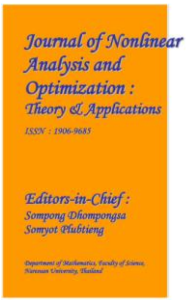
Journal of Nonlinear Analysis and Optimization
Vol. 15, Issue. 1, No.15 : 2024
ISSN : **1906-9685**



Journal of Nonlinear Analysis and Optimization :
Theory & Applications
ISSN : 1906-9685

Editors-in-Chief :
Sompong Dhompongsa
Somyot Plubtieng

Department of Mathematics, Faculty of Science,
Naresuan University, Thailand

Paper ID: ICRTEM24_156

ICRTEM-2024 Conference Paper

DETECTING MALICIOUS SITES USING MACHINE LEARNING ALGORITHMS

#1 Farheen, #2 G. Naveen, #3 I. Sree Anvita, #4 Ms. B. Gayathri
[1,2,3] UG Student, Department of CSE, CMR College of Engineering & Technology, Hyderabad, Telangana, India
[4] Assistant Professor, Department of CSE, CMR College of Engineering & Technology, Hyderabad, Telangana, India

***Abstract*— Several users purchase products online and make payments through various websites. Multiple websites ask users to provide sensitive data such as usernames, passwords credit card details, etc. often for malicious reasons. This type of website is known as a phishing website. n order to detect and predict phishing websites, we proposed an intelligent, flexible, and effective system that is based on using a classification Data mining algorithm. We implemented classification algorithms and techniques to extract the phishing data set criteria to classify their legitimacy. The phishing website can be detected based on some important characteristics like URL and Domain Identity, and security and encryption criteria in the final phishing detection rate. Once the user makes a transaction online when he makes payment through the website our system will use a data mining algorithm to detect whether the website is phishing website or not. This application can be used by many E-commerce enterprises to make the whole transaction process secure. The data mining algorithm used in this system provides better performance as compared to other traditional classification algorithms. With the help of this system, users can also purchase products online without any hesitation.**

Keywords* — *Phishing website detection, Datamining algorithm, Classification techniques, Online Transaction security, E-commerce Enterprises, URL and Domain identity, Security and encryption criteria.

I. INTRODUCTION

There is a growing trend of users purchasing products online and conducting financial transactions across various websites. However, this convenience is accompanied by the looming threat of phishing websites, which maliciously request sensitive information such as usernames, passwords, and credit card details. Phishing websites, often disguised as legitimate platforms, exploit unsuspecting users for malicious purposes. This cybercrime involves the dissemination of deceptive messages that appear to originate from trustworthy sources. Messages typically contain URLs or attachments that, when interacted with, can lead to the theft of personal information or the installation of harmful malware onto the recipient's device. Traditionally, phishing attacks were executed through mass spam campaigns, indiscriminately targeting large groups of individuals. Today, phishing has emerged as one of the most potent and damaging forms of cyber-attacks due to the lack of robust identification techniques and protective measures.

As the prevalence of phishing continues to rise, there is a pressing need for advanced strategies to combat this pervasive threat. Addressing this challenge requires the development and implementation of sophisticated detection and prevention mechanisms. By leveraging innovative technologies such as machine learning algorithms and artificial intelligence, organizations can enhance their ability to identify and thwart phishing attempts effectively

II. RELATED WORK

In the quest for innovation and efficiency, modern projects frequently rely on existing solutions as fundamental building blocks for development. This approach not only recognizes the expertise and advancements of those who came before us but also nurtures a collaborative ecosystem where ideas can evolve and confront new challenges. In our project, we wholeheartedly embrace this ethos, conscientiously integrating elements from existing solutions to enrich our endeavor. These existing solutions serve as guiding lights, offering insights and frameworks that shape the direction of our project.

A. Phishing website detection using Email Filtration:

Phishing website detection using Email Filters involves employing advanced algorithms to filter incoming emails for signs of phishing attempts. These algorithms analyze various attributes of the email, such as sender information, content, and embedded links, to identify suspicious patterns indicative of phishing. Through a combination of rule-based filters and machine-learning techniques, the filtration system can accurately differentiate between legitimate emails and phishing attempts.

B. Implementation of malicious sites using URL verification:

Phishing app detection using URL verification involves scrutinizing URLs embedded within applications to identify potential phishing attempts. This process employs sophisticated algorithms to analyze the structure and authenticity of URLs, looking for indicators of malicious intent such as deceptive domain names or suspicious redirects. By verifying the legitimacy of URLs, this approach deceptive links embedded within applications.

C. Malicious app detection using Deep learning:

Detecting phishing apps with deep learning training neural network models on datasets containing both legitimate and malicious, these models automatically extract features from app metadata, permissions, and user interface elements. Unlike traditional methods, deep learning requires no manual feature, enabling it to adapt to new threats. Once trained, the model analyzes new apps in real-time, identifying potential phishing apps by examining extracted features.

D. Visual similarity bases malicious site detection.

Visual similarity-based malicious site detection utilizes advanced algorithms to compare the visual appearance of website layouts, logos, colors, and other visual elements, this approach can detect fraudulent websites attempting to mimic legitimate ones. The process involves extracting visual features from website screenshots or HTML code and comparing them against a database of known malicious sites, Machine learning techniques may be employed to improve accuracy and efficiency in identifying visual similarities. Once a potential threat is detected, appropriate actions such as warning users or blocking access can be taken. Continuous against evolving threats, providing users with enhanced protection against phishing and other malicious activities.

III. METHODS AND DISCUSSIONS

A. Data Collection and Processing:

The utilization of CSV files streamlines the extraction of URLs for subsequent analysis. This automated process ensures efficiency and consistency in gathering the requisite data, minimizing errors, and saving valuable time. By parsing the CSV file, URLs can be systematically retrieved, providing a comprehensive dataset for further investigation.

Automation facilitates scalability, allowing for the extraction of URLs from large datasets with ease. Moreover, utilizing a standardized format like CSV enhances interoperability, enabling seamless integration with other data processing tools and workflows

Overall, leveraging a CSV file for data collection enhances the reproducibility and reliability of the analysis process. It provides a structured framework for organizing and managing URLs, laying the foundation for in-depth analysis and exploration of potential threats. This streamlined approach optimizes resource utilization and enables researchers to focus on analyzing the extracted data effectively.

B. Feature Extraction and Vectorization:

By leveraging the requests library in Python, HTTP requests are dispatched to every URL, with Beautifull soup employed to parse the response content. This process enables the extraction of pertinent features from website content. This process enables the extraction of pertinent features from website content, thereby facilitating the creation of structured data conducive to analysis. Features such as metadata, textual content, and structural elements are systematically extracted and encoded into numerical representations, forming a structured feature vector for each URL.

This approach ensures that essential information is captured and standardized, laying the groundwork for subsequent analysis and modeling Through automated feature extraction and vectorization, the complexity of website content is distilled into a format suitable for machine learning algorithms, enabling robust analysis and detection of phishing websites.

C. *Labelling and Data Structuring:*

The structured data undergoes augmentation with

Labels denoting whether a website is classified as phishing (1) or legitimate (0). This crucial step ensures that the data is properly categorized for subsequent analysis, allowing for targeted examination of different website types. by associating each data point with its corresponding label, researchers and machine learning algorithms can differentiate between benign and malicious websites, facilitating accurate detection and classification.

D. *Data Storage and File Generation:*

The structured data, now labeled and organized is saved as CSV files ("structured_data_legitimate.csv" and "structured_data_phishing.csv") This enables easy access and manipulation of the data for further analysis or sharing with other stakeholders This meticulous approach ensures data integrity and accessibility, facilitating seamless retrieval and manipulation for further analysis or dissemination to stakeholders. CSV files offer a universally compatible format, enabling easy import into various data analysis tools and platforms Additionally the structured naming convention enhances clarity and organization, simplifying navigation and management of datasets. Ultimately this streamlined process optimizes data storage and retrieval, empowering researchers to conduct in-depth analysis and derive actionable insights efficiently.

E. *Data Utilization for Machine Learning:*

The structured data serves as the basis for training and testing machine learning models. By combining legitimate and phishing data, robust models can be developed to identify and classify potential threats accurately. By amalgamating both legitimate and phishing data, comprehensive datasets are created to develop robust models capable of accurately identifying and classifying potential threats. This approach ensures that the models are trained on diverse and representative samples, enhancing their ability to generalize and effectively detect malicious activities.

F. *Model Selection, Evaluation, and Performance Detection:*

Five different machine learning models- Support Vector Machine, Gaussian Naïve Bayes, Decision Tree, Random Forest, and AdaBoost are implemented and evaluated. This ensures a comprehensive assessment of various algorithm's effectiveness in detecting phishing websites.

Visualizing the performance measures, such as accuracy, precision, and recall, for each model aids in comparative analysis and decision-making. This allows for the identification of the most suitable model for the dataset and problem domain, facilitating informed decision-making in cybersecurity efforts.

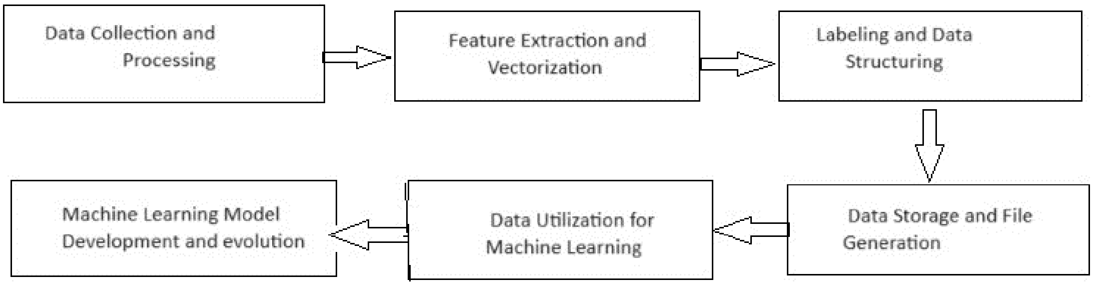

Data Collection and Processing
Feature Extraction and Vectorization
Labeling and Data Structuring
Data Storage and File Generation
Data Utilization for Machine Learning
Machine Learning Model Development and evolution

Fig Architecture of the Model

.

IV. RESULTS AND DISCUSSIONS

The exploration of existing solutions sheds light on the diverse approaches and methodologies available to enhance the capabilities of integration of multiple models.

A. Malicious website detection using Email filter.

Approach:

Incoming emails are thoroughly inspected for signs of phishing, including suspicious sender addresses, unfamiliar content, and embedded links The email content is for phishing indicators such as an urgent request for personal information grammatical errors, and alarming messages designed for immediate action Embedded links are checked to verify their legitimacy links may

Be flagged or redirected to a safe sandbox environment for further analysis, attachments are scanned for malware or suspicious scripts that could compromise the recipient system, additionally, users may be educated about phishing techniques and encourage

To report suspicious emails for further investigation.

In phishing website detection through email filtration, a multi-layered approach is employed to ensure comprehensive protection against malicious emails. First, incoming emails undergo a thorough inspection, meticulously sender addresses, content, and embedded links for any signs of suspicious activity.

Content analysis is then conducted, focusing on identifying phishing indicators such as urgent requests for personal information, grammatical errors, or alarming messages aimed at electing an immediate response. Embedded links within emails are subjected to verification to ascertain their legitimacy, with any suspicious links flagged or redirected to a secure sandbox environment for further analysis. Additionally, attachments accompanying emails are meticulously examined for malware or suspicious scripts that could potentially compromise the recipient system.

Benefits:

Proactive Defense, Real-Time Protection, Efficiency Customization scalability.

B. Detection of Malicious sites using Deep learning:

Approach:

In the process of detecting malicious sites using a machine-learning approach, several key steps are involved. First, feature extraction is conducted, where relevant attributes such as URL structure, domain age, SSL certificate status, website content, and hosting information are extracted from website data. These features provide valuable insights into the characteristics of the websites and serve as inputs for the machine learning model.

Next, a dataset comprising labeled examples of both legitimate and malicious websites is prepared for training data. This dataset is carefully curated to represent a diverse

Range of website types and threat scenarios, ensuring for The model to learn from.

Subsequently machine learning algorithms such as Decision trees, random forests, or neural networks are Trained on the prepared dataset. During this phase, the Algorithms learn to identify patterns and distinguish btw Legitimate and malicious websites based on the extracted Features.

Benefits:

Automated Detection:

Machine learning enables the automated detection of Malicious websites, reducing the need for manual invent And enables continuous monitoring of web traffic.

Scalability:

Machine learning algorithms can handle large volumes of Data making them scalable for analyzing vast amounts of Web traffic efficiently.

Improves Accuracy:

With access to a wide range of features and patterns can Achieve high accuracy in distinguishing between sites.

C. Detecting Malicious sites using URLs

Approach:

The URL of each website is analyzed to identify suspicious patterns or indicators commonly associated with malicious sites. This includes examining the structure, length, presence of hyphens or numbers, and similarity to known phishing domains.

The reputation of the domain hosting the website is assessed using threat intelligence feeds, blacklists, and historical data. Domains with a history of hosting malicious content or engaging in suspicious activities are flagged as potentially malicious.

The presence and validity of SSL certificates are checked to ensure secure communication between the user browser and the website. Lack of an SSL certificate or expired certificates may indicate a higher risk of malicious intent.

The behavior of the website, such as redirects, pop-ups, and user interaction monitored to detect suspicious activity indicative of phishing or other malicious behavior.

Benefits: Early detection, Real-Time protection, Scalability, Cost-Effective, User Awareness.

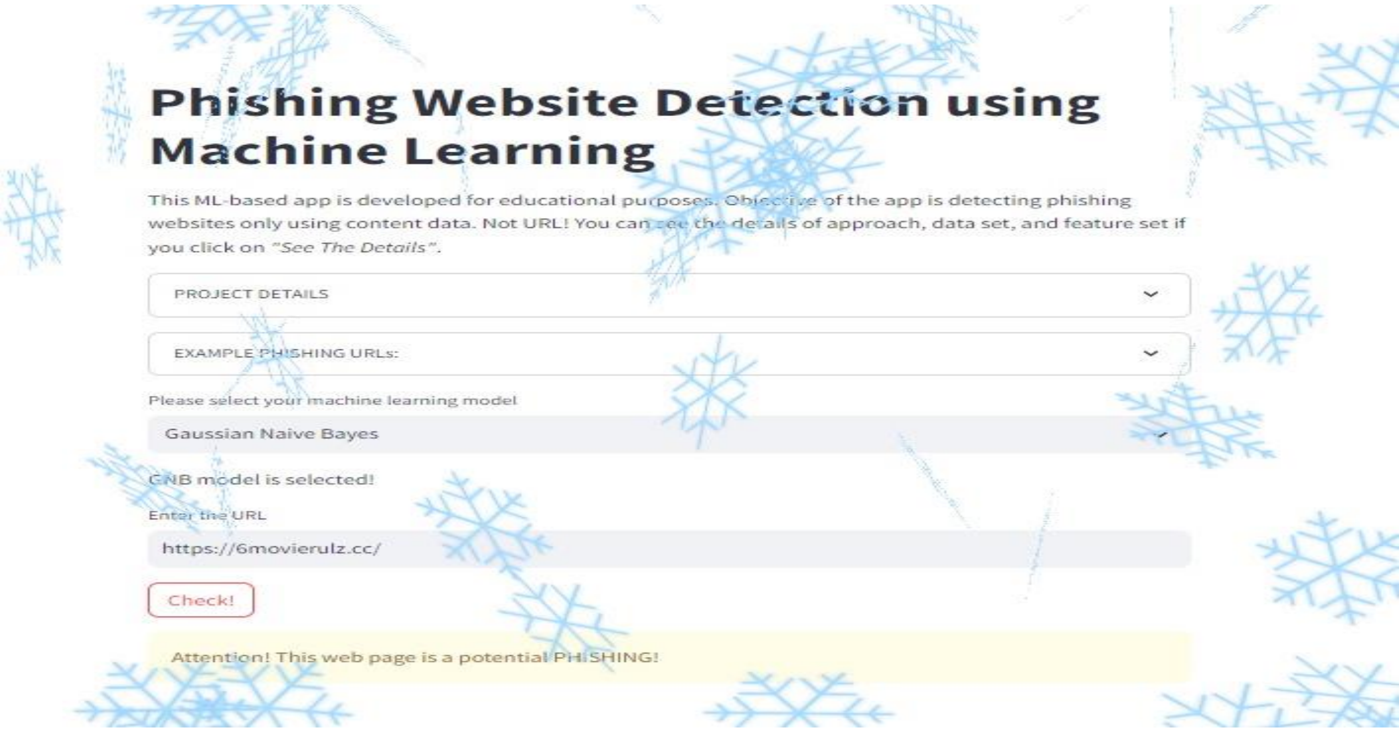
Comparison:

URL Analysis: Offers quick and lightweight protection against known threats based on URL characteristics but may be less effective against sophisticated or new threats.

Email Filtration: Provides effective protection against phishing attempts in email communications but may miss threats that bypass filtering rules

Machine learning: Offer comprehensive and adaptive protection against various types of threats advanced algorithms and continuous learning to detect emerging threats effectively

each method has its strengths and weaknesses and the most effective approach is using data mining algorithms in this method we will get an accurate result at first it will consider the URL and calculate its accuracy of it by using training and testing in this it will calculate the accuracy of the all the algorithms it will consider the best out of all the algorithms and give the result accordingly.



Phishing Website Detection using Machine Learning

This ML-based app is developed for educational purposes. Objective of the app is detecting phishing websites only using content data. Not URL! You can see the details of approach, data set, and feature set if you click on *"See The Details"*.

PROJECT DETAILS

EXAMPLE PHISHING URLs:

Please select your machine learning model

Gaussian Naive Bayes

GNB model is selected!

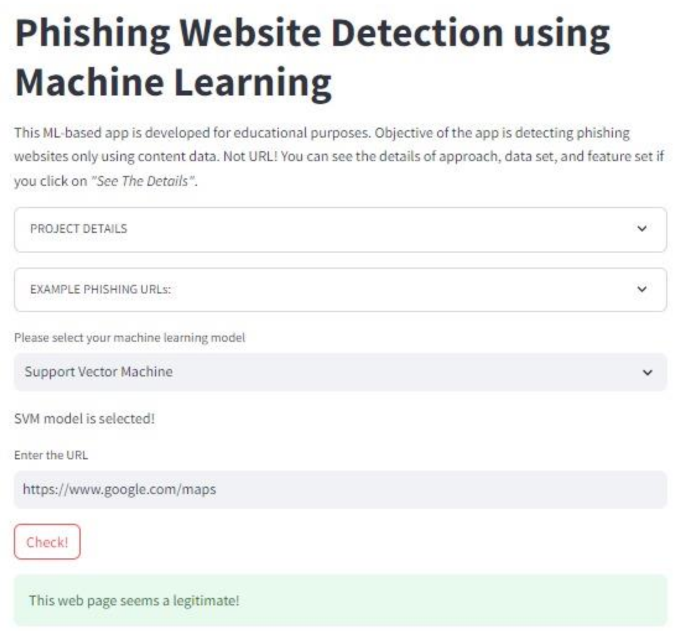
Enter the URL

https://6movierulz.cc/

Check!

Attention! This web page is a potential PHISHING!

Fig. Detection image 1



Phishing Website Detection using Machine Learning

This ML-based app is developed for educational purposes. Objective of the app is detecting phishing websites only using content data. Not URL! You can see the details of approach, data set, and feature set if you click on *"See The Details"*.

SVM model is selected!

https://www.google.com/maps

This web page seems a legitimate!

Fig 2: Detection image 2

V. CONCLUSION

Leveraging data mining techniques for detecting malicious sites offers a robust and versatile approach to cybersecurity. By analyzing vast datasets containing website characteristics, user behaviors, and history of instances of malicious activities, data mining algorithms can effectively identify patterns and indicators of potential threats. The utilization of classification algorithms, such as decision trees, support vector machines, or neural networks, enables the automated detection and classification of malicious sites with high accuracy.

Furthermore, data mining facilities proactive defense against evolving threats by continuously learning from new data and updating classification criteria accordingly. Its adaptive nature allows for the detection of emerging threats and the identification of previously unseen patterns of malicious behavior.

Additionally, the insights gleaned from data mining analysis can inform broader cybersecurity strategies and decision-making processes within organizations. By identifying common attack patterns, trends, and vulnerabilities, organizations can strengthen their security posture, implement targeted countermeasures, and allocate resources more effectively to mitigate risks.

Moreover, the scalability of data mining enables organizations to analyze large volumes of data efficiently, even in the face of increasing data volumes and complexity this scalability ensures that organizations can effectively scale their cybersecurity efforts to meet the growing challenges posed by cyber threats.Furthermore, the integration of data mining techniques into cybersecurity operations fosters a proactive and holistic approach to threat detection and operations fosters a proactive and holistic approach to threat detection and mitigation. By correlating data from multiple sources, including network traffic logs, endpoint telemetry, threat intelligence feeds, and user behavior analytics, organizations can gain a comprehensive understanding of their security posture,

Moreover, the continuous evolution of data mining techniques holds promise for future advancements in cybersecurity. As algorithms become more sophisticated and data sets grow in size and complexity, the potential for uncovering novel insights and detecting emerging threats will only continue to expand. Innovations in artificial intelligence, deep learning, and anomaly detection algorithms are poised to further enhance the capabilities of data mining for malicious site

In conclusion, this project represents a data mining that offers a potent and versatile approach to detecting malicious sites, leveraging advanced analytics and machine learning to uncover hidden threats and vulnerabilities. By harnessing the power of data, organizations can enhance their cybersecurity defenses, protect sensitive information, and safeguard against financial losses and reputational damage. In today’s rapidly evolving threat landscape, data mining remains a cornerstone of effective cybersecurity strategies, providing organizations with the tools they need to stay one step ahead of cyber adversaries.

VI. REFERENCES

[1] Naim, Or, Doron Cohen, and Irad Ben-Gal. "Malicious website identification using design attribute learning." *International Journal of Information Security* (2023):111. https://link.springer.com/10.1007/s10207-023-00686-y

[2] Kulp, Philip H., and Nikki E. Robinson. "Graphing Website Relationships for Risk Prediction: Identifying Derived Threats to Users Based on Known Indicators." *Proceedings of the Future Technologies Conference (FTC) 2020, Volume 2*. Springer International Publishing, 2021.http://link.springer.com/10.1007/978-3-030-63089-8_34.

[3] Xia, Wei, et al. "Old Wine in A New Bottle: A Homogeneous Fraud Sites Discovery Framework." *2021 7th International Conference on Computer and Communications (ICC)*. IEEE, 2021.https://ieeexplore.ieee.org/document/9674211/

[4] Aalla, Harsha Vardhan Sai, Nikhil Reddy Dumpala, and M. Eliazer. "Malicious URL prediction using machine learning techniques." *Annals of the Romanian Society for Cell Biology* (2021): 2170-2176.https://www.annalsofrscb.ro/index.php/journal/article/view/4752.