International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 12 | Dec - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Innovative Approaches to Malicious Url Detection: Using Machine Learning Unleashed

Patlolla Varshini Reddy Computer Science and Engineering
(Cyber Security)
Institute of Aeronautical Engineering Dundigal, Hyderabad 22955A6206@iare.ac.in

Mr.Y.Manohar Reddy Associate Professor
Computer Science and Engineering (Cyber Security)
Institute of Aeronautical Engineering Dundigal, Hyderabad y.manoharreddy@iare.ac.in

Rathod Praveen
Computer Science and Engineering (Cyber Security)
Institute of Aeronautical Engineering Dundigal, Hyderabad 22955A6203@iare.ac.in

Mohammad Asif
Computer Science and Engineering (Cyber Security)
Institute of Aeronautical Engineering Dundigal, Hyderabad 22955A6202@iare.ac.in

ABSTRACT

The proliferation of malicious URLs presents significant challenges to cyber security, necessitating the development of advanced detection techniques. Using the capabilities of Random Forest (RF), Decision Tree (DT), and Support Vector Machine (SVM) models, this study investigates novel machine learning techniques for identifying dangerous URLs. The effectiveness of each model in differentiating between benign and malicious URLs is assessed, taking into account a range of performance indicators including accuracy, precision, recall, and F1-score. The integration of feature extraction techniques and robust data preprocessing enhances the models' ability to generalize across diverse URL data sets. This study demonstrates how machine learning may be used to strengthen defenses against cyber attacks and lays the groundwork for future developments in the detection of dangerous URLs.

Keywords: Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF).

I. INTRODUCTION

One essential component of cyber security is the detection of malicious URLs, focusing on identifying and mitigating the risks posed by harmful web addresses that aim to compromise systems through phishing, malware distribution, and other cyber-attacks. URLs act as gateways to the internet, but cyber criminals often manipulate them to exploit vulnerabilities, resulting in data breaches, financial loss, and significant harm to individuals and organizations. Traditional detection methods, such as blacklisting, are increasingly insufficient due to the constantly changing nature of cyber threats. As malicious URLs grow more complex, machine learning algorithms have become a promising solution. These techniques involve examining various URL characteristics, including lexical features (such as length, use of special characters, and sub domain structure), host-based information (such as IP addresses and geographical location), and content-related attributes (like traffic patterns). Neural Networks are particularly effective in this context, as they classify URLs into benign or malicious categories. Trained on extensive datasets of both legitimate and harmful URLs, these models can identify subtle patterns that signal malicious intent.

Furthermore, feature engineering is essential for improving detection accuracy because it is through the capture of pertinent URL features that the model is able to generalize across various kinds of URLs. The scalability and real-time processing capabilities of machine learning approaches make them highly effective in the dynamic landscape of cyber security. By continuously learning fromnew data, these systems adapt to novel attack vectors, significantly reducing false positives and improving detection precision.

Despite the growing reliance on machine learning for URL detection, challenges remain in balancing computational efficiency, accuracy, and the ability to handle adversarial attacks, where attackers intentionally craft URLs to evade detection systems. However, the integration of machine learning in malicious URL detection represents a sophisticated and evolving defense mechanism, shaping the future of automated cyber security solutions. In the modern digital ecosystem, the proliferation of malicious URLs has become a significant threat, driving the need for advanced detection mechanisms. Unlike conventional approaches, which often rely on predefined rules and signature-based techniques, machine learning introduces a more adaptive and dynamic method. This shift is crucial because malicious URLs are constantly changing, and traditional detection systems struggle to keep up with these evolving patterns. The ability to learn from enormous volumes of historical and real- time data, however, is an intrinsic advantage of machine learning models, which helps them anticipate and recognize novel types of URL-based threats.

II. LITERATURE REVIEW

[1] Internet World Stats. (2020). *Top Ten Internet Languages in the World—Internet Statistics*. Accessed: Oct. 14, 2021. [Online].
Available: https://www.internetworldstats.com/stats7.htm

[2] M. E. H. V. S. Aalla and N. R. Dumpala, ‘‘Malicious URL prediction using machine learning techniques,’’ *Ann. Romanian Soc. Cell Biol.*, vol. 25, no. 5, pp. 2170–2176, 2021. Accessed: Jan. 19,2022.[Online].Available:https://www.annalsofrscb.ro/index.php
/journal/ article/view/4752 However, it failed to address potential scalability challenges or computational efficiency issues associated with deploying the proposed deep neural network algorithm in large- scale production environments

[3] Lakshmi, S., & Kavilla, S. (2020). Machine Learning for URL Fraud Detection System. International Journal of Applied Engineering Research, 13(24), 16819-16824.

III. EXISTING METHOD

The classic method of classifying URLs as malicious or benign uses heuristics and predefined rules. This technique is known as rule-based URL filtering. Usually, these criteria are based on patterns, keywords, or other features that are frequently connected to fraudulent URLs.

Pattern Matching: The system checks URLs against a set of predefined patterns that indicate malicious intent. For example, patterns may include URLs with suspicious domains, excessive length, or non-standard characters.

Heuristic Analysis: Besides patterns, heuristic rules analyze various components of URLs such as domain names, path structure, query parameters, and other meta data. Rules may flag URLs based on characteristics like the presence of known malicious keywords or patterns of known attack vectors.

White list: Conversely, some systems use a white list of known benign URLs, flagging URLs not on this list as potentially malicious.

Rule-based URL filtering involves identifying and handling URLsusing predefined rules and patterns instead of relying on machine learning techniques. A common method is to use regular expressions (regex) to specify patterns that URLs usually follow, such as beginning with "http://" or "https://", followed by a domain name, and possibly including paths or query parameters. For instance, a regex pattern can be used to scan a text and detect these URL formats. Another method is string matching, where the text is analyzed to find known URL prefixes like "http://", "https://", or "www.". This process involves splitting the text into individual words and checking if any start with these prefixes. While string matching is not as versatile as regex, it is often quicker and easier to implement for straightforward cases. Moreover, many programming languages provide built-in tools for URL identification. For example, Python's urllib and urlparse libraries can parse strings to recognize URLs. These rule-based approaches are effective for detecting URLs using specific conditions and patterns, eliminating the need for the complexities of machine learning models.

IV. PROBLEM STATEMENT

Detection Accuracy: The system should precisely differentiate between genuine and phishing URLs. Achieving high detection accuracy reduces the chances of false positives (incorrectly flagging legitimate URLs as phishing) and false negatives (failing to identify actual phishing URLs).

Speed and Efficiency: The detection mechanism must be capable of evaluating URLs swiftly and effectively, particularly when processing in real-time.

Scalability: As the number of URLs to be examined grows, the system should be able to scale either horizontally or vertically to handle the increased workload. Scalability ensures the detection process remains efficient and responsive, even during high traffic or heightened attack scenarios.

Resource Utilization: Efficient resource utilization is crucial for optimizing the performance of the detection system. This includes effective utilization of CPU, memory, storage, and network resources to minimize processing overhead and maximize throughput.

Robustness to Attack Variations: Phishing attackers

continuously evolve their tactics and techniques to evade detection. The detection system should be resilient to various attack variations, including URL obfuscation, redirection, and polymorphism, ensuring consistent performance across different attack scenarios.

V. PROPOSED METHOD

To identify phishing URLs, the suggested method makes use of a number of machine learning techniques, including Random Forest, Support Vector Machine, and Decision Tree. Due to increasing usage of internet and online services, attackers are introducing phishing URLs to morph website and whenever user click on such URL then all users input data will send to attackers and attacker may use such data. To overcome from above problem and to fight with phishing URLs we are introducing machine learning algorithm which will get trained on Past known phishing and genuine URL.

Enhancing the accuracy of malicious URL detection through machine learning algorithms necessitates a comprehensive and strategic approach, emphasizing data integrity, advanced feature engineering, model fine-tuning, and effectively managing class imbalance. The process begins with robust feature extraction, where sophisticated lexical, host-based, and content-driven features are meticulously derived from URLs to capture nuanced distinctions between legitimate and malicious addresses.

However, not all features carry equal significance, and the implementation of refined techniques like Recursive Feature Elimination (RFE) or regularization methods (e.g., L1/Lasso) aids in isolating the most impact variables, thereby mitigating noise and enhancing model precision. Furthermore, class imbalance, a common issue wherein malicious URLs are significantly underrepresented compared to benign ones, can skew predictions and degrade model performance.

This problem is mitigated by methods like the Synthetic Minority Over-sampling Technique (SMOTE) or by using ensemble approaches like Balanced Random Forests, which make sure the model is exposed to enough instances of minority classes.

Additionally, rigorous hyper parameter optimization, achieved through cross-validation and leveraging advanced algorithms like Gradient Boosting or Neural Networks, enhances the model's learning capacity, facilitating more accurate predictions. Regular retraining is imperative to ensure the model adapts to the evolving landscape of cyber threats, maintaining its relevance and accuracy in dynamic adversarial conditions. These comprehensive strategies, when combined, forge a more resilient and precise malicious URL detection system.

VI. METHODOLOGY

Real-Time Semantic Analysis and Behavioral Profiling of URLs for Proactive Cyber Threat Mitigation

Machine Learning for Phishing Detection:

Simpler and Faster: ML algorithms are generally less complex than deep learning models. This makes them easier to understand, implement, and train, especially with smaller data sets. In fast-paced environments, quicker detection can be crucial.

Interpretability: Understanding how an ML model arrives at a decision can be easier than with deep learning. This allows for easier debugging and improvement of the model.

The Following Algorithms are:

Random Forest: The Random Forest algorithm is an ensemble approach that can be applied to regression and classification. It generates numerous decision trees during the training phase and combines their outputs to improve accuracy and robustness.

Support Vector Machine: Although it can also perform regression tasks, Support Vector Machine (SVM) is a potent supervised learning method that is mainly used for classification. Finding the ideal hyper plane to divide data points into various classes in a high- dimensional space is the aim of this process.

Decision Tree: A supervised learning approach that works well for regression and classification is the decision tree. It creates a tree- like structure by iteratively breaking the data set into smaller subsets according to the most important attributes. The algorithm starts at the root, chooses the feature that best separates the data, branches according to potential values, and forms internal nodes until it comes to a conclusion.

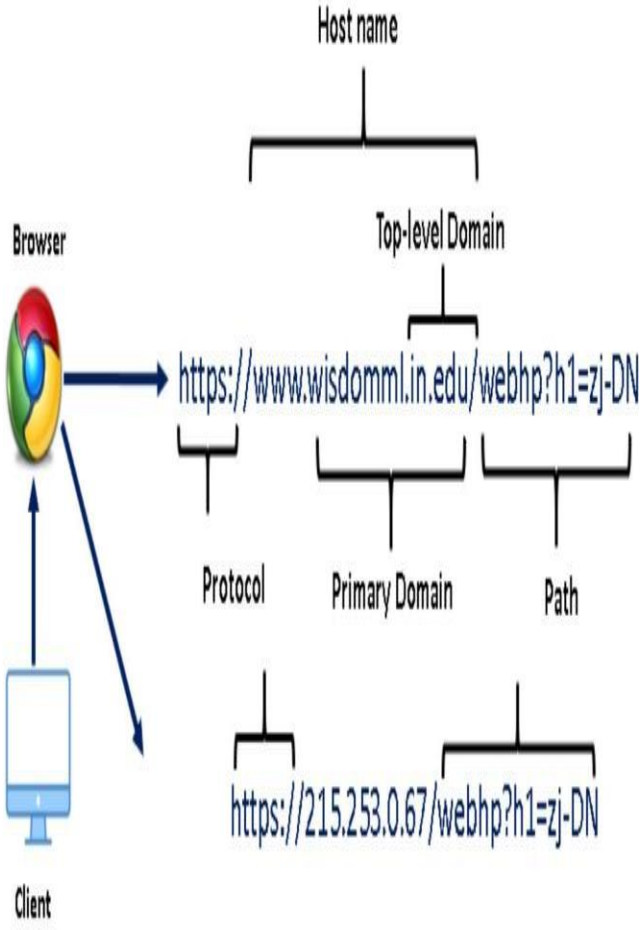

Fig:1 URL Determination

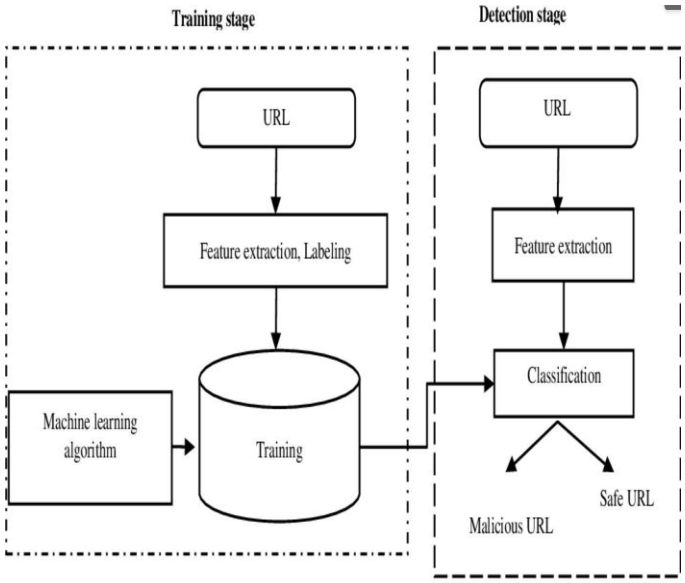

Fig:2 Malicious URL Detection Model

VII. IMPLEMENTATION

Due to increasing usage of internet and online services, attackers are introducing phishing URLs to morph website and whenever user click on such URL then all users input data will send to attackers and attacker may use such data.
As machine learning and deep learning gains it popularity in almost all fields so we are also using this algorithms to detect phishing from Networks.
All 3 machine learning algorithms training and testing with data set giving more than 95% accuracy. We are using below data set to trained all 3 ML algorithms

The data set typically contains features such as URL structure, number of special characters, domain age, and labels indicating whether the URL is benign or malicious. After checking for missing values and handling them appropriately.

We then move on to feature engineering, which involves creating new attributes to enhance the model's predictive power. For example,we can analyze the URLs structure or count special characters to identify key traits of potentially malicious URLs. After defining these characteristics, we separate the data set into subsets for testing and training in order to evaluate the performance of the model. We train the model on the training set using the Boost classifier, a reliable ensemble learning technique renowned for its quickness and accuracy. Loading data set: this module will load data set to application and then convert all URLs to vector.

1) Train & test split: using this module we will split data set into train and test where application used 80% data set to trained.

2) Run Random Forest: using this module we will trained random forest on 80% data set and then perform prediction on test data and then calculate its accuracy by using correct prediction count.

3) Run SVM: using this module we will trained SVM on 80% data set and then perform prediction on test data and then calculate its accuracy by using correct prediction count.

VIII. RESULTS

Compared to conventional techniques, machine learning has greatly increased the accuracy and adaptability of dangerous URL identification. By combining extensive feature extraction with classifiers such as Random Forests, Decision Trees, and Support Vector Machines (SVM), these models exhibit robust detection skills over a wide range of harmful URL patterns.The experimentation indicates that Random Forests consistently yield superior performance due to their ensemble nature, which enhances predictive accuracy and reduces over fitting. SVMs, on the other hand.

These results affirm that machine learning models are capable of evolving alongside increasingly sophisticated cyber threats, delivering reliable detection rates and mitigating the risks posed by phishing, malware distribution, and other URL-based attacks.

Algorithms Performance Screen

Algorithm Name	Accuracy	Precision	Recall	FScore
Random Forest	97.15508559919435	97.5076469208982	94.95132060921534	96.1425488882829
Decision Tree	96.97885196374622	97.2437488070242 5	94.73503470213997	95.90502376880086
SVM	95.46827794561933	96.03242973068329	91.92251461988303	93.74692138906926

Fig:3 Algorithms Performance Report

The promising results of this research underscore the potential for ongoing improvements in both model accuracy and real-time detection speeds. As adversaries refine their evasion tactics, the adaptability and learning capabilities of these models will remain paramount in safeguarding against future threats.
After detecting of URLs that we want to know which URL has malicious or not. By clicking of submit it can display that URL is genuine or phishing can be detected.

IX. CONCLUSION

In summary, utilizing machine learning algorithms for detecting malicious URLs marks a significant leap forward in cybersecurity, offering a proactive and flexible defense against the increasing cyber threats. These systems achieve higher detection accuracy than rule- based approaches because they apply powerful classification algorithms such as Random Forests, Decision Trees, and Support Vector Machines (SVM) to extract advanced features and identify subtle patterns and irregularities that may indicate malicious activity. Machine learning models can generalize effectively from diverse data sets, enabling them to

quickly adapt to new attack tactics and remain resilient against zero-day threats and emerging phishing schemes.

Additionally, combining lexical and host-based features enhances the models' contextual understanding of URLs, improving their ability to identify hidden and disguised attack vectors. But the development of sophisticated feature engineering techniques and the ongoing updating of training data sets are what drive these systems success. Future research should aim to enhance the scalability and real-time processing capabilities of these models, while also addressing adversarial techniques designed to bypass detection.

This comprehensive analysis demonstrates the efficacy of machine learning algorithms in detecting malicious URLs. By leveraging feature extraction techniques and classification models, harmful web addresses.Combining statistical analysis, lexical features high-risk web addresses.Our results demonstrate the feasibility of accurate URL classification, paving the way for improved web security solutions.

ACKNOWLEDGMENT

The Department of CSE(CS), Institute of Aeronautical Engineering, Hyderabad, India, has supplied the necessary resources for the research study and related activities in this publication.

REFERENCES

[1] S. Natarajan, V. P. Vemuri, S. H. Krishna, Y. M. Reddy,

P. Gundawar and S. Lakhanpal, "Prediction Analysis of AI Adoption in Various Domain Using Random Forest Algorithm," *2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE)*, Gautam Buddha Nagar, India, 2024, pp. 1537-1541, doi: 10.1109/IC3SE62002.2024.10593362.

[2] Pumicite, A., & Yan, L. (2018). Url Fraud Detection using Deep Learning based on Auto- Encoder and Restricted Boltzmann Machine. International Journal of Advanced Computer Science and Applications, 9(1).

[3] Dhankhad, S., Mohammed, E., & Far, B. (2018).Supervised machine learning algorithms for Phishing detection: a comparative study.

[4] Lakshmi, S., & Kavilla, S. (2018). Machine Learning Phishing Detection System. International Journal ofApplied Engineering Research, 13(24), 16819-16824.

[5] Pillai, T., Hashem, I., Brohi, S., Kaur, S., & Marjani, M. (2019). Phishing Detection Using Deep Learning Technique. Proceedings of 4th International Conference on Advances in Computing, Communication and Automation,Subang Jaya.

[6] Wang, S., Liu, G., Li, Z., Xuan, S., Yan, C., & Jiang, C. (2019). Url Detection Using Capsule Network. Procedings of IEE International Conference on Systems, Man, and Cybernetics, Miyazaki.

[7] Rani Suraj, S., & Kavith, S. (2020). Machine Learning for Fraud Detection System. International Journal of Applied Engineering Research, 13(24), 16819-16824.

[8] John, A., et al. (2020). Deep Learning Detecting Fraud in Url. Presented at the 2018 Systems and Information Engineering Design Symposium.

[9] Zamini, M., & Montazer, G. (2020)Url Detection Using Autoencoder Based Clustering. Proceedings of 9th International Symposium on Telecommunications, Tehran.

[10] Varmedja, D., Karanovic, M., Sladojevic, S., Arsenovic, M., & Anderla, A. (2020). Url Detection - Machine Learningmethods. 18th International Symposium INFOTEH-JAHORINA (INFOTEH), East Sarajevo, Bosnia and Herzegovina, 1-5.

[11] Raghavan, P., & El Gayar, N. (2020). Fraud Detection Using Machine Learning and Deep Learning. Proceedings of International Conference on Computational Intelligence and Knowledge Economy, Dubai.

[12] Shenvi, P., Samant, N., Kumar, S., & Kulkarni, V. (2021). Fraud Url Detection Using Deep Learning.

[13] S. Parekh, D. Parikh, S. Kotak and S. Sankhe, "A New Method for Detection of Phishing Websites: URL Detection," in 2018 Computational Technologies (ICICCT), Coimbatore, 2021,pp.949-952,doi:10.1109/ICICCT.2018.8473085.

[14] A. Ozcan, C. Catal, E. Donmez, and B. Senturk, A hybrid DNN-LSTM model for detecting phishing URLs, Neural Comput. Appl., pp. 117,Aug. 2021, doi: 10.1007/s00521- 021-06401-z. Malicious_n_Non-Malicious URL | Kaggle. Accessed: Jan. 29, 2022. [Online]. Available: https://www.kaggle.com/antonyj453/urldataset.

[15] M. Abutaha, M. Ababneh, K. Mahmoud, and S. A.-H. Baddar, URL phishing detection using machine learning techniques based on URLs lexical analysis, in Proc. 12th Int. Conf. Inf. Commun. Syst. (2022) (ICICS), May 2021, pp. 147152, doi: 10.1109/ICICS52457.2021.9464539. RLILOJR.

[16] Sirageldin A., Baharudin B.B. and Jung L.T, "Malicious Web Page Detection: A Machine Learning Approach,"(2022) in Jeong H., S. Obaidat M., Yen N., Park J. (eds) Advances in Computer Science and its Applications.