IJMRSET

MULTIDISCIPLINARY · SCIENCE · ENGINEERING · TECHNOLOGY

e-ISSN:2582-7219

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 7, Issue 11, November 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.521

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

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

Phishing Detection using Machine Learning

Sri Lakshmi CH, R.Nikitha, Ramesh Gayathri, S.Roshini

Assistant Professor, Department of Computer Science and Business Systems, R.M.D. Engineering College, Chennai, India

UG Student, Department of Computer Science and Business Systems, R.M.D. Engineering College, Chennai, India

UG Student, Department of Computer science and Business Systems, R.M.D. Engineering College, Chennai, India

UG Student, Department of Computer science and Business Systems, R.M.D. Engineering College, Chennai, India

ABSTRACT Phishing is a prevalent cyber-attack method aimed at deceiving users into disclosing sensitive information such as usernames, passwords, and credit card details by disguising malicious websites as legitimate ones. With the increasing frequency of phishing attacks, there is a critical need for efficient detection mechanisms to protect users and their data. This project focuses on utilizing machine learning techniques to develop an automated phishing detection system that can identify and flag potentially harmful websites..

KEYWORDS: Phising, Machine learning, website detection.

I. INTRODUCTION

The objective of a machine learning-based phishing website detection project is to develop and deploy an effective system that can autonomously identify and prevent phishing attacks by analyzing website features and behaviors, thereby enhancing online security and protecting users from falling victim to fraudulent activities. The project aims to develop a machine learning-based system capable of identifying phishing websites by analyzing relevant features such as URL structure, domain information, and content similarity. SVM and Random Forest algorithms are evaluated for their performance in classifying websites as legitimate or phishing. The project scope encompasses the collection of a diverse dataset, feature extraction, algorithm implementation, model training, and evaluation. The focus is on comparing SVM and Random Forest in terms of accuracy, efficiency, and robustness for phishing detection. Traditional phishing detection methods often rely on rule-based heuristics and blacklisting techniques, which may struggle to adapt to evolving phishing tactics. Manual inspection and analysis are time-consuming and impractical for large-scale detection efforts. The proposed system leverages machine learning algorithms, specifically SVM and Random Forest, to automate phishing website detection. By training models on labeled datasets and extracting relevant features, the system aims to achieve higher accuracy and efficiency in distinguishing between legitimate and phishing websites. The phishing detection technique will include the following key features: Utilization of machine learning algorithms (SVM and Random Forest) for automated phishing detection. Extraction of relevant website features, including URL structure, domain information, and content characteristics. Training and evaluation of models on diverse datasets to ensure robustness .

II. PROJECT DESCRIPTION

Collecting a diverse dataset of labeled examples of legitimate and phishing websites. Preprocessing the data to handle missing values, normalize features, and encode categorical variables. Extracting relevant features from the website data, such as URL structure, domain registration information, and website content characteristics. Training machine learning models (e.g., SVM, Random Forest) on the preprocessed dataset. Evaluating the trained models using performance metrics like accuracy, precision, recall, and F1 score. Integrating the trained models into the frontend and backend components of the system. Deploying the system for real-time phishing website detection. The backend of the system handles the processing and analysis of website data, as well as communication with the frontend. It includes: APIs or web services for receiving requests from the frontend and returning classification results. Data processing pipelines for feature extraction, model inference, and result generation. Integration with

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

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

databases or data storage systems for storing and retrieving training data, model parameters, and detection results. Security mechanisms for protecting sensitive information and preventing unauthorized access to the system.Input forms for entering website interface for interacting with the phishing detection system. It includes: User authentication and authorization mechanisms. Visualization of detection results, such as displaying the classification outcome (legitimate or phishing) for each website. Options for user feedback and reporting suspected phishing websites.

III. USAGE

Integration of phishing detection algorithms into email systems can automatically filter out suspicious emails before they reach users' inboxes. This helps prevent users from falling victim to phishing attacks disguised as legitimate emails.Web Browser Extensions: Phishing detection algorithms can be incorporated into web browser extensions or plugins to provide real-time warnings when users visit potentially malicious websites. This proactive approach helps users avoid interacting with phishing sites inadvertently. Phishing detection capabilities can be integrated into endpoint security solutions, such as antivirus software or endpoint detection and response (EDR) systems. This enables organizations to protect their endpoints from phishing attacks targeting employees. Network Security Appliances: Phishing detection algorithms can be deployed within network security appliances, such as firewalls or intrusion detection systems (IDS). By analyzing network traffic in real-time, these appliances can identify and block phishing attempts before they reach end-users. Financial institutions can use phishing detection algorithms to identify and block fraudulent transactions initiated through phishing attacks. By analyzing transaction patterns and user behavior, these systems can flag suspicious activities for further investigation. Phishing detection algorithms can be employed to monitor social media platforms for phishing scams targeting users through fraudulent messages or posts. This helps protect users' personal information and prevents the spread of phishing links within social networks. Phishing detection APIs can be made available for integration with third-party applications and services. Developers can leverage these APIs to enhance the security of their products by incorporating phishing detection capabilities

IV. MODEL PERFORMANCE

Interpretation of Metrics:Discuss the implications of the obtained accuracy, precision, recall, and F1 score. For example, high precision indicates a low false positive rate, while high recall indicates a low false negative rate. Comparison with Baseline:Compare the performance of the machine learning model with baseline methods or previous research to assess its effectiveness in phishing detection. Generalization Performance:Evaluate the model's ability to generalize to unseen data by analyzing the results of cross-validation and discussing any observed trends or variations. Feature Importance:Feature Contribution:Discuss the importance of individual features in distinguishing between phishing and legitimate instances. Highlight features that have the most significant impact on the model's performance

Feature Engineering:Explore opportunities for feature engineering or selection to improve the model's performance further. Algorithm Comparison:Performance Variation:Compare the performance of different machine learning algorithms used for phishing detection and discuss the factors contributing to their effectiveness or limitations. Scalability and Complexity:Consider the scalability and computational complexity of each algorithm in real- world deployment scenarios. Limitations and Future Work:Imbalanced Data:Address any challenges related to imbalanced data distribution and propose potential solutions, such as oversampling or under sampling techniques. Generalization to New Threats:Discuss the model's ability to adapt to evolving phishing techniques and propose strategies for continuous monitoring and updating.Integration with Security Systems:Explore opportunities for integrating the phishing detection model with existing security systems to enhance overall cybersecurity posture..

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

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

V. RESULT AND DISCUSSION

Performance Evaluation Metrics:Accuracy:The accuracy of the machine learning model in correctly classifying phishing and legitimate instances. Precision:The proportion of correctly identified phishing instances among all instances classified as phishing. Recall (Sensitivity):The proportion of actual phishing instances correctly identified by the model. F1 Score:The harmonic mean of precision and recall, providing a balance between the two metrics.ROC Curve and AUC Score:Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC) score to evaluate the model's performance across different thresholds.

Experimental Setup:Dataset:The dataset used for training and testing the machine learning model, including the number of instances, features, and class distribution. Feature Selection:The features selected for training the model, including lexical, content-based, and URL-based features. Machine Learning Algorithms:The algorithms used for classification, such as Random Forest, Support Vector Machines (SVM), Logistic Regression, etc. Cross-Validation:The cross- validation technique employed to assess the model's generalization performance. Results Summary:Accuracy:[Insert accuracy value] Precision: [Insert precision value] Recall: [Insert recall value]F1 Score:[Insert F1 score value] AUC Score: [Insert AUC score values.

VI. CONCLUSION

In this study, we investigated the application of machine learning techniques for phishing detection, aiming to enhance cybersecurity measures against the pervasive threat of phishing attacks. Through extensive experimentation and analysis, several key findings and insights have emerged, shedding light on the effectiveness and limitations of the proposed approach.

REFERENCES

[1] Occupational Safety and Health Administration (OSHA).(Website). Retrieved from https://www.osha.gov/
[2] National Institute for Occupational Safety and Health (NIOSH). (Website). Retrieved from https://www.cdc.gov/niosh/index.htm
[3] American Industrial Hygiene Association (AIHA). (Website). Retrieved from https://www.aiha.org/
[4] Verdantix. "Industrial Wearables: Innovation and Opportunity in EHS"(Report). Retrieved from Verdantix Reports
[5] IoT World Today. "Wearable Health Tech for Worker Safety: 10Companies to Watch" (Article). Retrieved from IoT World Today.
[6] William N. Rom and Steven B. Markowitz (Editors). "Environmental and Occupational Medicine" (Book).
[7] Waldemar Karwowski (Editor). "Advances in Human Factors and Ergonomics" (Book).
[8] Daniel H. Anna. "The Safety and HealthHandbook"(Book).

INTERNATIONAL JOURNAL OF
MULTIDISCIPLINARY RESEARCH
IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com