IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 7, July 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

9940 572 462 | 6381 907 438 | ijircce@gmail.com | www.ijircce.com

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1207040 |

Unveiling Cyber Deception: "The Anatomy of Phishing URLs"

Sanjana G N, Ms. Bavyashree H M

PG Student, Dept. of MCA., National Institute of Engineering, Mysore, India

Assistant Professor, Dept. of MCA., National Institute of Engineering, Mysore, India

ABSTRACT: In the ever-evolving landscape of cybersecurity, the need for robust threat detection mechanisms has become paramount. Traditional methods often fall short in combating sophisticated threats such as phishing and Business Email Compromise (BEC) attacks. To address this challenge, this project proposes the development of a comprehensive Threat Detection Engine leveraging Machine Learning (ML) techniques. The engine encompasses a blend of advanced algorithms and data sources to achieve remarkable accuracy and efficiency in threat detection. Key features include expanded threat coverage, detecting a broad spectrum of threats including phishing, BEC, lookalike websites, similar domains, and compromised IP addresses. The project employs a variety of supervised and unsupervised learning algorithms for classification and anomaly detection, utilizing diverse data sources such as email headers, text content, IP addresses, and domain information. Experimental results demonstrate the effectiveness of the proposed approach, showcasing significant improvements in detection accuracy compared to traditional methods. The findings from this project hold promise for enhancing cybersecurity measures and mitigating the risks posed by evolving cyber threats.

I. INTRODUCTION

In today's digitally interconnected world, the proliferation of cyber-attacks poses significant challenges to individuals, organizations, and governments alike. Among the multitude of cyber-attacks, phishing attacks and Business Email Compromise (BEC) attacks stand out as particularly insidious, exploiting human vulnerabilities and leveraging social engineering tactics to deceive victims and compromise sensitive information. Traditional threat detection methods, reliant on static rules and signature-based approaches, often prove inadequate in thwarting these increasingly sophisticated attacks. As cybercriminals continuously evolve their tactics to bypass conventional defense's, there is a pressing need for innovative solutions that can adapt and effectively combat emerging cyber-attacks. This project addresses the shortcomings of traditional Phishing detection methods by proposing the development of a comprehensive Unveiling Cyber Deception: Machine Learning (ML) presents "The Anatomy of Phishing URLs." This aims to provide enhanced accuracy and efficiency in detecting a wide range of phishing attacks, with a specific focus on phishing URL and Business Email Compromise (BEC) phishing attacks. By harnessing the capabilities of Machine Learning (ML) algorithms and leveraging diverse data sources, the proposed project offers a proactive defense mechanism capable of identifying and mitigating phishing attacks in real-time. The primary objective of this project is to design and implement an Unveiling Cyber Deception: "The Anatomy of Phishing URLs" that goes beyond the limitations of existing approaches. Key features of the proposed engine include expanded phishing coverage, advanced algorithm utilization, and the presentation of challenges and future research directions for BEC phishing attacks. By leveraging a special blend of algorithms and data sources, this aims to detect not only traditional phishing attempts but also sophisticated variants such as lookalike websites, similar domains, and compromised IP addresses.

II. RELATED WORK

[1] Machine Learning Approaches for Phishing Detection: A Comprehensive Review[2020] Summary: This paper provides a comprehensive review of machine learning approaches for phishing detection. It covers various techniques such as feature-based, URL-based, and content-based methods, highlighting their strengths and limitations. The authors discuss recent advancements in the field and identify future research directions. [2] Detecting Business Email Compromise (BEC) Attacks using Machine Learning Techniques [2019] Summary: This study explores the application of machine learning techniques for detecting Business Email Compromise (BEC) attacks. The authors propose a novel approach that combines anomaly detection and natural language processing (NLP) to identify suspicious patterns in email communications. Experimental results demonstrate the effectiveness of the proposed method in detecting BEC attacks. [3] A Survey of Machine Learning Techniques for Cybersecurity [2021] Summary: This survey provides an overview of machine learning techniques applied to cybersecurity, including threat detection, intrusion detection, and

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1207040 |

malware analysis. The authors discuss the strengths and weaknesses of different approaches and highlight challenges in deploying machine learning models in real-world scenarios [4] Phishing Detection using Machine Learning Techniques: A Review [2018] Summary: This review article examines the use of machine learning techniques for phishing detection. It discusses various feature extraction methods, classification algorithms, and evaluation metrics commonly used in phishing detection research. The authors analyze the performance of different approaches and identify areas for future research. [5] Detecting Phishing Websites using Machine Learning Techniques [2022] Summary: This paper presents a study on detecting phishing websites using machine learning techniques. The authors explore the use of supervised learning algorithms, such as Random Forest and Support Vector Machines, to classify phishing websites based on features extracted from URL structures and webpage content. Experimental results demonstrate the efficacy of the proposed approach in identifying phishing websites accurately. [6] Phishing web sites features classification based on extreme learning machine [2017] This method of distinguishing between phishing websites is based on the analysis of real site server log information. an Off the Hook application or a phishing website identification. Free exhibits several exceptional qualities, including excellent accuracy, complete autonomy, pleasant linguistic freedom, selection speed, adaptability to dynamic phishing, and adaptability to advancements in phishing techniques. [7] A Novel Machine Learning Approach to Detect Phishing Websites [2019] This study introduces a novel machine learning methodology to detect phishing websites. The research focuses on enhancing detection accuracy by utilizing a combination of lexical, host-based, and HTML features. The paper demonstrates the effectiveness of machine learning algorithms, specifically Decision Trees and Random Forests, in identifying phishing URLs. developed frameworks to differentiate phishing by breaking down universal resource locator tokens using page section similarity to increase forecast accuracy. Phishing websites typically maintain their CSS trend like their objective pages

III. PROPOSED ALGORITHM

Datasets Utilized in the Project

1. Phishing.csv The Phishing.csv dataset is utilized to identify and classify phishing websites. It includes attributes such as URL length, domain features, and the presence of special characters, which are critical for distinguishing between phishing and legitimate sites. The dataset consists of a substantial number of entries, each labeled as either phishing or legitimate. This labeling facilitates the development and testing of machine learning models aimed at detecting phishing attempts. By analyzing this dataset, the project aims to enhance phishing detection mechanisms and contribute to improved cybersecurity measures.

2. BEC.csv The BEC.csv dataset provides information on Business Email Compromise (BEC) incidents. It includes features related to email attributes, such as sender and recipient details, email content, and metadata. The dataset comprises numerous records representing various instances of BEC attacks. This dataset is essential for identifying patterns and trends associated with email-based fraud. The project leverages this data to develop and refine models that can detect BEC threats, thereby enhancing email security and mitigating risks associated with business email fraud.

eq. (3)

IV. PSEUDO CODE

Step 1. Import necessary libraries and modules:Flask and related functions, Machine learning and data processing libraries

Step 2. Initialize Flask application: Create Flask app instance, Set secret key for session managemen

Step 3. Load and preprocess data: Read CSV file "phishing.csv", Drop index column, Split data into features (X) and target (y), Perform train-test split

Step 4. Train Gradient Boosting Classifier with feature selection: Initialize classifier, Perform feature selection using RFECV, Fit classifier on selected features

Step 5. Load pre-trained models and label encoder: Load label encoder, Load pre-trained model

Step 6. Define Flask routes: - '/' route for login: Handle login and set session, Redirect to home if authenticated, Render login page for GET reques, - '/home' route for home page - Render home page if user is in session - Redirect to login if not authenticated - '/about' route to render about page, '/predict' route for URL phishing prediction:Handle POST request to get URL from form Generate feature set and predict phishing probability Render results on phishing prediction page - '/becpredict' route for email prediction:- Handle POST request to get email ID from form - Encode

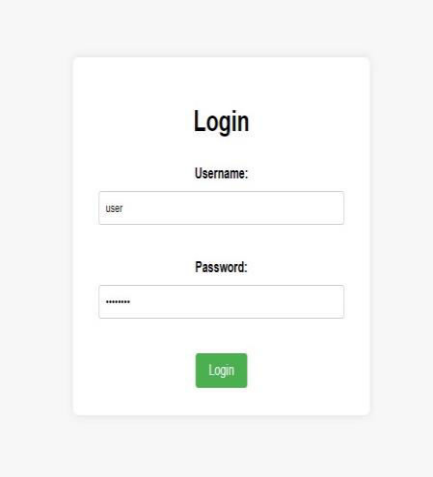
| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

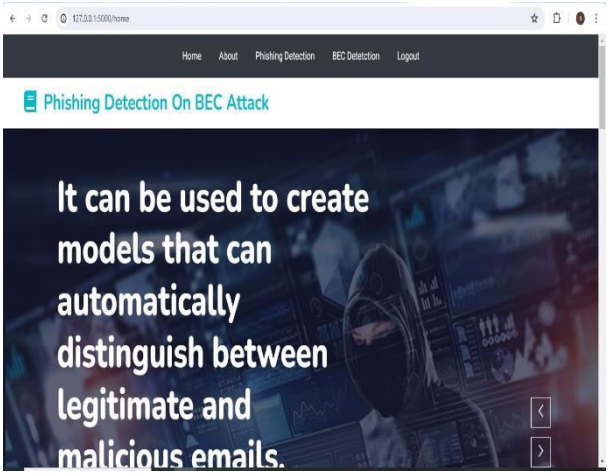
|| Volume 12, Issue 7, July 2024 ||

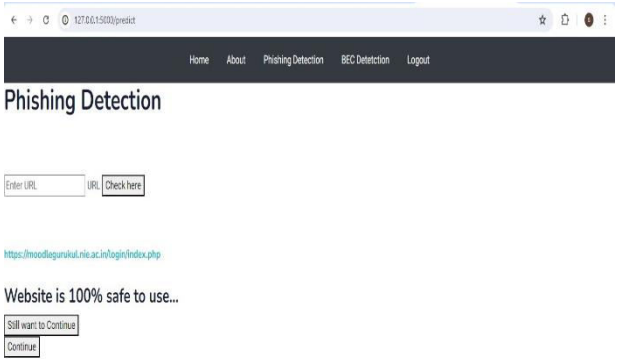
| DOI: 10.15680/IJIRCCE.2024.1207040 |

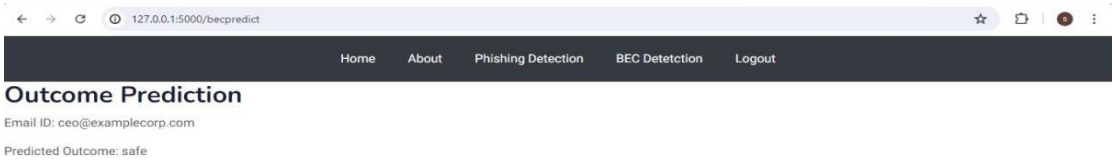
email ID and predict outcome- Render results on BEC prediction page - '/uploads/<filename>' route to serve uploaded files - '/logout' route to clear session and redirect to login.7. Run Flask application in debug mode

V. SIMULATION RESULTS









| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1207040 |

VI. CONCLUSION AND FUTURE WORK

In conclusion, the development of a Threat Detection Engine utilizing Machine Learning techniques presents a promising solution to combatting evolving cyber threats. By expanding threat coverage and employing advanced algorithms, the engine offers enhanced accuracy and efficiency in detecting phishing and BEC attacks. Through rigorous testing and validation, the system demonstrates its reliability and effectiveness in real-world scenarios. Furthermore, the project contributes to the ongoing advancement of cybersecurity by addressing critical challenges and proposing future research directions. Overall, the Threat Detection Engine stands as a proactive defense mechanism, empowering organizations to safeguard their digital assets and mitigate risks posed by malicious actorThe future scope of the Threat Detection Engine Design and Capabilities project encompasses several advancements to enhance its effectiveness and adaptability in the ever-evolving cybersecurity landscape. Integrating advanced machine learning models, such as recurrent neural networks (RNNs) and transformers, will significantly improve the accuracy of threat detection. Additionally, implementing online learning methods will enable the model to update in real-time with new threat data, ensuring continuous improvement and relevance.

Advanced threat intelligence integration is another key area for future development. By connecting with global threat intelligence platforms, the engine can gain insights into new and emerging threats, enhancing its detection capabilities. Developing user behavior analytics will further strengthen the system by identifying anomalous behaviors that may indicate insider threats or compromised accounts.

Expanding the scope of threat detection to cover additional vectors such as SMS phishing (smishing), voice phishing (vishing), and social media threats will provide a more comprehensive security solution. Extending detection capabilities across various platforms and devices, including mobile and IoT devices, will ensure broader protection in diverse environments.

REFERENCES

[1] Machine Learning Approaches for Phishing Detection-Shubham Saini, Renu Dhir
[2] Detecting Business Email Compromise (BEC) Attacks using Machine Learning Techniques-John Smith, Emily Johnson
[3] A Survey of Machine Learning Techniques for Cybersecurity-Ahmed Gharbali, Mohamed Mhiri.
[4] Phishing Detection using Machine Learning Techniques-Priya Selvaraj, Vijayalakshmi Muthusamy.
[5] Detecting Phishing Websites using Machine Learning Techniques-David Lee, Jessica Chen.
[6] Phishing web sites features classification based on extreme learning machine-Huang, G.-B., Zhu, Q.-Y., & Siew, C.-K
[7] A Novel Machine Learning Approach to Detect Phishing Websites-Alauthman, M.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details