ISSN: 2583-7346
DOI: https://doi-ds.org/doilink/05.2024-48867725/IJAEAST.2024.02.0004, February 2024, Volume-2, Issue-2, pp.69-74
International Journal of Advances in Engineering Architecture Science and Technology

A Smart Solution for Automated Phishing URL Detection Using Machine Learning

G. Divya[1], D. Evans David[2], B. Guna Sekhar[3], K. Suman[4], G. Chandrika[5]

[1,2,3,4]Student, Department of Computer Science & Engineering, Dadi Institute of Engineering and Technology (Autonomous), Andhra Pradesh, India

[5]Assistant Professor, Department of Computer Science & Engineering, Dadi Institute of Engineering and Technology (Autonomous), Andhra Pradesh, India

divyagonthina3@gmail.com[1], evansdavid1803@gmail.com[2], buddhagunasekhar2233@gmail.com[3], sumankonthala2002@gmail.com[4]

Received: 23-01-2024 **Accepted:** 25-02-2024 **Published:** 27-02-2024

Abstract

Background: The increasing integration of the Internet into our social lives has brought about a significant shift in how people learn and work, simultaneously exposing us to a rising threat of serious security attacks.
Objectives: Recognizing various network threats, especially novel attacks, has become a pressing issue that demands immediate attention. Phishing site URLs, with the goal of harvesting private information such as user identities, passwords, and online financial transactions, pose a significant risk.
Statistical Analysis: Phishers often create sites that closely mimic the appearance and semantics of legitimate websites, taking advantage of users who access government and financial services online. Consequently, there has been a notable surge in phishing threats and attacks over the past few years.
Findings: As technology evolves, phishing methods are advancing rapidly, necessitating the adoption of anti-phishing techniques to effectively detect and counter such attacks. This project focuses on the implementation of a system that achieves these objectives, employing four machine learning supervised classification models: K-Nearest Neighbor, Kernel Support Vector Machine, Decision Tree, and Random Forest Classifier.
Applications and Improvements: Through experimentation, it was determined that the Random Forest Classifier outperforms the others, providing the highest accuracy for the selected dataset, with an impressive accuracy score of 96.82%.

Keywords: Phishing attack, Semantic analysis methods, Database, Random Forest, Machine Learning.

1. Introduction

Phishing remains a highly dangerous criminal activity within the realm of cyberspace. The increasing reliance of users on online services provided by government and financial institutions has led to a significant uptick in phishing attacks over recent years. Phishers have transformed this illicit practice into a successful business venture, utilizing diverse methods to target vulnerable users, including messaging, VOIP, spoofed links, and counterfeit websites.The

ISSN: 2583-7346
DOI: https://doi-ds.org/doilink/05.2024-48867725/IJAEAST.2024.02.0004, February 2024, Volume-2, Issue-2, pp.69-74
International Journal of Advances in Engineering Architecture Science and Technology

creation of counterfeit websites, closely mimicking genuine layouts and content, has become a prevalent strategy employed by phishers. These deceptive sites mirror the appearance and information of legitimate platforms, making it challenging for users to discern between the authentic and fraudulent. The primary objective behind these counterfeit websites is to extract sensitive user data, such as account numbers, login credentials, and debit/credit card passwords. Additionally, attackers employ social engineering tactics, masquerading as high-level security measures and prompting users to respond to security questions, leading them to unwittingly disclose information and fall victim to phishing attacks.

2. Literature Review

Numerous scholars have conducted analyses on the statistics of phishing URLs, inspiring our current approach. Happy emphasizes phishing as a perilous method for hackers to clandestinely acquire users' account information. Users, often unaware of such traps, become victims of phishing scams due to a lack of financial aid, personal experience, market awareness, or brand trust. Mehmet et al. proposed a URL-based phishing detection method, employing eight different algorithms on three distinct datasets using various machine learning methods and hierarchical architectures.

Garera et al. employed logistic regression to classify phishing URLs, considering red flag keywords, Google's web page features, and Page Rank quality recommendations. Although direct comparisons are challenging due to differing datasets and features, our approach utilizes machine learning techniques to analyze URL and website properties.

Yong et al. introduced a novel approach to detect phishing websites, focusing on a URL-centric methodology. Their capsule-based neural network involves parallel components, removing shallow URL characteristics and constructing accurate feature representations. Their system competes effectively with cutting-edge detection methods.

Vahid Shahrivari et al. utilized machine learning for phishing detection, favouring the random forest algorithm for its accuracy. incorporated NLP tools for improved results, achieving high accuracy with Support Vector Machine.

Amani Alswailem et al. experimented with various machine learning models, achieving optimal accuracy with the random forest algorithm.

Hossein et al. developed the "Fresh-Phish" open-source framework, utilizing machine learning classifiers on a labelled dataset. The study by X. Zhang suggested a phishing detection model based on mining semantic characteristics, achieving successful results in Chinese web pages.

M. Aydin proposed a versatile and straightforward framework for extracting characteristics, utilizing data from Phish Tank and authentic URLs from Google. Feature selection methods and performance evaluation using Nave Bayes and Sequential Minimal Optimization were conducted, with SMO preferred for phishing detection.

In summary, these studies contribute diverse methodologies and insights into phishing detection, employing machine learning, neural networks, and innovative approaches to enhance accuracy and efficiency.

3. Methodology

A phishing website, employing social engineering techniques, replicates authentic webpages and Uniform Resource Locators (URLs). The URL, being the primary conduit for phishing attacks, offers phishers complete control over sub-domains, enabling manipulation through file components and directories. In this study, we adopted the linear-sequential model, commonly recognized as the waterfall model, to delineate our methodology. While the waterfall model is

ISSN: 2583-7346
DOI: https://doi-ds.org/doilink/05.2024-48867725/IJAEAST.2024.02.0004, February 2024, Volume-2, Issue-2, pp.69-74
International Journal of Advances in Engineering Architecture Science and Technology

considered conventional, it proves effective in scenarios with limited requirements. Our approach involved segmenting the application into smaller components, developed using both frameworks and hand-crafted code. This division facilitated a systematic and structured progression throughout the development process.

Data Collection

Phishing URLs were systematically gathered using the open-source tool Phish Tank. This platform offers a diverse range of phishing URLs in various formats, such as csv, json, and others, regularly updated on an hourly basis. The dataset derived from Phish Tank consists of 5000 randomly selected phishing URLs, serving as the foundation for training machine learning models.

Data Cleaning

To enhance the quality of the dataset, a rigorous data cleaning process was implemented. This involved tasks such as filling in missing values, smoothing out erratic data points, identifying and removing outliers, and rectifying anomalies. The objective was to ensure a refined and reliable dataset for subsequent analyses.

Data Pre-processing

Data pre-processing, a crucial step, involved transforming the raw, unstructured data into a well-organized and structured dataset. This prepares the data for further research and analysis by addressing irregularities and inconsistencies.

Data Splitting

The dataset was divided into 8000 training samples and 2000 testing samples. This segregation aimed to facilitate the training of the machine learning model effectively. Notably, the nature of this dataset indicates a supervised machine learning problem, specifically a classification problem. Given that the input URLs are categorized as legitimate or phishing, the focus is on classification.

Supervised Machine Learning Models

Several supervised machine learning models were evaluated for training on this dataset. The models considered included Decision Tree, Multilayer Perceptron, Random Forest, Autoencoder Neural Network, XGBoost, and Support Vector Machines. Each of these models underwent scrutiny to determine their efficacy in addressing the classification task posed by the dataset.

4. Structure of an URL

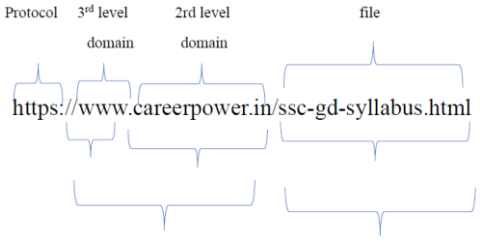

Figure 1. Structure of URL

ISSN: 2583-7346
DOI: https://doi-ds.org/doilink/05.2024-48867725/IJAEAST.2024.02.0004, February 2024, Volume-2, Issue-2, pp.69-74
International Journal of Advances in Engineering Architecture Science and Technology

Feature Extraction

Classification output: 0 = legitimate, 1 phishing
1. having IP Address 2. URL Length 3. Shortening Service 4. having At Symbol 5. double slash redirecting 6. Prefix Suffix 7. having Sub Domain 8. SSL State 9. Domain registration length 10. Favicon 11. Open ports 12. HTTPS token in_URL 13. Request URL 14. URL of Anchor 15. Links_in_tags 16. Server Form Handler 17. Submitting to_email 18. Abnormal URL 19. Site Redirect 20. on mouseover_changes 21. DNS Record 22. web traffic rank 23. Page Rank 24. RightClick Disabled 25. popUpWindow 26. Iframe redirection phishing domains 27. Google Index 28. Links pointing to page 29. Statistical report-top.

5. Accuracy Score

The figure 1 is a comparative plot that compares the accuracy of the four algorithms namely, Random Forest, Kernel SVM, KNN, Decision tree

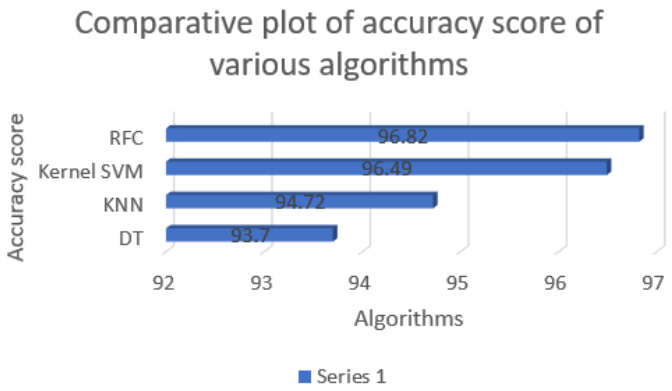

Figure 2. Comparative Plot of Accuracy Scores

Flow Diagram

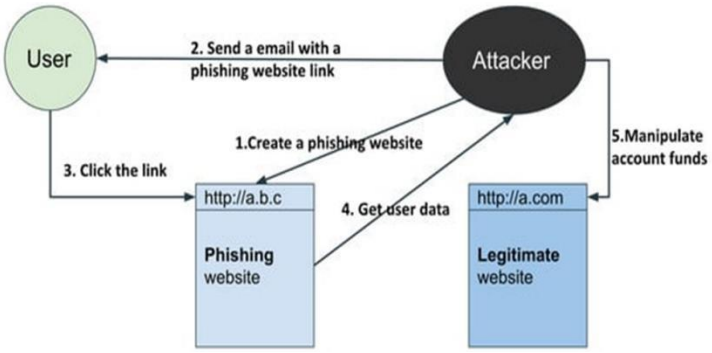

Figure 3. Flow Diagram

The phishing URL detection process using machine learning begins with the collection of diverse phishing URLs from sources like Phish Tank, ensuring regular updates for dataset relevance. Subsequently, data cleaning addresses missing values, outliers, and anomalies, enhancing dataset quality. Data pre-processing transforms raw data into a structured format. After splitting the dataset, various supervised machine learning models, including Decision Tree, Multilayer Perceptron, Random Forest, Autoencoder Neural Network, XGBoost, and

ISSN: 2583-7346
DOI: https://doi-ds.org/doilink/05.2024-48867725/IJAEAST.2024.02.0004, February 2024, Volume-2, Issue-2, pp.69-74
International Journal of Advances in Engineering Architecture Science and Technology

Support Vector Machines, are trained on the data. Model performance is evaluated using testing data, and the most effective model is selected based on metrics like accuracy and precision.

6. Result

The phishing URL detection process employing machine learning yields promising results through a systematic approach. Commencing with the collection of phishing URLs from Phish Tank and ensuring regular updates, the dataset undergoes thorough cleaning, addressing missing values, outliers, and anomalies. Data pre-processing transforms the raw information into a structured format, facilitating effective machine learning model training. Splitting the dataset into training and testing samples, various supervised models, including Decision Tree, Multilayer Perceptron, Random Forest, Autoencoder Neural Network, XGBoost, and Support Vector Machines, are evaluated. The selected model, based on metrics such as accuracy and precision, is deployed for real-time or batch processing, demonstrating robust capabilities in identifying phishing threats. Continuous monitoring, updates, and a feedback loop contribute to ongoing refinement, enhancing the model's accuracy and adaptability to emerging phishing techniques.

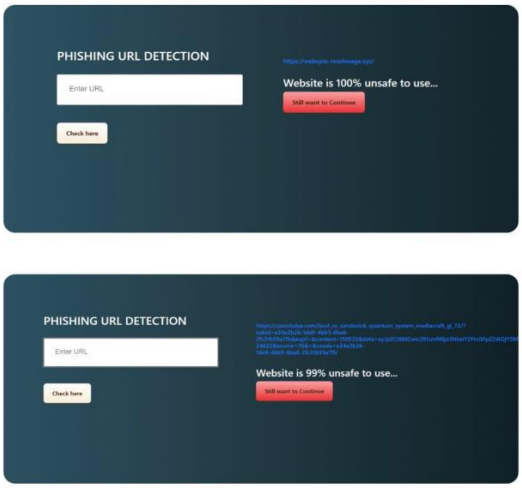

Figure 4. Final Result

7. Conclusion

Phishing URL detection employing machine learning techniques has proven to be a robust and effective approach in safeguarding users against cyber threats. The utilization of datasets sourced from platforms like Phish Tank, coupled with regular updates, ensures the model's relevance in the ever-evolving landscape of phishing attacks. Rigorous data cleaning and preprocessing contribute to the creation of a high-quality dataset, optimizing the performance of supervised machine learning models. The evaluation of various models, including Decision Tree, Multilayer Perceptron, Random Forest, Autoencoder Neural Network, XGBoost, and Support Vector Machines, enables the selection of the most effective tool for the task. Deployment of the chosen model demonstrates its prowess in real-time or batch processing, showcasing its ability to accurately identify phishing.

8. Future Scope

The future scope of phishing URL detection using machine learning holds immense promise in enhancing cybersecurity. By leveraging advanced algorithms and continuous learning, machine learning models can adapt to evolving phishing techniques, providing more robust and accurate detection of malicious URLs. This approach enables proactive identification of threats, reducing the risk of successful phishing attacks and bolstering overall digital security. As cyber threats continue to evolve, the integration of machine learning in phishing URL detection represents a crucial advancement in staying ahead of malicious actors and safeguarding sensitive information in the digital landscape.

Acknowledgments

The authors express their deep sense of gratitude to the Chairman, Dadi institute of Engineering & Technology Sri Dadi Ratanakar for the facilities provide in carrying out this work successfully.

References

1. Reid G. Smith and Joshua Eckroth delve into the evolution and future prospects of AI applications in their article "Building AI Applications: Yesterday, Today, and Tomorrow," published in AI Magazine in March 2017.
2. Panos Louridas and Christof Ebert explore machine learning in their IEEE Software article titled "Machine Learning," offering insights into the field's current state and developments in September 2016.
3. Michael Jordan and T.M. Mitchell provide an overview of machine learning trends, perspectives, and prospects in their science article published in July 2015.
4. Steven Aftergood discusses the online Cold War in the realm of cybersecurity in his article "Cybersecurity: The Cold War Online," featured in Nature in July 2017.
5. Aleksandar Milenkoski, Marco Vieira, Samuel Kounev, Alberto Avritzer, and Bryan Payne survey common practices in evaluating computer intrusion detection systems in their ACM Computing Surveys article from September 2015.
6. Chirag N. Modi and Kamatchi Acha conduct a comprehensive review of security challenges in the virtualization layer and intrusion detection/prevention systems in cloud computing in their article published in The Journal of Supercomputing in March 2017.
7. Eduardo Viegas et al. work towards an energy-efficient anomaly-based intrusion detection engine for embedded systems in their article published in IEEE Transactions on Computers in January 2016.
8. Y. Xin et al. explore machine learning and deep learning methods for cybersecurity in their article "Machine Learning and Deep Learning Methods for Cybersecurity," published in IEEE Access in 2018.
9. Neha R. Israni and Anil N. Jaiswal conduct a survey on various phishing and anti-phishing measures in their article published in the International Journal of Engineering Research and Technology in 2015.
10. Pingchuan Liu and Teng-Sheng Moh focus on content-based spam email filtering in their article from October 2016, addressing techniques to combat unwanted email content.