IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Phishing Url Detection Using Machine Learning

[1]Vijay Vitthal Patil, [2]Somraj Kumar Patil,

[1,2]Students, [3]Assistant Professor,

[1,2,3] Computer Science Department,

Abstract: Phishing attacks, designed to deceive users into divulging sensitive information by masquerading as trustworthy entities, are a significant threat in the digital landscape. Detecting phishing URLs (Uniform Resource Locators) is crucial to safeguarding online users and protecting sensitive data. Traditional approaches to phishing detection, often reliant on manual blacklisting and heuristic methods, struggle to keep pace with the rapidly evolving tactics of attackers. This study explores the application of machine learning techniques to improve the detection of phishing URLs, leveraging their ability to learn from data and identify patterns indicative of phishing activities.We propose a robust framework for phishing URL detection using machine learning algorithms, combining feature extraction techniques and classification models. Our methodology involves the extraction of key features from URLs, including lexical characteristics, domain-based features, and URL metadata. These features serve as inputs to various machine learning classifiers, including Logistic Regression, Support Vector Machines (SVM), Random Forest, and Gradient Boosting, among others.

Keywords - URL detection, Machine learning, Cybersecurity, Feature extraction, Lexical analysis, Domain attributes, Web page content.

I. Introduction

Exploiting any type of weakness in the computing environment, cybercrime is on the rise worldwide. Ethical hackers focus more on identifying vulnerabilities and suggesting solutions for remediation. Within the cyber security world, there has been a pressing need for the development of efficient techniques. The majority of IDS approaches in use today are unable to handle the complex and dynamic nature of cyberattacks on computer networks. Due to machine learning's efficiency in solving cyber security problems, machine learning has recently gained significant attention.

Machine learning approaches have been used to address important cyber security difficulties such as spam, phishing, malware categorization and detection, and intrusion detection.

Though it is unable to automate a whole cyber security system, machine learning helps identify threats to cyber security more effectively than other software-oriented approaches, which reduces the pressure for security analysts. We want to show that the problem of identifying attacks is very different from these other applications, and that this difference makes it much harder for the intrusion detection sector to make effective use of machine learning.

Heuristic criteria and blacklist-based techniques are two ways from the past that have limits when it comes to identifying phishing URLs. Blacklists have to be updated and maintained frequently, and they have trouble detecting recently established phishing websites. Conversely, heuristic approaches usually depend on pre-established patterns that are easily circumvented by adversaries that regularly modify their tactics to get around detection systems. As a result, more advanced and flexible solutions are desperately needed to properly combat the dynamic nature of phishing attempts.

OBJECTIVE :

- To Accurately Identify Phishing URLs.
- To Evaluate Detection Effectiveness.
- To Generate Contextually Accurate Prediction.
- To Simplify URL Analysis.
- To Enable Quick Decision-Making.
- To Identify Key Indicators of Phishing.
- To Improve Security and Efficiency.

II. SYSTEM ARCHITECTURE

Machine learning-based phishing URL detection is a vital cybersecurity feature that involves detecting malicious URLs intended to steal confidential data. The architecture that is being given here combines a number of different parts and procedures to effectively gather data, preprocess it, extract features, train and assess machine learning models, and apply continuous updates and real-time detection.

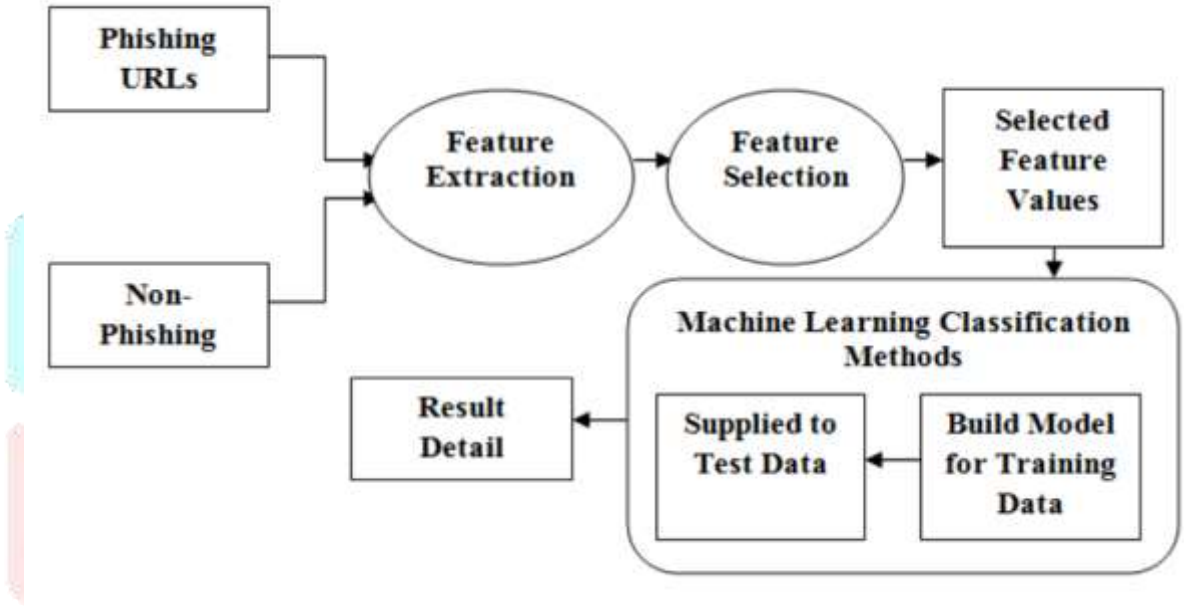

Fig 1 : System Architecture

Data Gathering and Preprocessing Layer: URL Gathering: URLs are gathered from a variety of sources, including threat intelligence databases, web crawlers, and user contributions. By doing this step, a complete dataset with authentic and phishing URLs is ensured.Data Labeling: To guarantee high accuracy, a combination of automatic techniques (such as matching against lists of known phishers) and human inspection is used to classify the gathered URLs as either "phishing" or "legitimate." Data cleaning is the process of eliminating duplicate entries, eliminating unnecessary data, and making sure the dataset is prepared for analysis. Extraction of Relevant Features: Lexical features (such as URL length and special characters), domain features (such as domain age and registration details), network features (such as IP address), and content features are among the features that are extracted from URLs.

Machine Learning Model Layer: Model Training: This step involves training various machine learning models such as Support Vector Machines (SVM), Random Forests, and Gradient Boosting on the prepared dataset. Feature selection and hyperparameter tuning are performed to optimize model performance. Model Evaluation: The trained models are evaluated using metrics like accuracy, precision, recall, and F1-score to select the best-performing model for deployment. Model Storage: The trained models are stored in a secure, version-controlled repository for easy access and updates.

Real-Time Detection Layer: URL Input Interface: A user-friendly interface (web or API-based) is provided for submitting URLs for real-time analysis. Batch processing options are available for analyzing multiple URLs simultaneously. Feature Extraction: In real-time, features are dynamically extracted from submitted URLs, mirroring the preprocessing steps used during model training. Model Inference: The deployed

machine learning model analyzes the features to classify URLs as 'phishing' or 'legitimate.' The system provides a confidence score along with the classification result. Feedback Loop: User feedback on the accuracy of the classifications is collected, which helps in refining and improving the detection model over time.

I. RESEARCH METHODOLOGY

3.1Data Flow Diagram

DFDs are applicable to various types of systems, including information systems, business processes, and software systems. They aid in visualizing and analyzing data flow, identifying areas of congestion and inefficiency, and effectively conveying system design to others. data flow diagram can be created using different notations, such as the Gane-Sarson notation and the Yourdon-DeMarco notation, depending on the designer's preferences and conventions.

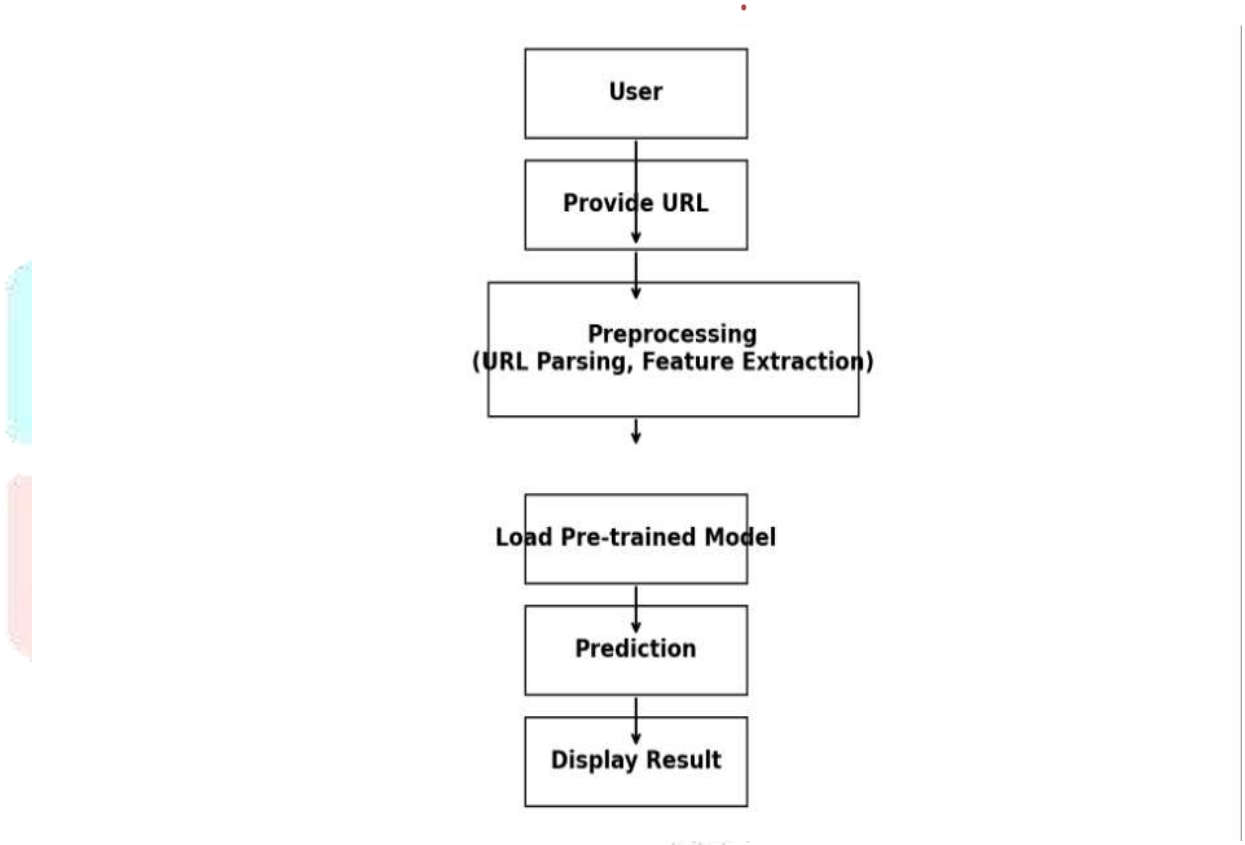

Fig 2 : Data Flow Diagram

Components of the Data Flow Diagram:

1. URL Source:
   - Represents external sources of URLs which can be websites, user submissions, or third-party databases.
2. Data Collection:
   - Collects URLs from various sources for analysis. Data is ingested into the system.
3. Data Preprocessing:
   - Cleans and prepares the collected URLs by removing duplicates, filtering noise, and formatting data.
4. Feature Extraction:
   - Extracts relevant features from URLs, such as lexical characteristics, domain information, and content features.
5. Model Training:
   - Uses the extracted features to train a machine learning model to distinguish between phishing and legitimate URLs.

6. Model Evaluation:
    - Evaluates the trained model's performance using metrics like accuracy, precision, and recall.
7. Real-Time Detection:
    - Applies the trained model to new URLs for real-time phishing detection. Classifies URLs as either phishing or legitimate.
8. User Interface:
    - Provides an interface for users to submit URLs for analysis and receive feedback.

IV. RESULTS AND DISCUSSION

4.1 Register Page :

Fig 3: Register page

A phishing URL detection home page is the main interface of a web application designed to help users identify and avoid malicious websites. This page typically provides several key features and functionalities to ensure the user can effectively assess the safety of URLs. Here's a detailed look at the elements and structure of a phishing URL detection home page.

4.2 Malware Detection page :

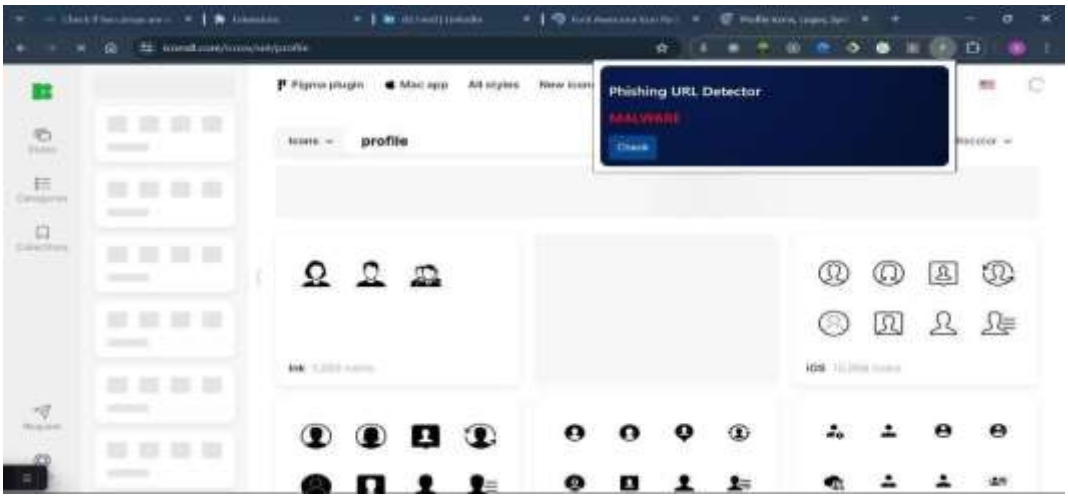

Fig 4: Malware Detection page

A malware detection page serves as the main interface for a web application that helps users detect and analyze potential malware threats on their devices or within files. This page is designed to be user-friendly while providing powerful tools for malware identification and reporting. Here's an in-depth look at the components and structure of a malware detection page.

4.3 : Data set which detect fishing website

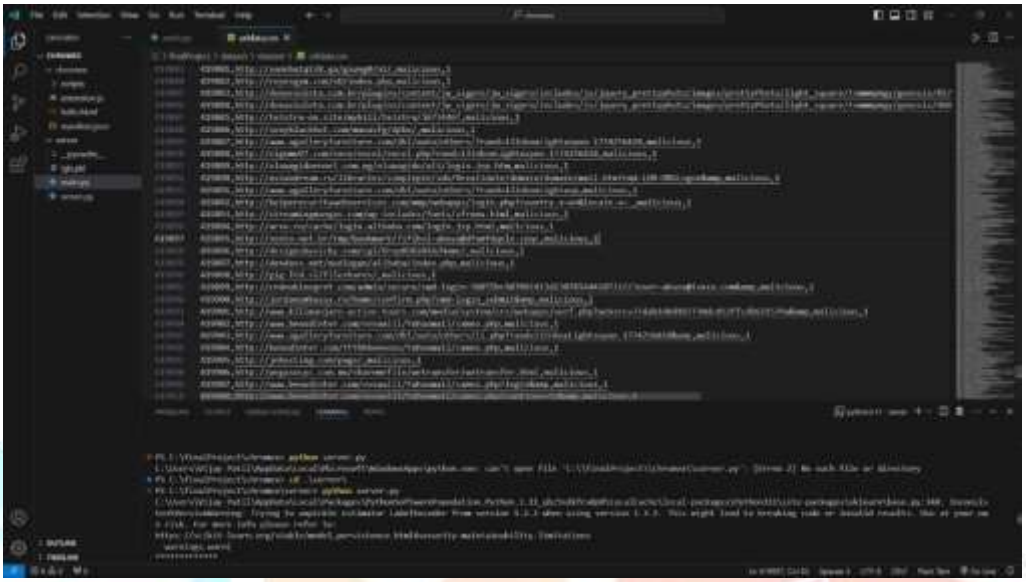

Fig 5 : Data set which detect fishing website

This are various data set which helps server to detect the fishing website. The web send the message to server and server blocks the website ASP.

III. CONCLUSION

THE PROJECT OF PHISHING URL DETECTION USING MACHINE LEARNING PROVIDES A COMPREHENSIVE SOLUTION FOR IDENTIFYING AND MITIGATING PHISHING THREATS. BY INTEGRATING ADVANCED MACHINE LEARNING TECHNIQUES, THE SYSTEM CAN EFFICIENTLY ANALYZE AND CLASSIFY URLS IN REAL-TIME, OFFERING A HIGH DEGREE OF ACCURACY AND ADAPTABILITY. THIS AUTOMATED APPROACH ENHANCES SECURITY BY QUICKLY IDENTIFYING MALICIOUS URLS, REDUCING THE RISK OF CYBER ATTACKS, AND PROTECTING USERS ACROSS VARIOUS PLATFORMS. THE PROJECT UNDERSCORES THE SIGNIFICANT ROLE OF MACHINE LEARNING IN MODERN CYBERSECURITY, OFFERING A SCALABLE AND EFFECTIVE DEFENSE AGAINST PHISHING ATTEMPTS.

III. REFERENCES

[1] Bindu J, Ravi Kumar K. "Phishing URL Detection using Machine Learning: A Review of Features and Algorithms." International Journal of Advanced Research in Computer Science, March 2023.
[2] Rahul R, S. Sundar, "A Comprehensive Survey on Machine Learning Techniques for Phishing Detection." International Journal of Computer Applications, April 2023.
[3] Mahesh S, Megha M, "Phishing Detection Using Machine Learning: Techniques, Challenges, and Future Directions." Journal of Network Security, May 2023.
[4] Anirudh K, Deepika R, "Automated Detection of Phishing Websites Using Machine Learning: A Comparative Study." International Journal of Cyber Security and Digital Forensics (IJCSDF), July 2023.
[5] Pallavi R, Suresh N, "An In-Depth Review on Phishing URL Detection Techniques Using Machine Learning." International Journal of Engineering Research and Technology (IJERT), August 2022.
[6] Santosh M, Anusha R, "Phishing URL Detection Using Supervised Machine Learning Algorithms." International Journal of Information Technology and Computer Science (IJITCS), November 2021.
[7] Divya S, Praveen K, "A Survey on Phishing URL Detection Using Machine Learning Techniques." IEEE Conference on Innovations in Computing, September 2021.
[8] Ramesh K, Sanjay R, "Phishing Website Detection Using Machine Learning Algorithms: A Detailed Analysis." International Journal of Advanced Trends in Computer Science and Engineering, June 2022.
[9] Prashant A, Naveen P, "Comparative Study of Machine Learning Models for Phishing URL Detection." International Journal of Computer Science and Information Security (IJCSIS), December 2021.
[10] Karthik N, Rohit S, "Efficient Phishing Detection System Using Machine Learning: Current Trends and Future Prospects." Journal of Internet Technology and Security, January 2022.