



## MALLICOUS URL DETECTION USING MACHINE LEARNING

**G.Jyothi**, Assistant Professor, CSE (Data Science), Sreyas Institute of Engineering and Technology, Hyderabad, India

**Nunna Harshavardhan**, Sreyas Institute of Engineering and Technology, Hyderabad, India

**Gorla Anjali, Sammeta Sravani**, Sreyas Institute of Engineering and Technology, Hyderabad, India

**Prekki Venkata Sriram**, UG Scholar, Sreyas Institute of Engineering and Technology, Hyderabad, India. [jyothi@sreyas.ac.in](mailto:jyothi@sreyas.ac.in)

### ABSTRACT :

Fraud attacks are important threat in cybersecurity. The attackers change their tactics frequently to deceive the users and steal sensitive information. Among all phishing techniques, phishing URLs masquerading as legitimate login pages are very insidious. They target to deceive the users to give away their credentials. This abstract presents an analysis of real-case scenarios with a focus on the detection of phishing URLs masquerading as login pages. In this paper, we collected a dataset of the legitimate login URL and its phishing counterpart. Using machine learning techniques, we designed a multi-layered model for detecting the phishing URLs. Our method included feature extraction from the URL, which encompasses lexical analysis, structural analysis, and behavioural analysis. We have also used various machine learning classifiers, such as decision trees, random forests, and SVM, to train and test our detection model's accuracy. In return, these features extracted were used for training classifiers so that we can easily separate the phishing URL from legitimate URL. We have executed our experiment over various login URLs that come under diverse domains.

**keywords:** MalliciousURLs, SVM, Random Forest, phishing attacks, Machine Learning

### INTRODUCTION :

The greatest threat in cyber security, as phishing attacks are still ongoing. human vulnerability to deceive users and compromise sensitive information of the various phishing techniques, the impersonation of legitimate login pages is one of the most commonly used and successful methods by cybercriminals. these malicious acts aim to trick users and they try to capture the users personal data like bank details, Accountnumbers, photos etc. Detection of phishing URLs, especially login page spoofing, is still a critical problem in the security space. Classic techniques to determine phishing attempts often involve blacklisting known malicious URLs or relying on user awareness training, which has proven ineffective in dealing with sophisticated attacks by the attackers. Since the phishing attacks are becoming more sophisticated in nature, using obfuscation techniques of URLs, domain spoofing, and all other social engineering tactics, advanced and robust mechanisms for detection are needed.

### LITERATURE SURVEY :

[1] **Sood, S. K., & Enbody, R. J. "Detecting Phishing Websites Using Machine Learning":**

Sood and Enbody suggested a method for detecting fraud websites by leveraging machine learning algorithms. Their research focuses on feature extraction from URLs, including lexical and structural analysis, combined with the use of ensemble learning methods for classification. The paper emphasizes the importance of feature engineering in improving detection accuracy.

[2] **Ramanuja, V., & Khemani, D. "A smart system designed to identify Fraud URLs using machine learning methods"** Ramanuja and Khemani present an intelligent system for detecting phishing URLs employing machine learning techniques. Their study explores the utilization of feature

engineering and classification algorithms to differentiate between good and fraud URLs. The research emphasizes the role of machine learning in enhancing cybersecurity measures against phishing attacks.

[3] **Al-Nemrat, A., & Fergus, P. "Detection of Phishing Websites Using URL Analysis and Machine Learning Techniques"**: Al-Nemrat and Fergus present Extensive investigation into the detection of phishing websites through URL analysis and machine learning techniques. Their study underscores the significance of URL analysis in phishing Detection.

### ALGORITHM AND WORKING PRINCIPLE :

#### Support Vector Machine :

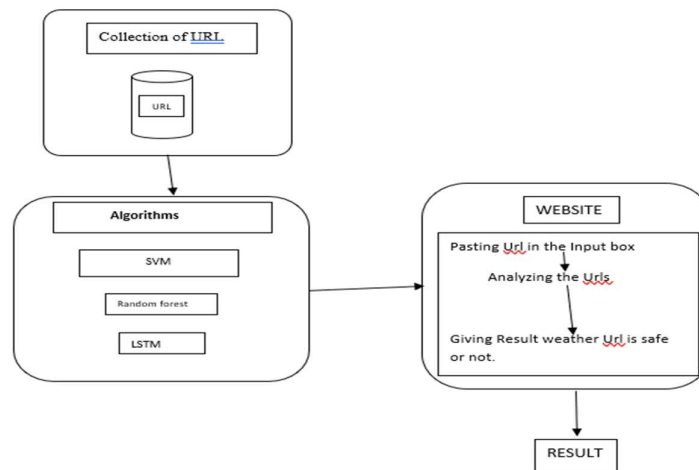
A Support Vector Machine is a supervised machine learning algorithm employed primarily for classification. The SVM seeks to find the best fitting plane which separates the data belonging to different classes with the widest margin. This is a decision boundary. The goal is to increase the margin, which is described as the distance to the nearest data point from either class the support vectors, which are critical for the definition of the optimal hyperplane.

#### Random Forest :

Random Forest is a machine learning technique that builds a collection of decision trees to solve problems like classification and regression. During the training process, it constructs multiple decision trees, each from a randomly selected subset of the data and features (bagging and feature randomness). Aggregation helps improve the final prediction by combining the outputs of multiple models. For classification, regression task. This approach increases model robustness, reduces overfitting, and increases accuracy. Random Forest is highly Meaningful because it can give feature importance scores and is widely applied in domains such as cybersecurity, healthcare, and finance.

#### Long Short Term Memory :

Long Short Term Memory networks are a class of neural networks that are most suited for sequential data analysis; thus, they are quite capable of successfully detecting malicious URLs. URLs are nothing but sequences of characters or words, and the malicious ones have, in most cases, uncommon patterns, like suspicious domain names, weird structure, or abnormal usage of characters. These patterns might be easily captured by LSTMs, which are able to retain relevant information over long sequences while eliminating the irrelevant background noise. Once trained on the dataset containing safe and malicious URLs, the LSTM model can learn to differentiate between the two types of URLs based on their structural and contextual attributes.



### METHODOLOGY :

The methodology for detecting malicious URLs given below.

**Data Collection and Preprocessing:** A URL dataset is downloaded from public databases and security logs. The downloaded URLs are then labelled as benign or malicious. **Preprocessing:** normalize the URLs; remove duplicates, and handle missing data.

**Feature Extraction:** Derived features from a URL to present its characteristics like:

- o lexical features - URL length, presence of special characters and suspicious keywords
- o host-based features- domain age, IP usage and server location.
- o Behavioural features- redirection patterns, file downloads-if available.

**Feature Selection:** The appropriate features are selected to enhance model performance, thus reducing the complexity of the model by making a correlation analysis or feature importance ranking.

**Model Selection:** Models from machine learning algorithms are chosen depending on the problem. Models' selection will help to understand the model and we will use different algorithms Each have its individual process and steps.

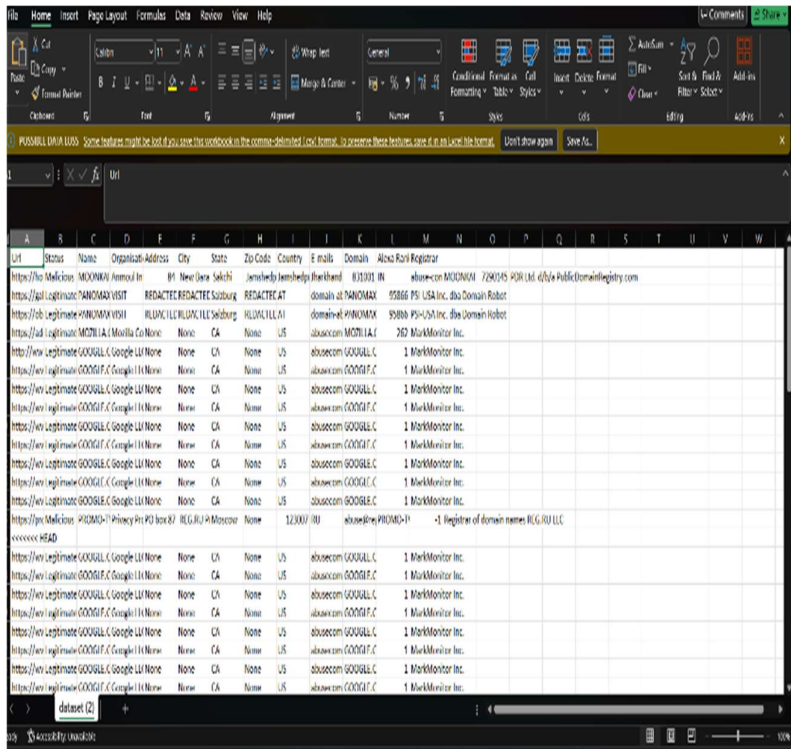
**Model Training:** The datasets are splitted for training, testing. Models are trained and optimized using cross-validation and hyperparameter tuning. Techniques like oversampling are applied to handle imbalanced datasets.

**Model Evaluation:** The model is observed by comparing accuracy, precision, recall to ensure it generalizes well to unseen data.

**Deployment:** The learned model is inserted inside the web-browsing softwares, such as email or web browsers, for live detection of malicious URLs. More incessantly, the model updates itself on newly acquired data.

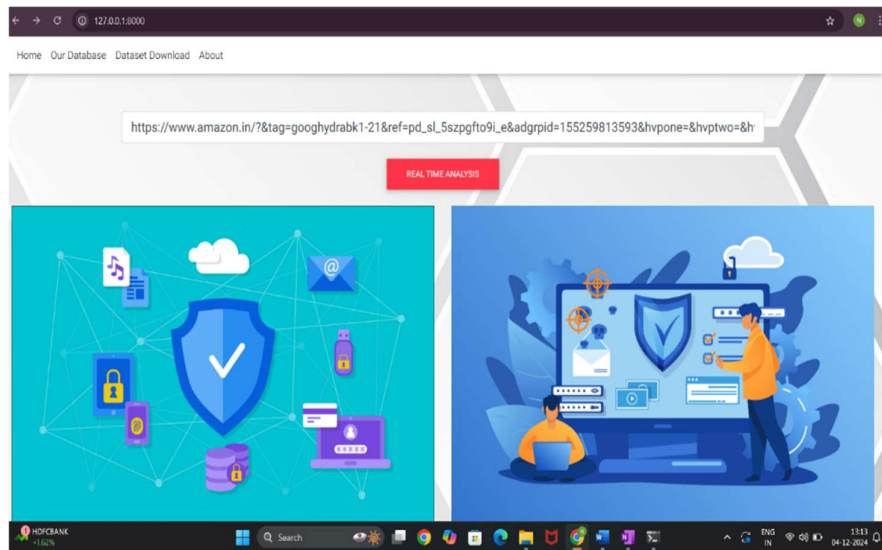
**RESULTS AND DISCUSSION :**

Collecting the genuine URL's and Mallicious URL's in a Data set.

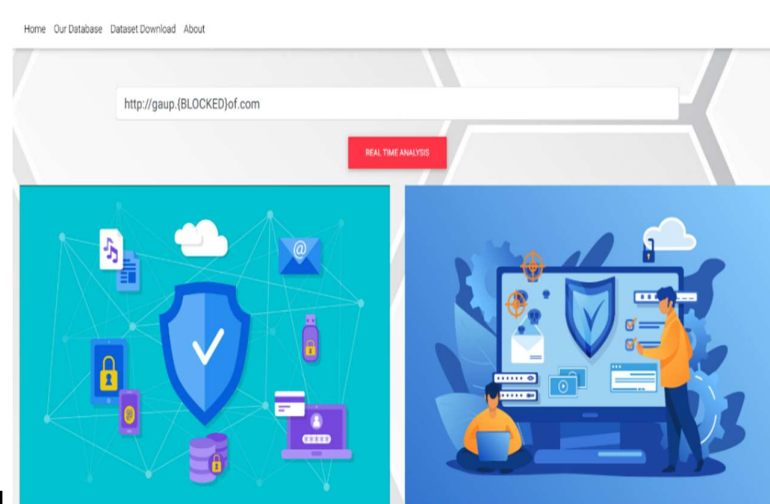
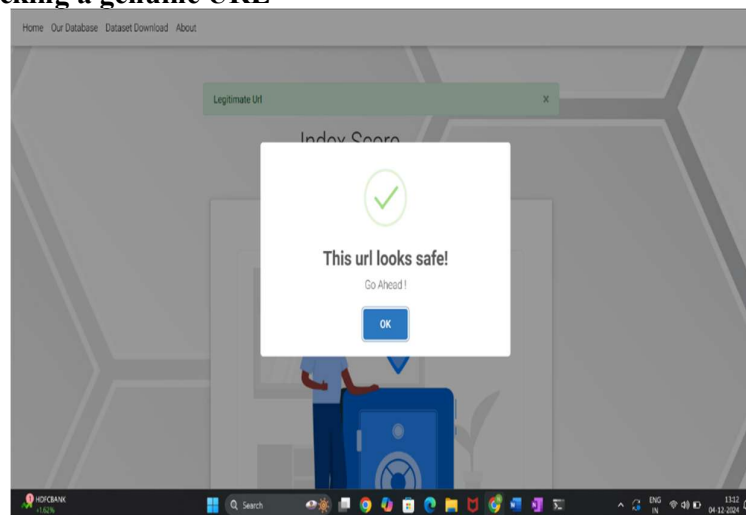


URL	Status	Name	Organization	Address	City	State	Zip Code	Country	E-mail	Domain	Alexa Rank	Registrar
https://www.malicious-moon.com/	Malicious	MOON.COM	Amool In	191 New Dera Sakshi	Jamshedpur	Jharkhand	831001	IN	abuse@moon.com	720045	PSI-DK	Public Domain Registry.com
https://legitimate-panorama.com/visit	Legitimate	PANORAMA	VISHI	REDACTED	Redacted	AT			domains@panorama.com	95866	PSI-USA	Inc. Domain Robot
https://legitimate-panorama.com/visit	Legitimate	PANORAMA	VISHI	REDACTED	Redacted	AT			domains@panorama.com	95866	PSI-USA	Inc. Domain Robot
https://legitimate-mozilla.com/	Legitimate	Mozilla	CA	None	None	US			abuse@mozilla.com	363		MarkMonitor Inc.
https://legitimate-google.com/	Legitimate	Google	LLP	None	CA	None	US		abuse@google.com			MarkMonitor Inc.
https://legitimate-google.com/	Legitimate	Google	LLP	None	CA	None	US		abuse@google.com			MarkMonitor Inc.
https://legitimate-google.com/	Legitimate	Google	LLP	None	CA	None	US		abuse@google.com			MarkMonitor Inc.
https://legitimate-google.com/	Legitimate	Google	LLP	None	CA	None	US		abuse@google.com			MarkMonitor Inc.
https://legitimate-google.com/	Legitimate	Google	LLP	None	CA	None	US		abuse@google.com			MarkMonitor Inc.
https://legitimate-google.com/	Legitimate	Google	LLP	None	CA	None	US		abuse@google.com			MarkMonitor Inc.
https://legitimate-google.com/	Legitimate	Google	LLP	None	CA	None	US		abuse@google.com			MarkMonitor Inc.
https://legitimate-google.com/	Legitimate	Google	LLP	None	CA	None	US		abuse@google.com			MarkMonitor Inc.
https://legitimate-google.com/	Legitimate	Google	LLP	None	CA	None	US		abuse@google.com			MarkMonitor Inc.
https://legitimate-google.com/	Legitimate	Google	LLP	None	CA	None	US		abuse@google.com			MarkMonitor Inc.
https://legitimate-google.com/	Legitimate	Google	LLP	None	CA	None	US		abuse@google.com			MarkMonitor Inc.
https://legitimate-google.com/	Legitimate	Google	LLP	None	CA	None	US		abuse@google.com			MarkMonitor Inc.
https://legitimate-google.com/	Legitimate	Google	LLP	None	CA	None	US		abuse@google.com			MarkMonitor Inc.
https://legitimate-google.com/	Legitimate	Google	LLP	None	CA	None	US		abuse@google.com			MarkMonitor Inc.
https://legitimate-google.com/	Legitimate	Google	LLP	None	CA	None	US		abuse@google.com			MarkMonitor Inc.
https://legitimate-google.com/	Legitimate	Google	LLP	None	CA	None	US		abuse@google.com			MarkMonitor Inc.
https://legitimate-google.com/	Legitimate	Google	LLP	None	CA	None	US		abuse@google.com			MarkMonitor Inc.
https://legitimate-google.com/	Legitimate	Google	LLP	None	CA	None	US		abuse@google.com			MarkMonitor Inc.
https://legitimate-google.com/	Legitimate	Google	LLP	None	CA	None	US		abuse@google.com			MarkMonitor Inc.

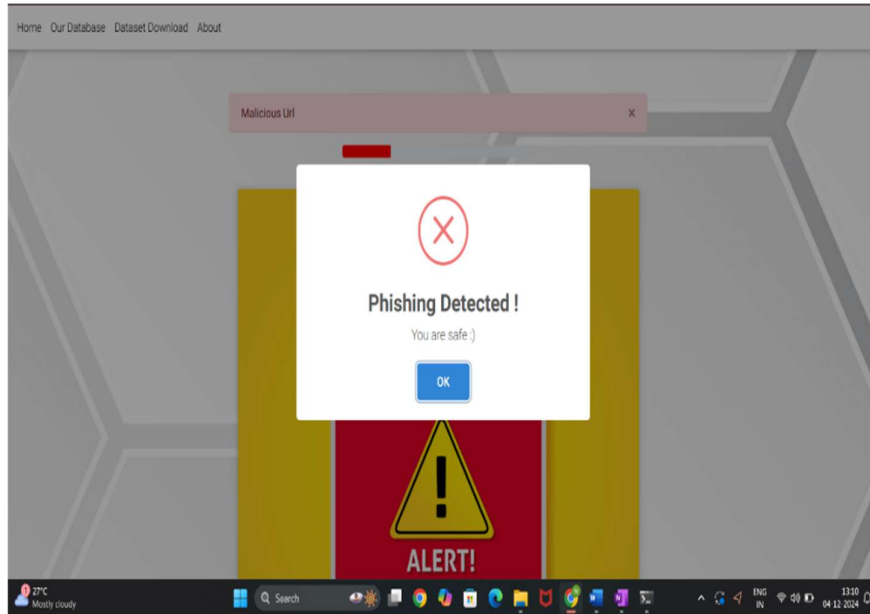
Checking Amazon website Url



**Result by checking a genuine URL**



**Checking a malicious Url.**

**RESULT FOR THE MALLICIOUS URL:****CONCLUSION :**

In conclusion, the detection of phishing URLs targeting login pages is a critical aspect of cybersecurity, given the prevalence and sophistication of phishing attacks in today's digital landscape. Through the implementation of advanced machine learning techniques and comprehensive feature analysis, the system presented in this study offers a robust solution to identify and mitigate phishing threats effectively. By collecting a diverse dataset of legitimate login URLs and phishing counterparts, the system leverages feature extraction techniques, including lexical, structural, and behavioural analysis, to identify unique characteristics indicative of phishing attempts. Machine learning models, such as decision trees, random forests, and deep neural networks, are employed to train and evaluate the detection model, achieving high accuracy and performance metrics. Real-world applicability testing demonstrates the effectiveness of the detection model against a diverse range of login URLs from different domains and industries, including emerging phishing techniques such as URL obfuscation and domain spoofing. Continuous monitoring and updates ensure adaptability to evolving phishing strategies, enhancing the system's resilience against cyber threats.

**REFERENCES :**

- [1] NP Mankar, PE Sakunde, S Zurange... - ... MIT Art, Design and ..., 2024 - iceeexplore.ieee.or
- [2] Survey on Malicious URL Detection Techniques BY Saleem Raja A; Madhubala R; Rajesh N; Shaheetha L; Arulkumar N published at published at 9th oct 2023
- [3] Finding effective classifier for malicious URL detection by youn zhou ,Bo lang ,chunlin published at 18th may 2022
- [4] Malicious URL Detection: A Comparative Study by Shantanu; B Janet; R Joshua Arul Kumar published in 2 April 2021
- [5] Kumar, V., & Dhiman, G. (2020). "A Machine Learning Approach for Phishing URL Detection."
- [6] Liu, F., Du, W., & Zhang, X. (2020). "Detecting Phishing Websites Based on Machine Learning Algorithms." IEEE Access, 8, 209490-209498.
- [7] Khammas, A., Ben-Youssef, A., & Toumi, F. (2019). "Phishing Websites Detection using Machine Learning Algorithms." Procedia Computer Science, 151, 846-853.



- [8] Al-Nemrat, A., & Fergus, P. (2019). "Detection of Phishing Websites Using URL Analysis and Machine Learning Techniques."
- [9] Wang, Z., Wu, W., & Wang, L. (2019). "A Machine Learning Based Approach for Phishing URL Detection." International Conference on Security, Privacy and Anonymity in Computation, Communication and Storage. Springer, Cham.
- [10] Thanh, H. T., & Nguyen, V. H. (2018). "An Adaptive Machine Learning System for Phishing URL Detection." Journal of Network and Computer Applications, 117, 90-100.
- [11] Srinivas Kalime, Naresh Boddula- "A Study on Detection of Distributed Denial of Services Attacks Using Machine Learning Techniques" IJR Available at [https:// edupediapublications .org/journals](https://edupediapublications.org/journals), e-ISSN: 2348-6848 p-ISSN: 2348-795X Volume 05 Issue 12 April 2018.