IJIRCCE

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

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 5, May 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 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205295 |

Detection of Phishing Websites Using Machine Learning

Prof. A.N. Kalal, Sugat Sunil Ingle, Prince Kumar Anil Gupta, Shreyash Sadashiv Bhole, Kamlesh Nandkumar Janawale

Department of Information Technology, Anantrao Pawar College of Engineering & Research, Pune, India

ABSTRACT: The emergence of phishing websites presents a significant challenge in the digital landscape, as they masquerade as authentic platforms to illicitly acquire sensitive user data. In response, machine learning methods have become increasingly prominent for countering such threats. This study proposes a novel approach centered around the utilization of extreme learning machine (ELM) algorithms to bolster the detection of phishing websites. These algorithms, especially when integrated with version information, have demonstrated remarkable efficacy in discerning fraudulent attempts. Given the diverse array of features exhibited by web pages, a varied set of attributes is imperative for effectively identifying phishing endeavors. While specifics regarding the dataset employed are undisclosed, it encompasses four principal categories of URL attributes: domain, address, base anomaly, and HTML/JavaScript attributes. These attributes play a pivotal role in the accurate identification of phishing attempts and the development of resilient defense mechanisms. Leveraging preprocessed and analyzed data, URL attributes are extracted and associated values are generated. Subsequently, machine learning methodologies, including ELM, are employed to scrutinize these attributes and establish threshold and range values for precise phishing website detection. The overarching objective of this endeavor is to devise an ELM classifier capable of adeptly identifying phishing websites by harnessing the multitude of features within the database. By harnessing machine learning techniques and an exhaustive set of URL attributes, the aim is to fortify cybersecurity protocols and mitigate the hazards posed by online phishing assaults.

KEYWORDS: Extreme Learning Machine (ELM), Support Vector Machine (SVM), Random Forest Algorithm, URL Phishing Websites, Browser add-ons.

I. INTRODUCTION

Phishing is a deceitful tactic wherein malicious actors impersonate reputable entities in online interactions, aiming to fraudulently acquire sensitive data such as usernames, passwords, and credit card details for nefarious purposes. This practice has raised alarm among security professionals due to its simplicity in execution—perpetrators can effortlessly create counterfeit websites that closely resemble legitimate ones. While experts may be adept at identifying these deceptive sites, the average individual often struggles to discern the discrepancy, making them susceptible to phishing attacks. Typically, attackers target financial institution accounts to pilfer login credentials. The financial ramifications of phishing for US businesses are considerable, estimated to amount to approximately $2 billion annually.

In February 2014, Microsoft's third annual Computing Security Index report revealed that phishing's annual global impact could reach or surpass $5 billion. A key factor contributing to the effectiveness of these attacks is the users' lack of awareness, rendering them susceptible to exploitation. This underscores the difficulty in combating phishing and safeguarding users' sensitive information, highlighting the necessity for continually evolving detection strategies

In phishing incidents, attackers create fake web pages mirroring the content of authentic websites, complicating users' ability to distinguish between genuine and fraudulent sites. They commonly exploit social engineering strategies to trick unsuspecting individuals into believing they are engaging with reputable entities. This includes the creation of bogus email accounts and messages to interact with legitimate organizations.

A reliable approach to detecting phishing websites is through the utilization of blacklists containing URLs and IP addresses linked to phishing schemes. These blacklists are commonly incorporated into antivirus databases, serving as a form of "blacklisting" method. Nonetheless, attackers employ deceptive methods to evade detection and circumvent blacklisting. They may employ obfuscation techniques to manipulate URLs, making them appear legitimate. Furthermore, attackers may leverage fast flux hosting and automatically generated proxies to mask their malicious endeavors as part of their evasion strategies.

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

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205295 |

Machine learning presents a promising avenue for effectively identifying and countering phishing websites. A prevalent use case of machine learning lies in the detection of phishing websites, where the goal is to differentiate between genuine and fraudulent sites. This task encompasses various techniques, such as identifying spear phishing and email phishing scams, aimed at discerning malicious entities from legitimate ones.

Therefore, it is crucial for users to acknowledge the potential repercussions of phishing attacks and not rely solely on unofficial sources. Embracing security software driven by machine learning (ML) holds promise in overcoming the deficiencies of current technologies.

Machine learning, a branch of artificial intelligence, operates on implicit programming and possesses the capability to learn autonomously, often without direct supervision. While certain techniques like supervised learning require training models with labeled data, others, such as reinforcement learning, function without explicit instructions. There are numerous machine learning methods available, including:

A. Supervised learning
B. Unsupervised learning
C. Reinforcement learning

II. MACHINE LEARNING ALGORITHMS

Extreme learning machine (ELM):
Extreme learning machine (ELM) is a type of artificial neural network (ANN) model characterized by a single hidden layer. Parameters such as threshold, weight, and activation values within the ANN necessitate suitable settings to facilitate effective learning. To achieve optimal learning outcomes, it is essential to refine the parameters iteratively. This process typically involves employing a gradient-based learning system to adjust the parameters to their appropriate values.

Random forest algorithm:
Random Forests (RF) encompass a set of regression and classification methods widely employed to tackle diverse challenges. In RF, data classification is accomplished through decision trees, wherein numerous trees are built during the training stage, typically predetermined by the programmer. Subsequently, these trees are employed collectively for prediction purposes. Each individual tree autonomously assigns a class label to input data, and the ultimate output is derived by amalgamating the predictions of all trees, often through averaging or voting mechanisms.

SVM (Support Vector Machine):
Random Forests (RF) comprise a collection of regression and classification techniques extensively utilized to address a variety of tasks. Within RF, data classification is achieved via decision trees, where multiple trees are constructed during the training phase, usually specified by the programmer. These trees are then utilized collectively to make predictions. Each tree independently assigns a class label to input data, and the final output is determined by aggregating the predictions of all trees, commonly through averaging or voting procedures.

III. LITERATURE REVIEW

SK Hasane Ahammad , Sunil D. Kate , Gopal D. Upaddhye , Sandeep Dwarkanath Pnade el al , ELSEVIER Jan 1 2022 . [1] a machine learning model can be created with the help of all the algorithms discussed above, and for testing and training, the machine learing model and 80% of the dataset were used for training and 20% for testing. [2] Learning dataset gives Acuuracy measure's by decision tree , randomforest and other algorithms.

Mohammed Hazim Alkawaz, Stephanie Joanne Steven, AsifIqbal Hajamydeen, Detecting Phishing Websites Using Machine Learning, 2020 16th IEEE International Colloquium on Signal Processing & its Applications, 28-29 Feb. 2020 . [1] After reviewing and researching for appropriate monitoring tools, proposed system has been identified and chosen to address the complexity of monitoring requirement for current situation. This software is designed to show awareness of the extensive level of its functionality, features that can be displayed in the monitoring era . [2] In conclusion, this system is designed for resources are used as intended, prevents from valuable information from leaks out, produce better control mechanism and alerts the user to keep their private information safe.

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

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205295 |

Meenu, Sunila Godara, Phishing Detection using Machine Learning Techniques, IJEAT, ISSN: 2249 – 8958, 2, December,2019. [1] This investigation proposes framework that utilization machine learning systems to beat the spam issue. A model of the framework has been produced on the Azure stage and the conduct of email servers has been examined. Develop a phishing detection model by using various data mining techniques to enhance the phishing detection accuracy and a feature selection method
[2] Finally, the comparison various machine learning techniques like two class logistic regression technique and two class boosted decision tree (DT) ,two class neural network(NN) and two class support vector machine (SVM) and improved logistic regression is proposed to detect spam .

Ankit Kumar Jain and B.B. Gupta EURASIP Journal On Information Security (2016)2016:9. [1] In this paper, we leraned a novel approach to protect against phishing attack using auto-updated white-list of legitimate sites accessed by the individual user. Furthermore, there approach is able to check the legitimacy of a webpage using hyperlink features. There experimental results showed that the proposed approach is very effective in protecting against phishing attacks as it has 86.02 % true positive rate with a very less false positive rate of 1.48 %.

IV. ANALYSIS AND RELATED WORK

To strengthen existing defense mechanisms against URL phishing attacks, various strategies have been proposed. These strategies, outlined in [2], focus on enhancing the accuracy of the classification model by prioritizing superior features and refining the selection process. Additionally, this study introduces a phishing detection model that utilizes diverse data processing techniques, notably leveraging the distinctive approach of VowPal Wabbit [4]. Protective measures such as maintaining a whitelist of legitimate websites, monitoring user visits, implementing authentication via link function, and identifying phishing tactics like DNS spoofing, embedded objects, and zero-hour attacks are recommended to reinforce defense against phishing attempts.

Moreover, the study introduces a 1D CNN-assisted deep learning model for identifying phishing websites. Results indicate that the proposed CNN-based algorithms excel at identifying previously unknown phishing websites [8]. Additionally, a multi-agent predicate architecture and ML classifier are deployed to monitor and thwart online phishing attacks [9], providing an intelligent ML-based system for phishing detection on websites. Furthermore, integrating a web browser extension that alerts users to phishing attacks in real-time is suggested [6]. This system supports various features, including recording banned URLs and directly verifying the legitimacy of online sites from the browser, warning users of blacklisted sites.

To implement email alerts when users attempt to access a website via a pop-up window, perceptual models based on machine learning techniques can be utilized to identify phishing web pages in real-time. Importing a dataset of valid phishing data from a database is recommended for this purpose. The preprocessing plan for the imported dataset includes extracting domain-based features, address-based features, anomaly-based features, and HTML and JavaScript features. These features serve as key factors for predicting whether a website is phishing. Preprocessing aims to extract relevant information and transform it into a format suitable for machine learning algorithms. By analyzing these features, machine learning models can effectively identify and classify phishing websites in real-time, enabling the implementation of proactive measures such as email alerts for users accessing potentially harmful web pages

V. ARCHITECTURE

To tackle the growing menace of phishing scams aimed at unsuspecting individuals, deploying effective phishing detection technology is imperative. The suggested solution entails preprocessing the input data and generating a dataset comprising both phishing information and canonical URLs. Machine learning methods are subsequently employed to analyze this dataset. Illustrated in Figure 1, the flexible approach and development timeline outline the architecture for diagnosing phishing. The process commences with users installing an extension in their Chrome browser window. Upon entering a URL or navigating to a webpage, the initial phase of the diagnosis involves "feature extraction," encompassing various types such as:

- **Address-based feature extraction.**
- **Extraction of abnormal features.**
- **Feature extraction from HTML and JavaScript.**
- **Feature extraction from the domain.**

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 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205295 |

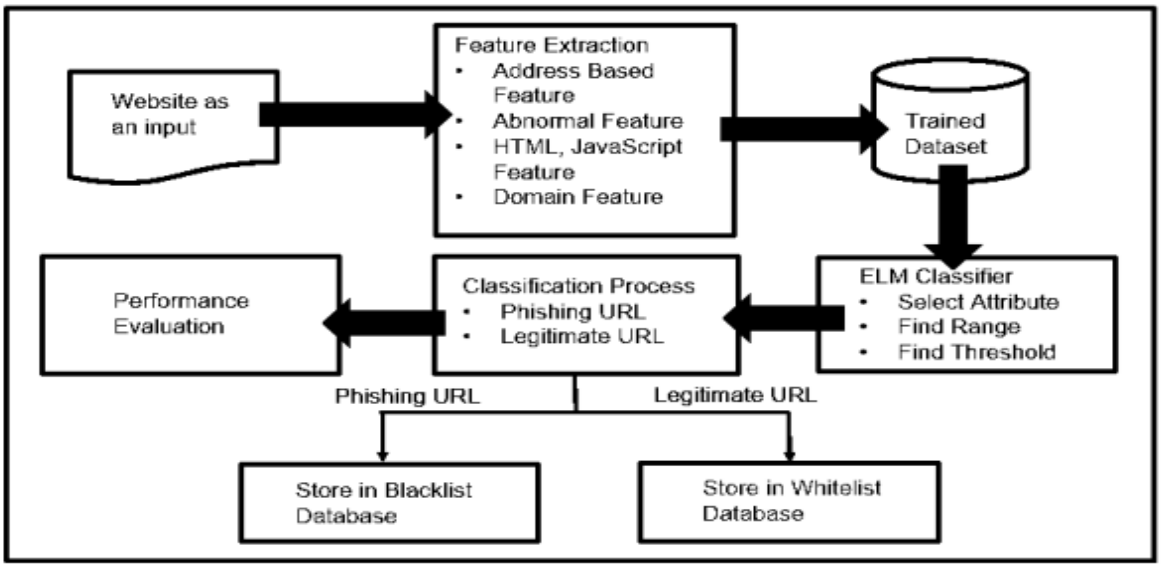

Fig 1: System Architecture

Feature extraction involves computing feature values to assess the legitimacy of a URL. For phishing detection, a specific phishing feature function is generated for each URL. Furthermore, overlap values are determined by examining URLs in anchor tag attributes, and this value is amalgamated with other attributes for a comprehensive evaluation. For example, if the URL contains the "@" symbol, it is assigned a value of "1"; otherwise, it is assigned a value of "0". Similarly, the length of the URL parameter is classified as follows: if the length falls between 51 and 75 characters, it is assigned a value of "0"; if it falls between 51 and 75 characters, it is assigned a value of "1"; and if it exceeds 75 characters, it is assigned a value of "-1".

Machine learning is utilized to analyze URLs by computing range and threshold values for various URL variables. These variables are associated with four algorithms: Extreme Learning Machine (ELM), Support Vector Machine (SVM), Random Forest (RF), and Logistic Regression (LR). The algorithm with the highest accuracy is chosen for classification purposes. The classification process involves assigning aspect values to each URL, based on a predefined set of associated aspect values (-1, 0, and 1). The X-axis represents the cost of the URL anchor tag, while the Y-axis represents the cost of the prefix extension. Both prefix extension URLs and anchor tag aspect URLs incur connection costs. A threshold value is established to determine the classification of URLs. For example, a URL resulting in '0102' may be classified as normal, '00010' as suspicious, and '11100' as indicative of a phishing attempt. Phishing URLs are added to a blacklist database, while valid URLs are added to a whitelist database. Upon completion of the analysis, users receive a result indicating the authenticity of the entered URL. If the URL is identified as a phishing site, users are alerted via a pop-up window displaying the warning message: "The URL is phishing. Do not proceed."

VI. ALGORITHM AND SEQUENCE FLOW

The Support Vector Machine (SVM) classifier and logistic regression are among the most commonly employed machine learning classifiers for establishing robust associations between two subjects. Random forests, on the other hand, are preferred for classification tasks due to their ability to evaluate randomly selected predictors to determine optimal divisions.

Here are the outlined steps for both the decision tree-based method and the Extreme Learning Machine (ELM) approach:

1. Randomly sample from the given dataset.
2. Construct a decision tree for each pattern and obtain prediction results for each tree.
3. Aggregate predictions from all decision trees.
4. Select the prediction with the highest score as the final result.

Extreme Learning Machine (ELM) Approach:

1. Visit a website or web page online.

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

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205295 |

2. Review the policy and extract 30 supported input properties.
3. Group samples in the dataset.
4. Randomly split the dataset into 90% training samples and 10% test samples.
5. Perform classification using ELM:
   5.1. Randomly assign hidden nodes and create parameters for hidden nodes.
   5.2. Determine the output matrix of the hidden layer.
   5.3. Calculate the output weight matrix of the section.
6. Make predictions regarding the website's authenticity.

The process of phishing web page detection involves preprocessing the actual computer address and collecting phishing information, similar to the imported facts. Detection methods include domain-based, address-based, random-based, and HTML and JavaScript capability approaches. Processed data is then used to extract URL attributes, generating a value for each attribute. A machine learning algorithm calculates the winning cost, resulting in edge values for URL attributes, thereby completing URL analysis. URLs are subsequently categorized as phishing or legitimate based on these features.

Advantage:
1. Aid users in steering clear of phishing scams.
2. Implement simple and effective procedures.
3. Attainable objective.
4. Stay flexible to keep pace with evolving trends.
5. Employ a rapid classification technique.
6. Demands a shorter timeframe than alternative approaches.

Drawbacks
1. Users will be alerted to attacks through pop-up notifications, as well as via email or text messages. The alert mechanism functions similarly in both cases.
2. The proposed system is confined to desktop or laptop usage, potentially excluding users of mobile devices like smartphones.

VII. CONCLUSION AND FUTURE WORK

Websites have become integral tools across various domains, facilitating tasks ranging from data management to scientific analysis, and they play a pivotal role in processing input data to derive valuable insights. They find application in diverse fields such as medicine, technology, business, commerce, education, and economics. Despite their widespread utility, websites are susceptible to exploitation by hackers who may employ them for nefarious activities, including phishing attacks.

Numerous research endeavors propose novel methodologies and strategies for detecting phishing attempts, with a particular focus on analyzing URLs and implementing corresponding detection techniques. The overarching objective is to classify phishing attacks among the myriad cyber threats. The proposed systems aim to alert users about potential phishing URLs and recommend safe alternatives even before users access them. Ultimately, the goal is to proactively thwart phishing attacks. To achieve this, advanced machine learning tools are leveraged, often utilizing datasets like the UCI dataset for validation and testing purposes.

REFERENCES

1. SK Hasane Ahammad , Sunil D. Kate , Gopal D. Upaddhye , Sandeep Dwarkanath Pnade el al , ELSEVIER Jan 1 2022
2. Mohammed Hazim Alkawaz, Stephanie Joanne Steven, AsifIqbalHajamydeen, Detecting Phishing Websites Using Machine Learning, 2020 16th IEEE International Colloquium on Signal Processing & its Applications, 28-29 Feb. 2020.
3. Meenu, Sunila Godara, Phishing Detection using Machine Learning Techniques, IJEAT, ISSN: 2249 – 8958, 2, December,2019.
4. Ankit Kumar Jain and B.B. Gupta EURASIP Journal On Information Security (2016)2016:9.
5. Joby James, Sandhya L, Ciza Thomas, Detection of Phishing URLs Using Machine Learning Techniques, 2013 International Conference on Control Communication and Computing (ICCC), December 2013.

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

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205295 |

6. Suleiman Y. Yerima, Mohammed K. Alzaylaee, High Accuracy Phishing Detection Based on Convolutional Neural Networks, IEEEXplore.
7. Megha N, KR Ramesh Babu, Elizabeth Sherly, An Intelligent System for Phishing Attack Detection and Prevention, IEEE Xplore ISBN: 978-1-7281-1261-9, 2019 IEEE.
8. AmaniAlswailem, BashayrAlabdullah, Norah Alrumayh, Dr. Aram Alsedrani, Detecting Phishing Websites Using Machine Learning 978-1-7281-0108- 8/19/ 2019 IEEE.
9. https://www.hindawi.com/journals/jam/2014/425731/(randomforest)
10. https://pdfs.semanticscholar.org/41ca/257920b5b5e6c1 cf4f4417bb85ac5a875935.pdf
11. https://archive.ics.uci.edu/ml/index.php 13. https://www.google.com/

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