IJIRCCE

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

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 3, March 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

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 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1203173 |

URL Based Phishing Website Detection Using Machine Learning

Esther Lipni [1], Kalluri Lakshmi Prasanna [2], Kancharla Bhavya [3], Koduru Supriya [4]

Assistant Professor, Department of Computer Science and Engineering, Prathyusha Engineering College, Thiruvallur, Tamil Nadu, India[1]

Department of Computer Science and Engineering, Prathyusha Engineering College, Thiruvallur, Tamil Nadu, India[2]

Department of Computer Science and Engineering, Prathyusha Engineering College, Thiruvallur, Tamil Nadu, India[3]

Department of Computer Science and Engineering, Prathyusha Engineering College, Thiruvallur, Tamil Nadu, India[4]

ABSTRACT: In this project, our main factor is a website, whether it is a fraudulent one or a legit one. Conventionally, a website can be detected whether it is harmful or not by the browser protection service, i Even though, the browser's firewall is enabled, it can never detect a phishing website. Because, Phishing site is not malicious site it steals data without the user even knowing it. So, to detect such sites we are training an ML model using different algorithms to determine the phishing site based on URL feature extraction. Based on different features of URL, such as like Domain length, character length etc., we will train the model with one algorithm at a time store their results and compare to find the more accurate one and display the results using the approved algorithm. Detecting such fraudulent websites is crucial to safeguarding users and organizations from financial and data loss. This paper proposes a novel approach for URL-based phishing website detection, leveraging machine learning techniques and feature engineering to analyze various characteristics of URLs.

KEYWORDS: Phishing, Machine learning, Feature extraction.

I. INTRODUCTION

Background

Phishing is a type of extensive fraud that happens when a malicious website act like a real one keeping in mind that theend goal to obtain touchy data, for example, passwords, account points of interest, or MasterCard numbers. phishes think of new and half breed strategies to go around the accessible programming and systems.

Challenges in Traditional Agriculture

In spite of the fact that there are a few contrary to phishing programming and methods for distinguishing potential phishing endeavours in messages and identifying phishing substance on sites, phishes think of new and half breed strategies to go around the accessible programming and systems.

Phishing is a trickery system that uses a blend of social designing what's more, innovation to assemble delicate and individual data, for example, passwords and charge card subtle elements by taking on the appearance of a dependable individual or business in an electronic correspondence.

Phishing is a trickey system that uses a blend of social designing what's more, innovation to assemble delicate and individual data, for example, passwords and charge card subtle elements by taking on the appearance of a dependable individual or business in an electronic correspondence.

URL-based phishing website detection using machine
learning involves gathering a dataset containing both legitimate and phishing URLs, labeling them accordingly. Features such as URL length, domain age, Presence of suspicious characters, subdomains, and HTTPS encryption status are extracted.

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

|| Volume 12, Issue 3, March 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1203173 |

II. OBJECTIVE

The objective of this project is to develop a ML model that to detect the phishing websites with accuracy. Expected steps and procedures to be followed to fulfil the objectives

The intention of this task is to expand a gadget learning model that can correctly discover phishing web sites. Expected steps and strategies to achieve the goals:
Load the dataset into the model and extract attributes from all URLs. Set facts without values and null values.

Visualize the statistics to find commonalities inside the features considered in the chart. Train your ML models using unique algorithms and evaluate their consequences up-to-date.

Compare all modules to determine the maximum accurate algorithm for the identity method. Hook to the end result you want to discover with the quest web page and arrange your statistics to the right.

III. EXISTING SYSTEM

A. Challenges Addressed by MHSA:

Complexity: MHSA involves complex computations which could be resource intensive,especially when processing large datasets of website features.

Training Data Requirement: Effective utilization of MHSA typically requires large amount of labeled training data.Due to the dynamic nature of phishing attacks it became challenging Feature Representation: MHSA may struggle effectively represent certain type of features commonly used in phishing website detection such as visual elements and contextual information.

Generalization: The ability of MHSA to generalize to unseen phishing attacks or variations in attack techniques maybe limited. Making it challenging for algorithms to keep pase and accurate the detect new phishing attempts.

B. Challenges Addressed by CNN:

Limited contextual information: CNN process input data in fixed-size windows, which may not capture the entire context of a URL. Phishing detection often requires understanding the context of the URL, such as the domain name, subdomains, and path structure, which CNN might not handle well.

Adversarial attacks: CNN are susceptible to adversarial attacks, where small, imperceptible perturbations to the input can cause missclassification. Phishing attackers can leverage this vulnerability to craft URLs that evade detection by CNN-based models.

Generalization issues: CNN trained on a specific dataset may not generalize well to unseen phishing URLs, especially if the distribution of phishing URLs changes over time. Continuous updates and retraining of the model are necessary to maintain effectiveness.

IV. PROPOSED SYSTEM

A. Innovative Features:

Here we are developing our model with few techniques:

1. URL analysis Module:

This module examines the structure and components of the URL to identify any anomalies or patterns commonly associate with phishing such as misspelled domain names , long urls with multiple subdomains or the presence of ip addresses instead of domain names.

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

|| Volume 12, Issue 3, March 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1203173 |

2. Page content analysis Module:

Scan the webpage content for phishing elements such as fake login forms, requests for sensitive information, or misleading content.

3. Visual similarity Detection Module:

Utilizes image processing techniques to compare the visual appearance of the suspected website with known legitimate websites to identify visual spoofing attempts.

4. Phishing Kit Detection Module:

Phishing websites, such as clones of popular Identifies common phishing toolkits and frameworks used to create login pages.

5. User Reporting and Feedback Module:

Allows users to report suspicious websites and provide feedback to improve the detecting system.User feedback will help to understand the drawback of executed system.

6. Real-time Threat Intelligence Feeds Module:

A Integrates with threat intelligence feeds to receive real-timeupdates on emerging phishing threats and malicious domains.

B. Addressing Existing System Limitations:

The proposed system directly addresses the limitations identified in the existing system:

1. Wrapping Technique:

wrapping technique which is a powerful and flexible technique for feature selection and capture interaction between features.

2. Class Balancing:

In order to reduce the datasets processing time we are using “class balancing” technique which helps to balance number of rows and number of columns in both phishing and non phishing datasets.

3. Distributed Oriented Analysis:

Distributed oriented analysis helps to reduce the processing time of datasets.It describes how data is clustered which means group the entire similar features.

V. CONCLUSION

Phishing is a cyber crime procedure utilizing both social building and specialized deception to take individual sensitive data. Besides, Phishing is considered as another extensive type of fraud.
Experimentations against recent dependable phishing data sets utilizing different classification algorithm have been performed which received different learning methods. The base of the experiments is accuracy measure.
The aim of this research work is to predict whether a given URL is phishing website or not.As a future work we might use this model to other Phishing dataset with larger size then now and then testing the performance of those classification algorithm’s in terms of classification accuracy.

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

|| Volume 12, Issue 3, March 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1203173 |

VI. FUTURE WORK

A. Behavioral Analysis:

- Incorporating behavioral analysis techniques can help in identifying anomalies in user behaviour that may indicate a phishing attack. Analysing mouse movements,keystrokes, and browsing patterns can provide valuable insight into whether a website is real or malicious.

B. Real-Time Monitoring:

- Implementing real-time monitoring systems that continuously scan websites for suspicious activity can help in detecting phishing websites as soon as they are created or modified.

C. Implementation of Farmer Feedback Mechanism:

- User Feedback Integration: Implementing a system for users to provide feedback on the website performance and suggest improvements. This user- centric approach ensures that created web evolves based on the actual needs and experiences of its users.

REFERENCES

[1] Katerya Burbela, Oleksii Baranovskyi, “Model of detection of phishing URLs Based on machine learning, ” May 2023.
[2] FAmogh J, Rajendra KN, Deekshith NG, S Parikshith, Asst. Prof. Suma L,“Malicious URL detection using mchine learning, ” IRJMETS International research journal of Modernization in Engineering Technology and Science , August 2022.
[3] Shantanu, B Janet, R Joshua Arul Kumar, “Malicious URL Detection” IEEE May 2021
[4] Ankesh Anand ,Kshitij Gorde,Joel Ruben Antony Moniz,Noseong Park,Tanmoy Chakraborty,Bei-Tseng Chu, “Phishing URL Detection with Oversampling based on Text Generative Adversarial Networks,” IEEE International Conference on Big Data (Big Data), December 2018.
[5] Muhammad Teaseer Sulemal,Shahid Mahmood Awan, “Optimization of URL Based phishing Websites Detection Through Genetic Algorithms” , April 2019.
[6] Mahdieh Zabihimayvan, Derek Doran, “Fuzzy Rough Set Feature Selection to Enhance Phishing Attack Detection,” International Conference on Fuzzy Systems (FUZZ-IEEE), New Orleans, LA, USA, June 2019.
[7] Abdul karim, Mobeen Shahroz, Khabib Mustofa, Samir Brahim Belhavouri, S Ramana Kumar Joga, “Phishing detection system hybrid machine learning based on URL” IEEE january 2019.
[8] Qasem Abu AI-Haija, Ahmad AI Badawi, “URL-based Phishing website detection” IEEE,April 2021.
[9] Adarsh Mandadi, Saikiran Boppana, Vishnu Ravella, R Kavitha, “Phishing website Detection Using Machine Learning” IEEE , May 2022.
[10] SK Hasane Ahammad, Sunil D.Kale, Gopal D.Upadhye, Sandeep Dwarakanth Pande, E. Venkatesh Babu, Amol V. Dhumane, Mr. Dilip Kumar Jang Bahadur, “Phishing URL Detection Using Machine Learning Methods” November 2022.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.379

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

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