinforcement learning. Advances in neural information processing systems, 26.

[26] Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. Advances in neural information processing systems, 30.

[27] Aburomman, A. A., & Reaz, M. B. I. (2017). A survey of intrusion detection systems based on ensemble and hybrid classifiers. Computers & security, 65, 135-152.

[28] Peddabachigari, S., Abraham, A., Grosan, C., & Thomas, J. (2007). Modeling intrusion detection system using hybrid intelligent systems. Journal of network and computer applications, 30(1), 114-132.

[29] Garg, A., & Maheshwari, P. (2016, January). A hybrid intrusion detection system: A review. In 2016 10th International Conference on Intelligent Systems and Control (ISCO) (pp. 1-5). IEEE.