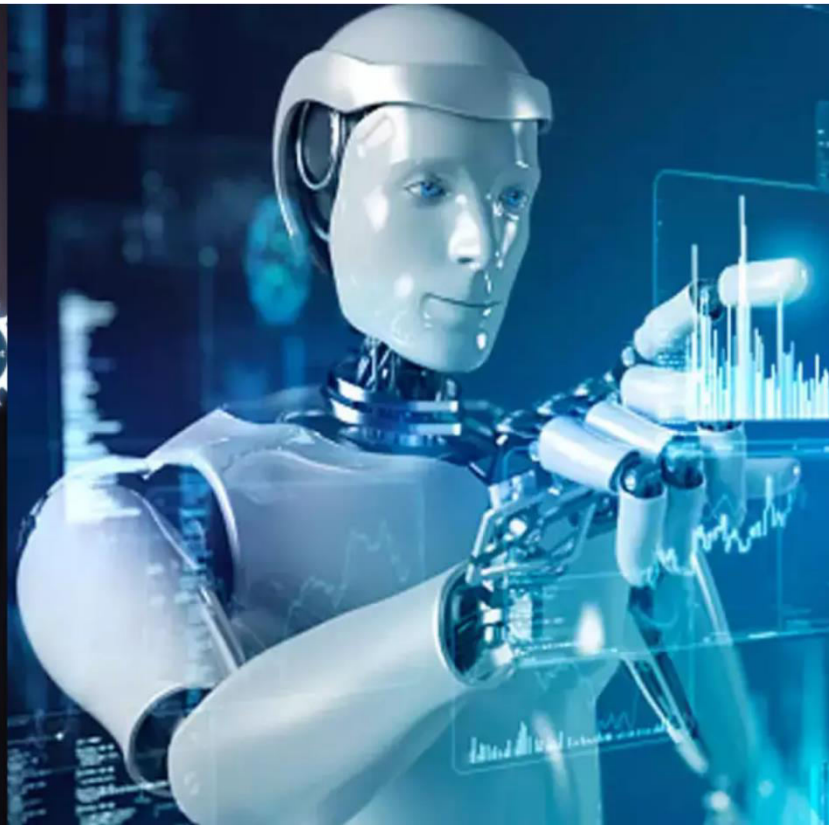
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 3, March 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403032|

Phishing Kit Attacks Dataset for Phishing Websites Identification

Mr.B.Arjun Balaji B.E., M.E.[1], K.Kameshwaran[2], A.Soundhra Perumal[3], S.Nandha Kumar[4], M.Ram Prasath[5]

Assistant Professor, Department of Computer Science and Design, Sethu Institute of Technology, Virudhunagar District, Tamil Nadu, India[1]

Department of Computer Science and Design, Sethu Institute of Technology, Virudhunagar District, Tamil Nadu, India[2,3,4,5]

ABSTRACT: Phishing attack is a simplest way to obtain sensitive information from innocent users. Aim of the phishers is to acquire critical information like username, password and bank account details. Cyber security persons are now looking for trustworthy and steady detection techniques for phishing websites detection. This paper deals with machine learning technology for detection of phishing URLs by extracting and analyzing various features of legitimate and phishing URLs. Decision Tree, random forest and Support vector machine algorithms are used to detect phishing websites. Aim of the paper is to detect phishing URLs as well as narrow down to best machine learning algorithm by comparing accuracy rate, false positive and false negative rate of each algorithm.

KEYWORDS: Phishing Detection , Machine Learning, URL Analysis, Decision Tree, Random Forest, Support Vector Machine, Cyber security, False Positive Rate, False Negative Rate, Phishing Websites

I. INTRODUCTION

In today's world, technology has become an integral part of the twenty-first century. The internet is one of these technologies, which is growing rapidly every year and plays an important role in individuals' lives. It has become a valuable and a convenient mechanism for supporting public transactions such as e-banking and e-commerce transactions. That has led the users to trust it is convenient to provide their private information to the Internet. As a result, the security thieves that have started to target this information have become a major security problem.

Phishing websites are considered to be one of these problems. They are using a social engineering trick, which can be described as fraudsters that try to manipulate the user into giving them their personal information based on exploiting human vulnerabilities rather than software vulnerabilities. Statistics have shown that the number of phishing attacks keeps increasing, which presents a security risk to the user information according to the Anti-Phishing Working Group (APWG) [1] and recorded phishing attacks by Kaspersky Lab [2], which stated that it has increased by 47.48% from all of the phishing attacks that have been detected during 2016. Recently, there have been several studies that tried to solve the phishing problem. Some researchers used the URL and compared it with existing blacklists that contain lists of malicious websites, which they have been creating, and there are others that have used the URL in an opposite manner, namely comparing the URL with a whitelist of legitimate websites. The latter approach uses heuristics, which uses a signature database of any known attacks that match the signature of the heuristic pattern to decide if it is a phishing website. Additionally, measuring website traffic using Alexa is another way that has been implemented by researchers to detect phishing websites.

Moreover, other researchers have used machine learning techniques. Machine learning is a field of computer science, which is also a branch of artificial intelligence (AI) that performs tasks and is capable of learning or acting in an intelligent way. It has two different types of learning: supervised learning and unsupervised learning. Supervised learning is based on training a model by giving it a set of measured features of data associated with a target label related to these data, and once the model is trained it can generate a new target label with unknown data. On the other hand, unsupervised learning is based on generating new data without giving any target label in the training process

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403032|

[22]. In this paper, the focus will be on the features combination that we get from Random Forest (RF) technique, as it has high accuracy, is relatively robust, and has a good performance.

II. LITERATURE SURVEY

Sahoo et al. (2020) presented a formal formulation of malicious URL detection as a machine learning task. Their study categorized and reviewed various feature representation techniques and algorithm designs used for phishing detection. Similarly, Wang et al. (2020) proposed a one-dimensional convolutional neural network (CNN) for encrypted traffic classification. Their end-to-end framework integrated feature extraction, selection, and classification, improving the detection of malicious traffic patterns.

Brar and Kumar (2020) conducted a comprehensive survey on cybersecurity threats and challenges. Their study classified recent incidents based on fundamental cybersecurity principles, providing insights into the evolving nature of cyber threats. In another study, Tan et al. (2021) introduced an eigenvalue assignment approach for attack detection in DC microgrids. Their model effectively identified attacks while minimizing the impact of load and voltage variations. Dawson and Thomson (2020) examined the future cybersecurity workforce, emphasizing the need for multidisciplinary skills beyond technical expertise. Their research highlighted the increasing complexity of cyber infrastructure and the growing number of vulnerable devices due to expanding connectivity.

These studies demonstrate the effectiveness of machine learning and AI-driven approaches in phishing detection and broader cybersecurity challenges. Future research can focus on refining detection techniques by incorporating deep learning, adaptive models, and real-time threat analysis to enhance accuracy and resilience against evolving cyber threats.

III. PROPOSED METHODOLOGY

The proposed system aims to develop an advanced Phishing URL Detection System using machine learning techniques. This system will classify URLs as legitimate or phishing based on extracted features and trained models, improving cybersecurity by identifying malicious websites before users interact with them. The methodology includes feature extraction, machine learning classification, model evaluation, and deployment for real-time detection. The solution will be implemented using Python, Scikit-learn, and deep learning frameworks, ensuring high accuracy and efficiency.

The architecture of the phishing detection system consists of four main stages: data preprocessing, feature extraction, model training, and classification. The system processes large datasets containing phishing and legitimate URLs, extracting key attributes that differentiate them. Lexical features such as URL length, presence of special characters, and domain structure are analyzed. Host-based features, including domain registration age and SSL certificate status, are considered for deeper security analysis. Content-based features, which involve examining the webpage's HTML structure, embedded scripts, and links, further enhance the detection accuracy.

For classification, multiple machine learning models will be employed, including Decision Trees, Random Forest, Support Vector Machines (SVM), and Neural Networks. These models are trained using labeled datasets of phishing and legitimate URLs. The system utilizes supervised learning, where historical data helps the model recognize phishing patterns. Additionally, ensemble learning techniques will be applied to enhance classification performance by combining the strengths of multiple models. The best-performing model will be selected based on evaluation metrics such as accuracy, precision, recall, and F1-score.

To ensure real-time detection and minimize false positives, the system incorporates adaptive thresholding and feature selection techniques. Gradient-based feature selection is used to prioritize the most significant attributes that contribute to phishing detection, improving model efficiency. Additionally, real-time URL analysis APIs can be integrated to verify a website’s credibility dynamically. The system will also include an automated retraining mechanism, where models are periodically updated with new phishing attack patterns to maintain effectiveness against evolving threats.

For real-world implementation, the phishing detection system will be deployed as a browser extension, email security filter, and API for cybersecurity applications. A simple Graphical User Interface (GUI) will display the classification results, showing users whether a URL is safe or suspicious. Future enhancements may include deep learning-based

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403032|

phishing detection, integration with AI-powered cybersecurity tools, and real-time phishing prevention mechanisms that alert users before clicking malicious links. By leveraging advanced machine learning techniques, the proposed system aims to provide an efficient and scalable solution for combating phishing threats in the digital world.

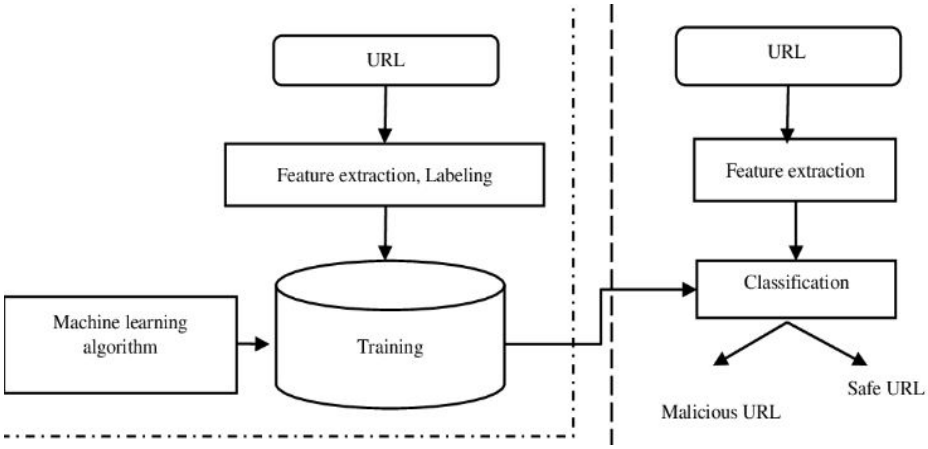


Fig1: Architecture Diagram

IV. RESULT AND DISCUSSION

Upload URL

In this module, if the user wants to classify the URL means they first give the URL. Based on the trained dataset, user can view the result. The result is categories as spam, malware, phishing, defacement.

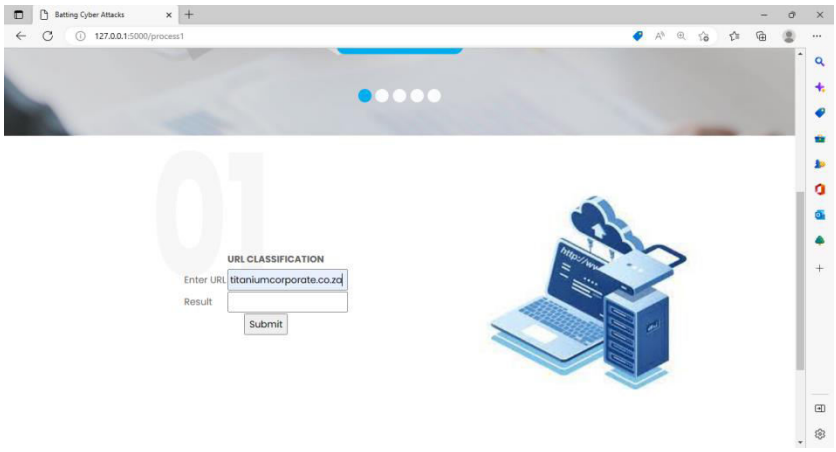


Fig 2: Upload Single URL

Prediction

In this module, if the user want to predict the URL based on three machine learning algorithms. Namely, XGBoost, Random Forest, LGB classifier. This prediction result having precision, recall, f1-score and support values.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403032|

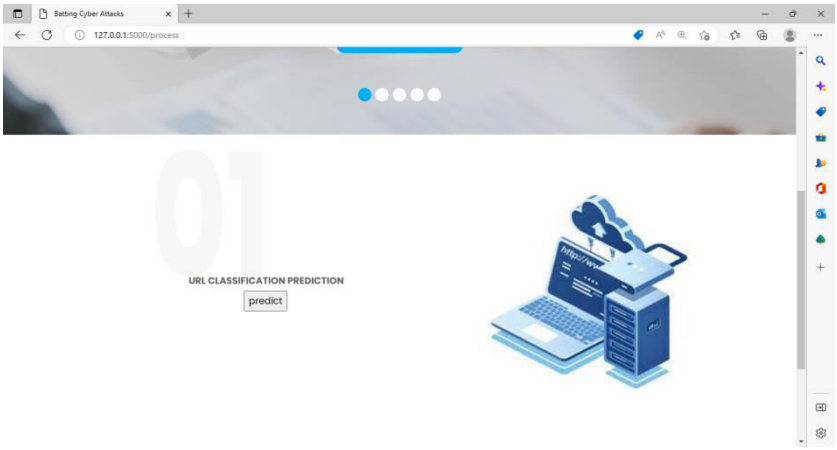


Fig 3: Prediction

Graphical Representation

In this module, if the user want to view the graphical representation of all URLs which are available in dataset.

Fig 4: Graphical Representation

V.CONCLUSION

This project presents PhishStorm, an effective phishing URL detection system based on lexical URL analysis. This approach is based on affinities in the URL. This affinity reflects the relationship between the words in the URL, especially the freely definable part of the URL, and the registered domain. Using search engine query data, we extract 12 features from URLs that characterize internal relevance and popularity. The proposed feature was used in a supervised classification of a ground truth data set of 96,018 phishing and legitimate URLs. This experiment yielded a classification accuracy of 94.91% with a low false positive rate of 1.44%. This experiment was extended by introducing a URL scoring system, Phish Storm, which dynamically calculates a URL's risk score. The test dataset risk assessment can correctly identify 99.22% of the canonical URLs and 83.97% of the dataset's phishing URLs. We extended the original approach to real-time analytics by leveraging modern big data streaming architectures and patterns based on STORM and Bloom filters.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403032|

REFERENCES

[1] C. Harrison, B. Eckman, R. Hamilton, P. Hartswick, J. Kalagnanam, J. Paraszczak, and P. Williams, ‘‘Foundations for smarter cities,’’ IBM J. Res. Develop., vol. 54, no. 4, pp. 1–16, Jul./Aug. 2023.
[2] S. P. Mohanty, U. Choppali, and E. Kougianos, ‘‘Everything you wanted to know about smart cities: The Internet of Things is the backbone,’’ IEEE Consum. Electron. Mag., vol. 5, no. 3, pp. 60–70, Jul. 2022.
[3] O. Shafaiy and N. Farzaneh, ‘‘Smart speed advisory system for drivers with priority assignment for smart city,’’ Int. J. Sensor Netw., vol. 36, no. 2, p. 68, 2021.
[4] UNESCO, ‘‘The united nations world water development report 2019,’’ UNESCO, Paris, France, United Nations World Water Develop. Rep. 3, 2019, p. 36.
[5] M. Mutchek and E. Williams, ‘‘Moving towards sustainable and resilient smart water grids,’’ Challenges, vol. 5, no. 1, pp. 123–137, Mar. 2014.
[6] A. Gonzalez-Vidal, J. Cuenca-Jara, and A. F. Skarmeta, ‘‘IoT for water management: Towards intelligent anomaly detection,’’ in Proc. IEEE 5th World Forum Internet Things (WF-IoT), Apr. 2019, pp. 858–863.
[7] V. Hopman, P. Kruiver, A. Koelewijn, and T. Peters, ‘‘How to create a smart levee,’’ in Proc. 8th Int. Symp. Field Meas. GeoMechanics, 2010, pp. 12–16.
[8] J. Li, X. Yang, and R. Sitzenfrei, ‘‘Rethinking the framework of smart water system: A review,’’ Water, vol. 12, no. 2, p. 412, Feb. 2020.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details