e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

AN INTELLIGENT IDENTIFICATION AND CLASSIFICATION SYSTEM FOR PHISHING UNIFORM RESOURCE LOCATORS (URLS)

B. Nithesh Kumar*1, N. Keerthana*2, S. Kevin Andrews*3

*1Student, Department Of Computer Application, Dr. M.G.R. Educational And Research Institute, Chennai, Tamil Nadu, India.

*2Associate Professor, Department Of Computer Application, Dr. M.G.R. Educational And Research Institute, Chennai, Tamil Nadu, India.

*3Professor, Department Of Computer Application, Dr. M.G.R. Educational And Research Institute, Chennai, Tamil Nadu, India.

DOI : https://www.doi.org/10.56726/IRJMETS51921

ABSTRACT

Cybersecurity is an issue of most significance today, and it is important to detect and mitigate such phishing Uniform Resource Locators (URLs). One approach for detecting phishing URLs using machine learning techniques is presented in this research. The proposed system uses a combination of benign and phishing URLs as a dataset to train various machine learning models. Feature engineering can be used to extract important attributes from URLs, while different algorithms are combined in order to find the best model. To assess how efficacious the suggested solution is, this study measures its performance by means of accuracy, precision, recall, and F1 score. As illustrated by the findings, the authors have successfully applied the machine-learning-based method for distinguishing between safe and fake URLs, which can be utilized as valuable tools leading to better online security measures. This article offers an intelligent algorithm that detects phishing URLs, thereby contributing to strengthening the cybersecurity framework in use presently.

Keywords: Cybersecurity, URL, Phishing URL Detection, Online Threats, Feature Engineering.

I. INTRODUCTION

In the contemporary landscape of evolving online threats and the incessant surge in cyberattacks, safeguarding digital environments has become an imperative task. A significant facet of this defense strategy involves the early identification and mitigation of phishing Uniform Resource Locators (URLs), which serve as gateways for cyber threats. Phishing URLs often disguise themselves within the vast expanse of the internet, necessitating advanced and adaptive detection mechanisms.

This research introduces a comprehensive approach to address the challenges posed by phishing URLs through the application of machine learning techniques. Leveraging the power of data-driven models, we present a system that utilizes a carefully curated dataset containing both benign and phishing URLs. The primary objective is to train and evaluate various machine learning models to effectively discern between harmless and phishing web addresses.

Our methodology encompasses feature engineering, a critical aspect in extracting relevant characteristics from URLs, which serves as the foundation for model training. This study conducts an in-depth evaluation of several machine learning methods to figure out the best effective approach to stable and precise recognition. The performance evaluation of the system is carried out using key metrics such as accuracy, precision, recall, and F1 score, providing nuanced insights into the system's efficacy.

Beyond the quantitative assessment, this study delves into the qualitative aspects of the proposed solution, emphasizing its robustness in handling the dynamic nature of phishing URL patterns. The results obtained not only showcase the effectiveness of the machine learning-based approach in distinguishing between benign and phishing URLs but also highlight its potential as a valuable and intelligent tool for enhancing online security measures.

By contributing a novel and adaptive solution for the proactive identification of phishing URLs, this research aligns with the ongoing efforts to fortify cybersecurity infrastructure. As cyber threats continue to evolve, our intelligent system demonstrates a forward-looking approach, empowering organizations and individuals with an advanced defense mechanism against the ever-growing sophistication of online threats.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

II. LITRERATURE SURVEY

[1] TITLE: Intelligent Techniques for Detecting Network Attacks: Review and Research Directions

AUTHOR: Hanan S.Atamimi

DESCRIPTION: The significant growth in the use of the Internet and the rapid development of network technologies are associated with an increased risk of network attacks. Network attacks refer to all types of unauthorized access to a network including any attempts to damage and disrupt the network, often leading to serious consequences. Network attack detection is an active area of research in the community of cybersecurity. In the literature, there are various descriptions of network attack detection systems involving various intelligent-based techniques including machine learning (ML) and deep learning (DL) models. However, although such techniques have proved useful within specific domains, no technique has proved useful in mitigating all kinds of network attacks. This is because some intelligent-based approaches lack essential capabilities that render them reliable systems that are able to confront different types of network attacks. This was the main motivation behind this research, which evaluates contemporary intelligent-based research directions to address the gap that still exists in the field. The main components of any intelligent-based system are the training datasets, the algorithms, and the evaluation metrics; these were the main benchmark criteria used to assess the intelligent-based systems included in this research article. This research provides a rich source of references for scholars seeking to determine their scope of research in this field. Furthermore, although the paper does present a set of suggestions about future inductive directions, it leaves the reader free to derive additional insights about how to develop intelligent-based systems to counter current and future network attacks.

[2] TITLE: Phishing Attacks Detection using Machine Learning and Deep Learning Models

AUTHOR: Malak Aljabri

DESCRIPTION: Because of the fast expansion of internet users, phishing attacks have become a significant menace where the attacker poses as a trusted entity in order to steal sensitive data, causing reputational damage, loss of money, ransomware, or other malware infections. Intelligent techniques mainly Machine Learning (ML) and Deep Learning (D L) are increasingly applied in the field of cybersecurity due to their ability to learn from available data in order to extract useful insight and predict future events. The effectiveness of applying such intelligent approaches in detecting phishing web sites is investigated in this paper. We used two separate datasets and selected the highest correlated features which comprised of a combination of content-based, URL lexical-based, and domain-based features. A set of ML models were then applied, and a comparative performance evaluation was conducted. Results proved the importance of features selection in improving the models' performance. Furthermore, the results also aimed to identify the best features that influence the model in identifying phishing websites. For classification performance, Random Forest (RF) algorithm achieved the highest accuracy for both datasets.

[3] TITLE: Phishing Websites Detection Based on Hybrid Model of Deep Belief Network and Support Vector Machine

AUTHOR: Xuqiao Yu

DESCRIPTION: The boosting of financial crimes that employ technical methods has become a critical issue that is urgent to be solved. However, the performance of most of the traditional classification methods are dependent on the quality of the prior knowledge of features. To address these problems, this paper proposed a hybrid model that combines the advantages of deep learning neural network of Deep Belief Network and machine learning method of Support Vector Machines. Firstly, the unidentified URLs from blacklist filtering are processed to have the URLs features extracted, the features are including statistical features, webpage code features and webpage text features. Secondly, deep features are extracted by the quick classification of deep learning model. Lastly, the resulting feature vectors combining with URL statistical features, webpage code features, webpage text features are fed into SVM model for classification. The model was tested on a dataset containing millions of phishing URLs and legitimate URLs, and have achieved the accuracy of 99.96%, the precision rate of 99.94% and the false positive rate of 51.32% which showed better performance than other comparison models.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

[4] TITLE: A heuristic technique to detect phishing website using TWSVM classifier

AUTHOR: Routhu Srinivasa Rao

DESCRIPTION: Phishing websites are on the rise and are hosted on compromised domains such that legitimate behavior is embedded into the designed phishing site to overcome the detection. The traditional heuristic techniques using HTTPS, search engine, Page Ranking and WHOIS information may fail in detecting phishing sites hosted on the compromised domain. Moreover, list-based techniques fail to detect phishing sites when the target website is not in the whitelisted data. In this paper, we propose a novel heuristic technique using TWSVM to detect malicious registered phishing sites and also sites which are hosted on compromised servers, to overcome the aforementioned limitations. Our technique detects the phishing websites hosted on compromised domains by comparing the log-in page and home page of the visiting website. The hyperlink and URL-based features are used to detect phishing sites which are maliciously registered. We have used different versions of support vector machines (SVMs) for the classification of phishing websites. We found that twin support vector machine classifier (TWSVM) outperformed the other versions with a significant accuracy of 98.05% and recall of 98.33%.

III. PROBLEM STATEMENT

EXISTING SYSTEM:

Traditional security measures heavily rely on signature-based methods to detect known phishing URLs, revealing inherent limitations. Their effectiveness diminishes against novel or zero-day threats, as they lack adaptability. Vulnerable to evasion techniques, these methods falter when attackers modify URLs to bypass existing signatures. Furthermore, their inefficiency in handling large-scale datasets results in slower detection rates. Dependency on manual rule creation and updates introduces delays in response, hindering real-time threat mitigation. The struggle to detect polymorphic URLs further compounds the issue, potentially leading to high false positives. The shortcomings of signature-based approaches underscore the need for more adaptive and scalable strategies in the ever-evolving landscape of cybersecurity.

DISADVANTAGES OF EXISTINTG SYSTEM:

- ✓ Signature-based methods, reliant on known patterns, are less effective against novel or zero-day threats due to their lack of adaptability.
- ✓ Manual rule creation and updates in signature-based methods cause delays in responding to emerging threats, impeding real-time protection against evolving cyber threats.
- ✓ Efficiency issues in processing large datasets with signature-based methods result in slower detection rates and potential scalability challenges as data volume increases.

IV. PROPOSED SYSTEM

- ✓ Gather a labeled dataset of URLs, where each URL is classified as either benign or malicious. This dataset should be diverse and representative of the types of URLs the system will encounter in the real world.
- ✓ Extract relevant features from the URLs. Potential features might include URL length, domain age, presence of special characters, use of IP addresses, and frequency of certain words or patterns.
- ✓ Choose a Decision Tree algorithm, these libraries offer efficient implementations of Gradient Boosting and are widely used in practice.
- ✓ Train the Gradient Boosting model on the preprocessed dataset. Use techniques like cross-validation to optimize hyperparameters and avoid overfitting.
- ✓ Deploy the trained Gradient Boosting model to a production environment. Integrate it into the cybersecurity infrastructure to analyze and classify URLs in real-time.
- ✓ Develop a user interface or integration for security analysts to interact with the system. Provide clear indications of why a URL is classified as malicious or benign.

ADVANTAGE OF PROPOSED SYSTEM:

- ✓ **Flexible Feature Extraction:** By extracting relevant features from the URLs, the system can adapt to new threats and variations in malicious URL patterns. Features such as URL length, domain age, and presence of special characters provide valuable insights into the nature of the URLs being analyzed.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

- ✓ **Robust Training Process:** Utilizing techniques like cross-validation ensures that the model is trained effectively on the dataset and prevents overfitting. This leads to a more robust and generalizable model that performs well on unseen data.
- ✓ **Real-Time Analysis:** Deploying the trained model to a production environment enables real-time analysis of URLs, allowing for rapid detection and response to potential threats. This enhances the cybersecurity infrastructure's ability to protect against malicious activities.

V. BLOCK DIAGRAM

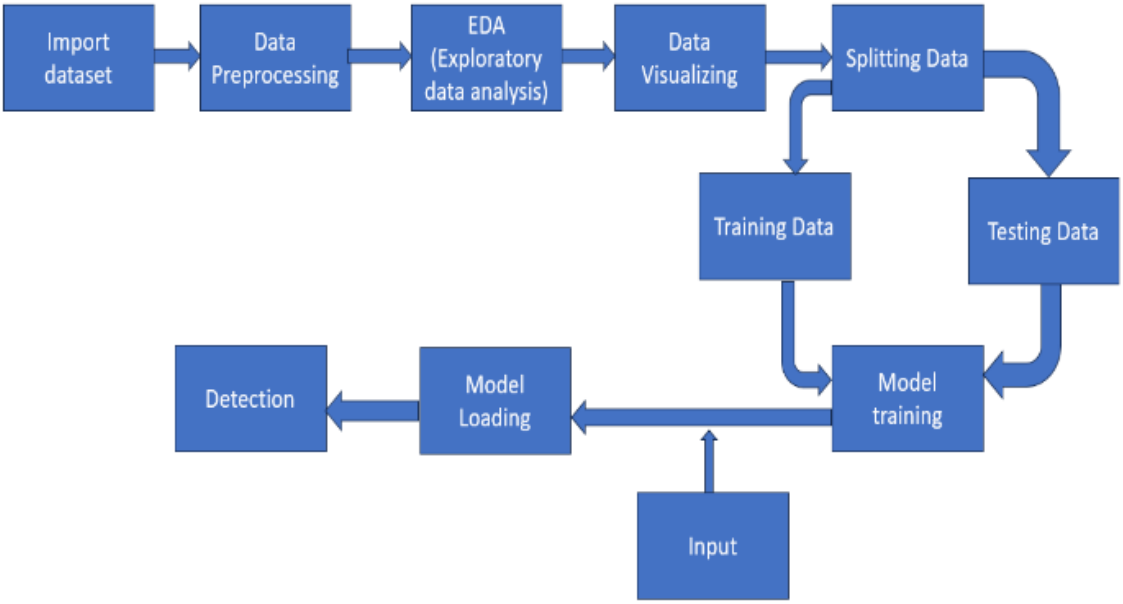

Figure 1: Block Diagram

ADVANTAGES OF BLOCK DIAGRAM:

- ✓ **Visualization of the Process:** A block diagram provides a visual representation of the entire process of phishing URL detection, allowing security analysts and researchers to understand how different components interact and influence each other.
- ✓ **Modularity:** A block diagram can help in breaking down the phishing URL detection process into modular components or stages. Each block represents a particular stage or component, such as URL extraction, feature extraction, machine learning models, and decision making.
- ✓ **Clarity:** Block diagrams help in clarifying complex processes by simplifying them into distinct stages or components. This can make it easier to explain and understand the phishing URL detection process, especially for non-technical stakeholders.
- ✓ **Analysis and Optimization:** A block diagram can be used to analyze and optimize the phishing URL detection process by identifying bottlenecks, redundant stages, or inefficiencies. It provides a high-level view of the process, making it easier to identify areas for improvement.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

VI. IMPLEMENTATION

ARCHITECTURE DIAGRAM:

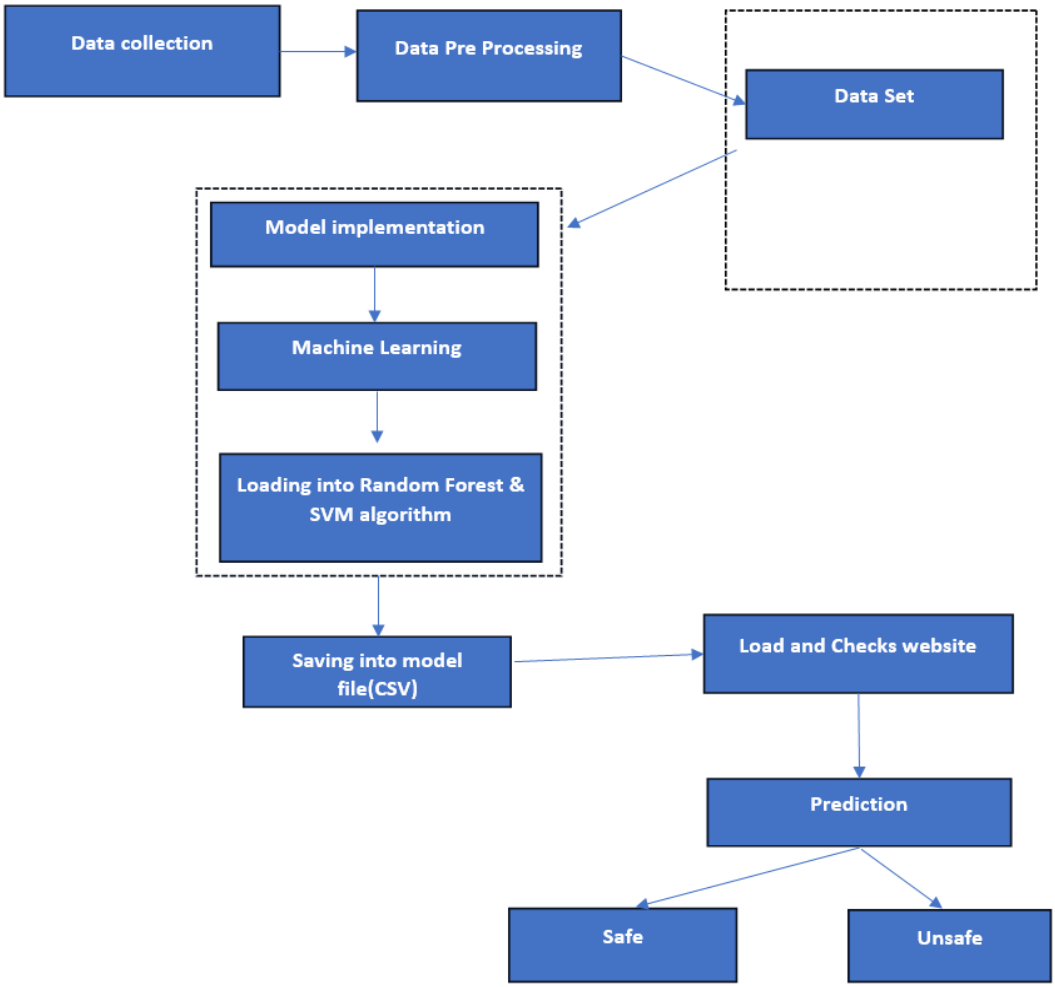

Figure 2: Architecture Diagram

SVM ALGORITHM:

An effective supervised machine learning method to solve regression and classification problems is called Support Vector Machine (SVM). Finding the hyperplane in the given input space that best differentiates the classes is the fundamental concept behind support vector machines (SVM).

TESTING:

Testing gives quantitative validation that the functionalities being tested are available in accordance with the technical and company specifications, software records, and user guides.

The following areas are the focus of the testing:

Valid Input: Recognized valid input classes need to be approved.

Invalid Input: identified classes of incorrect input is necessary.

Functions: It is necessary to perform the identified functions.

Output: It is necessary to exercise the designated types of application outputs.

Procedures: Invoking procedures or integrating systems is necessary. Test preparation and organization are centered on requirements, critical features, or unique test cases. Furthermore, testing needs to take into account data fields, specified procedures, sequential processes, and systematic coverage related to identifying business process flows. Additional tests are identified prior to testing completion, and the efficaciousness of the current tests is evaluated.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

VII. RESULT

After the implementation of testing model the machine learning algorithm is ready to execute the result. In the execution process of result, it loads the trained data-set from the particular location, which it stored and it compiled with the user given input data. This model perform by an machine learning algorithm, which it represent the user given URL (input) is safe or unsafe at the final stage of result execution.

VIII. CONCLUSION

The escalating digital storage of personal information on mobile devices underscores the critical need for robust cybersecurity measures, particularly in detecting phishing URLs. The prevalence of phishing attacks, leading to substantial data and financial losses, necessitates effective detection mechanisms. This project employs supervised learning algorithms, specifically random forest and SVM, revealing that Random forest achieves a commendable 90.61% accuracy in identifying phishing URLs. The research emphasizes the significance of URL attributes in distinguishing between phishing and benign entities, highlighting specific characteristics crucial for encryption and disguise. The visualization of relationships between these attributes provides valuable insights into patterns associated with phishing URLs, contributing to the ongoing efforts to enhance cybersecurity in the digital era.

IX. FUTURE ENHANCEMENT

Dynamic Feature Selection:

Investigate the effectiveness of dynamic feature selection techniques in improving model performance. This could involve methods such as recursive feature elimination or feature importance ranking during model training.

Ensemble Methods:

Explore the integration of ensemble methods such as Random Forest or XGBoost alongside Gradient Boosting to further enhance the classification accuracy and robustness of the model.

Deep Learning Architectures:

Investigate the applicability and performance of deep learning architectures, such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs), in URL classification tasks. Deep learning models might capture more complex patterns and relationships within URLs.

X. REFERENCES

[1] Patil, Dharmaraj R., and Jayantro B. Patil. "Malicious URLs detection using decision tree classifiers and majority voting technique." Cybernetics and Information Technologies 18, no. 1 (2018): 11-29.

[2] Vinayakumar, R., K. P. Soman, and Prabaharan Poornachandran. "Evaluating deep learning approaches to characterize and classify malicious URL's." Journal of Intelligent Fuzzy Systems 34.3 (2018): 1333-1343.

[3] Garera, Sujata, et al. "A framework for detection and measurement of phishing attacks." Proceedings of the 2007 ACM workshop on Recurring malcode. 2007.

[4] Mohammad, Rami M., Fadi Thabtah, and Lee McCluskey. "An assessment of features related to phishing websites using an automated technique." 2012 international conference for internet technology and secured transactions. IEEE, 2012.

[5] Zhiwang, Cen, Xu Jungang, and Sun Jian. "A multi-layer bloom filter for duplicated URL detection." 2010 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE). Vol. 1. IEEE, 2010.

[6] Khonji, Mahmoud, Youssef Iraqi, and Andrew Jones. "Phishing detection: a literature survey." IEEE Communications Surveys Tutorials 15.4 (2013): 2091-2121

[7] Sahoo, Doyen, Chenghao Liu, and Steven CH Hoi. "Malicious URL detection using machine learning: a survey. CoRR abs/1701.07179 (2017)." (2017).

[8] Vazhayil, Anu, R. Vinayakumar, and K. P. Soman. "Comparative study of the detection of malicious URLs using shallow and deep networks." 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT). IEEE, 2018

[9] Srinivasan, S., Vinayakumar, R., Arunachalam, A., Alazab, M., Soman, K. P. (2021). DURLD: Malicious URL detection using deep learningbased character level representations. Malware analysis using artificial intelligence and deep learning, 535-554

[10] Menon, R.R.K., Akhil Dev, R., Bhattathiri, S.G., "An insight into the relevance of word ordering for text data analysis." 2020 fourth international conference on computing methodologies and communication (ICCMC). IEEE, 2020

[11] Menon, R. R., Kaartik, J., Nambiar, E. K., TK, A. K., Kumar, A. (2020, June). Improving ranking in document based search systems.2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184) (pp. 914-921). IEEE.

[12] Darling, Michael, Greg Heileman, Gilad Gressel, Aravind Ashok, and Prabaharan Poornachandran. "A lexical approach for classifying malicious URLs." In 2015 international conference on high performance computing simulation (HPCS), pp. 195-202. IEEE, 2015.

[13] G. Chakraborty and T. T. Lin, "A URL address aware classification of malicious websites for online security during websurfing," 2017 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS), 2017, pp. 1-6, doi: 10.1109/ANTS.2017.8384155.

[14] J. Xu and H. Lan, "Darknet Web URL Detection without URL Content Leakage," 2021 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS), 2021, pp. 377-382, doi: 10.1109/TOCS53301.2021.9688967.

[15] B. Janet, A. Nikam and J. A. Kumar R, "Real Time Malicious URL Detection on twitch using Machine Learning," 2022 International Conference on Electronics and Renewable Systems (ICEARS), 2022, pp. 1185-1189, doi: 10.1109/ICEARS53579.2022.9751862.