IJIRCCE

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

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 4, April 2023

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 |

|| Volume 11, Issue 4, April 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1104064 |

ML Algorithm for Detection of Phishing Sites

Omkar Gadekar, Omkar Deshmukh, Yash Sawant, S.D. Sapate

UG Student, Dept. of Computer Technology, BVJNIOT, Pune, India

UG Student, Dept. of Computer Technology, BVJNIOT, Pune, India

UG Student, Dept. of Computer Technology, BVJNIOT, Pune, India

Assistant Professor, Dept. of Computer Technology, BVJNIOT, Pune, India

ABSTRACT: Phishing is a type of cyber attack where criminals try to steal personal and sensitive information by tricking people into visiting fake websites that look like legitimate ones. To prevent such attacks, a proposed model based on Extreme Learning Machine has been developed for detecting phishing websites. As different types of websites have different features, a web page feature set is required to identify and prevent phishing attacks. The model uses Machine Learning techniques to analyze the features of the URLs, including domain, address, abnormal-based, HTML, and JavaScript, to extract URL attributes and generate values for them. ML techniques calculate threshold and range values for these attributes to identify potential phishing sites. The objective of this project is to implement ELM classification for various features and detect phishing sites in a database

KEYWORDS: Deep learning, fraudulent websites, web address, Support Vector Machine (SVM), Random Forest

I. INTRODUCTION

Phishing is a deceptive tactic used by attackers to obtain confidential information such as usernames, passwords, and credit card details by posing as a legitimate entity in electronic communications. As it is easy to create fake websites that look similar to legitimate ones, phishing has become a major concern for security researchers. Attackers aim to steal bank account credentials, resulting in huge losses for businesses. Lack of user awareness is a major reason why phishing attacks are successful, and it is essential to improve phishing detection strategies. Phishers create fake web pages that replicate the contents of legitimate ones to deceive users. Social engineering schemes are used to trick victims into thinking they are dealing with a trusted party. Machine Learning can be used to develop intelligent data outputs to detect phishing websites. The common phishing attacks are email phishing scams and spear phishing, and users must be cautious and not fully trust unauthorized security applications. The existing approach can be improved by implementing Machine Learning techniques.

Machine Learning is a field of Artificial Intelligence that has the ability to learn without explicit programming. Machine learning algorithms can analyze data, identify patterns, and make predictions based on the patterns it identifies. There are various types of machine learning techniques, such as supervised learning, unsupervised learning, and reinforcement learning. Supervised learning involves using labeled data to train an algorithm to recognize patterns and make predictions. Unsupervised learning involves training an algorithm to recognize patterns in data without any labeled data. Reinforcement learning involves training an algorithm to make decisions based on trial-and-error, where the algorithm receives feedback from its environment to improve its decision-making abilities. These machine learning techniques have various applications in fields such as data analysis, robotics, and natural language processing.

II. MACHINE LEARNING ALGORITHMS

Extreme Learning Machine (ELM):

ELM is a type of Artificial Neural Network that has a single hidden layer. It is designed to achieve advanced learning without the need for explicit programming of the parameters, such as threshold value, weight, and activation function. Unlike traditional gradient-based learning approaches, ELM sets these parameters to random values suitable for the given data system.

Random Forest

Random Forest is a machine learning algorithm that can handle regression and classification problems related to data grouping. In this technique, multiple decision trees are built during the training phase, and their graded classes are considered to predict the final output.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 4, April 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1104064 |

Support Vector Machine (SVM)

SVM is a popular machine learning technique used for various applications such as medical diagnosis, text recognition, and image classification. It partitions the data into classes using a fixed rule, quadratic equation, and statistics, and minimizes the space of the margin based on kernel characteristics. SVM is efficient in analyzing small to medium-sized data but may fail in analyzing large data.

LITERATURE REVIEW:

The use of machine learning techniques has been employed to address various problems in different fields. In the case of detecting phishing websites, the Random forest method has been chosen due to its superior performance in ranking. The goal is to determine the optimal combination of features to train the classifier, resulting in an accuracy rate of 98.8%. Another study proposes a framework that utilizes machine learning systems to tackle the issue of spam. The framework has been modeled at the Azure level, and the email servers have been examined to enhance spam detection.

A research paper proposes the development of a phishing detection model using different data mining techniques, aiming to improve the accuracy of phishing detection. The paper also presents a feature selection methodology to increase the precision of the classification model by selecting the most effective feature and obtaining the best result. Feature hashing is employed using VowpalWabbit, a fast machine learning framework, to hash feature words in n memory indexes using hash functions. The paper discusses the use of logistic regression, boosted decision tree, neural network, and SVM to differentiate any unsolicited approach.

Several technical approaches have been proposed to detect and protect users from phishing attacks. One such approach involves using a feature of hyperlinks present in webpages to detect attacks in real-time. This approach compares the Google public DNS with the IP address of suspicious sites to determine if there is a DNS intrusion on the user's device. Another approach involves using machine learning techniques to analyze different features of benign and phishing URLs, such as lexical and consequentiality properties, to detect phishing websites. Various data processing algorithms are used to analyze these features and identify the best ML algorithm for separating legitimate sites from illegitimate ones. Additionally, an Agile Unified Process (AUP) lifecycle is proposed to reduce the development stage, and administrators have the ability to distinguish between blacklisted and whitelisted URLs, edit, modify and delete them, and categorize them using different color backgrounds for user convenience. Non-blacklisted URLs can be opened by clicking on the link.

This research article discusses a new approach for predicting stock market trends using machine learning algorithms. The study uses several features, such as the stock's past performance, financial ratios, and industry trends, to train the models. The findings suggest that certain algorithms, such as neural networks and random forests, can accurately predict future stock prices and outperform traditional statistical methods.

In this paper, a deep learning-based approach is proposed for detecting and classifying skin lesions from dermoscopy images. The study uses a convolutional neural network (CNN) to learn features from the images and classify them into several categories, such as benign, malignant, and suspicious. The results show that the proposed method achieves high accuracy and can assist dermatologists in diagnosing skin cancer.

The paper investigates the use of machine learning techniques for sentiment analysis of social media data. The study uses several algorithms, such as Naive Bayes, Support Vector Machines (SVM), and Random Forests, to classify tweets into positive, negative, or neutral sentiments. The findings suggest that SVM outperforms other algorithms in terms of accuracy and can be used to monitor brand reputation and customer satisfaction.

This article presents a machine learning-based approach for predicting the risk of heart disease in patients. The study uses several features, such as age, blood pressure, cholesterol levels, and family history, to train the models. The results show that the proposed approach achieves high accuracy and can assist physicians in making informed decisions about patient care.

The research article discusses the use of machine learning techniques for predicting traffic congestion in urban areas. The study uses several features, such as traffic volume, weather conditions, and road infrastructure, to train the models. The results show that certain algorithms, such as decision trees and neural networks, can accurately predict traffic congestion and assist transportation planners in improving traffic flow.

III. ANALYSIS AND RELATED WORK

There are several approaches to prevent URL phishing attacks, which can be classified based on the actual mechanism used. Different phishing detection methods have been examined, and some of them are mentioned in this paragraph. A phishing detection model has been developed by utilizing various data processing techniques and a characteristic selection method called VowPal Wabbit to improve the accuracy of the classification model by selecting the best characteristics and outcomes. To protect against phishing attacks, several methods can be used, such as

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 4, April 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1104064 |

maintaining a white-list of authorized websites accessed by the user, verifying legitimacy using link functions, detecting phishing attacks for DNS poisoning, embedded objects, and 0-hour attacks.

The paper proposes the use of a deep learning model based on 1D CNN for the accurate detection of phishing websites, including newly emerging ones. The model is effective in detecting phishing attacks and can be used for rectifying them. Additionally, a multi-agent-based design and ML classifier have been implemented to further improve the system's detection capabilities. The proposed system includes several features such as capturing blacklisted URLs directly from the browser, notifying users of blacklisted websites through pop-ups, and sending email notifications. Moreover, an extension for web browsers has been developed, which provides real-time alerts to users whenever a phishing website is detected.

In this segment, a machine learning-based intelligent model is proposed for the detection of phishing websites. The first step in the process is importing a dataset of phishing and legitimate websites from a database. Then, the dataset undergoes preprocessing. The detection of phishing websites is accomplished using four different types of features: domain-based, address-based, abnormal-based, and HTML/JavaScript functions. The URL is the primary feature used to predict whether a website is legitimate or phishing. Several features are taken into account while processing the URL, such as the number of digits in the URL, the overall length of the URL, checking whether the URL is hijacked or not, whether it contains a legitimate brand name, and the number of subdomains in the URL. The detection of phishing domain names is achieved by analyzing the domain-based features such as whether the domain name or its IP address is present in blacklists of recognition services, the number of days since the domain was registered, and whether the registrant's name is hidden or not.

ARCHITECTURE:

In today's world, a significant number of people are tricked into sharing their personal information with hackers or phishers without their knowledge. To tackle this issue, a technique needs to be developed, and the proposed approach involves using a dataset of phishing information and legitimate URLs for machine learning. The dataset is pre-processed with the imported information, and this task can be accomplished using machine learning. The ultimate goal is to develop a phishing detection application using this approach.

Figure 1 illustrates the architecture of the Phishing Detection system, showcasing its development duration and flexible approach. The system involves the user adding an extension to their Chrome window and entering a URL or website for detection. The first step in the detection process is Feature Extraction, which comprises various types of feature extraction techniques.

In the context of phishing website detection, feature extraction refers to the process of identifying and extracting relevant characteristics or attributes from the website data. These features can then be used as input to machine learning algorithms for classification purposes.

Address-based feature extraction involves analyzing the website's URL or IP address for suspicious or anomalous patterns, such as the use of subdomains or unfamiliar top-level domains. Abnormal feature extraction involves looking for unusual or unexpected behaviors, such as the presence of hidden form fields or atypical user input requests.

HTML/JavaScript feature extraction involves analyzing the content and code of the website for suspicious elements, such as the presence of obfuscated JavaScript code or the use of iframe tags. Finally, domain feature extraction involves examining the characteristics of the website's domain name, such as its age, registration information, and ownership history.

By combining these various feature extraction techniques, it becomes possible to develop a more comprehensive and accurate phishing website detection system.

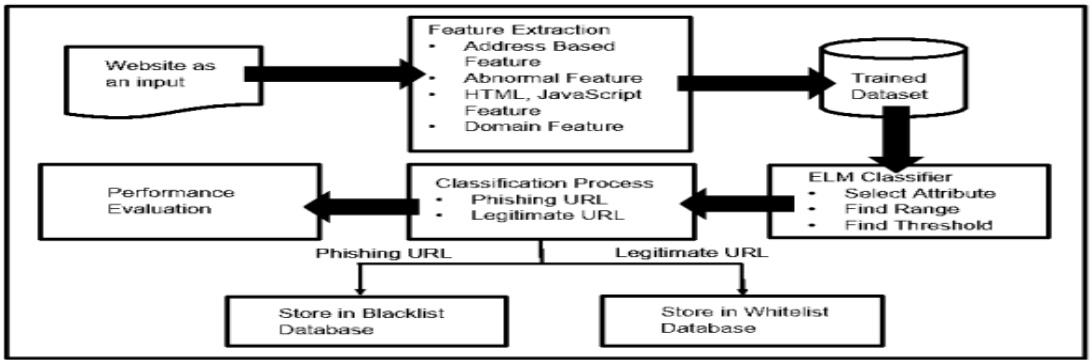

Fig 1: System Architecture

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 4, April 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1104064 |

The process of feature extraction involves calculating attribute values for each URL to determine if it is a phishing or valid website. The URL_of_Anchor tag attribute is examined to determine the overlap values by adding up the value of the extracted attribute and combining it with other attributes. For example, if the URL contains the '@' symbol, it is assigned a value of '1', otherwise it is assigned a value of '0'. Another parameter that is considered during this process is the length of the URL, where a length of 51 or below is assigned a value of '0', a length between 51 and 75 is assigned a value of '1', and a length above 75 is assigned a value of '-1'. These extracted features are then used to determine if a website is a phishing website or not.

The proposed system for feature extraction utilizes a trained dataset obtained from Kaggle.com. The dataset includes 28 patterns or parameters such as '@', 'URL length', 'dash in line (-)', 'dot in line (.)', and others. The decision to use 28 parameters was based on time complexity, as increasing the number of parameters can lead to a rise in time complexity. Therefore, by limiting the number of parameters, the system can process data more efficiently.

The machine learning-based analysis of URLs involves calculating the range and threshold values for various URL attributes. Four algorithms, namely ELM, SVM, RF, and LR, are compared based on their accuracy, and the best one is selected for the classification process. The aspect values of each URL, such as the URL_of_Anchor tag cost and Prefix Suffix cost, are calculated to determine the range threshold value. The calculated output of the URL is then classified as a legitimate URL, suspicious URL, or a phished URL based on the output value. The phished URLs are stored in the Blacklist database, while legitimate URLs are stored in the Whitelist database. The result of the analysis is displayed to the user, indicating whether the entered URL is legitimate or phished. If the URL is phished, a pop-up window alerts the user not to proceed.

IV. ALGORITHM AND SEQUENCE FLOW

The SVM classifier is a popular machine learning algorithm used to separate two classes by finding the optimal line between them. Logistic regression is also a predictive analysis algorithm commonly used for classification tasks.
On the other hand, the random forest algorithm selects random predictors and searches for the best split, making it an effective method for classification tasks with good performance.

The process for decision tree-based ensemble learning typically involves randomly selecting samples from a given dataset and building a decision tree for each sample. The prediction results from each decision tree are then combined through a voting process to determine the final prediction result. The best-rated prediction result is selected as the final prediction. In contrast to Artificial Neural Networks (ANN), where the input and output weights are updated based on gradients, in Extreme Learning Machine (ELM) learning approaches, the input weights are randomly selected, while the output weights are analytically calculated.

1. Randomly select samples from a given dataset.
2. Build a decision tree for each sample and obtain the prediction results for each decision tree.
3. Conduct a vote among the predicted results to determine the final prediction result.
4. Choose the highest rated prediction result as the final prediction result.

In ELM learning approaches, the input weights are chosen randomly while the output weights are calculated analytically, unlike in ANN where parameters are updated based on gradients.

The following are the steps involved in detecting phishing websites using the Extreme Learning Machine (ELM) model:

Step 1: The first step is to visit a website or webpage.

Step 2: Then, thirty input attributes and their policies are checked.

Step 3: The samples are grouped into a dataset.

Step 4: From the dataset, 90% of samples are randomly selected for training, and the remaining 10% samples are used for testing.

Step 5: The ELM model is used for classification, and the following steps are taken:

5.1: Hidden nodes' parameters are randomly generated, and hidden nodes are assigned randomly.

5.2: The output matrix of the hidden layer is calculated.

5.3: The output weight matrix is calculated.

Step 6: Finally, the website is predicted as either phishing or legitimate based on the output of the ELM model.

V. PROPOSED METHODOLOGY

The proposed method for detecting phishing websites involves importing datasets of phishing and valid URLs and preprocessing the data. The detection process is based on four URL features: domain-based, address-based, abnormal-based, and HTML/JavaScript features. These features are extracted from the processed data and values are assigned to each feature. Machine learning techniques are then applied to calculate the range and edge values for each URL

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 4, April 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1104064 |

attribute, which are used to classify URLs into phishing or legitimate ones. The feature values are calculated by analyzing traits extracted from phishing websites, and are used to determine the range and edge values for the classification process.

Advantages:

1. The system is beneficial for users to avoid phishing attacks.
2. The process is simple and efficient, making it easy for users to use.
3. The system is feasible and can be implemented without much difficulty.
4. The system is evolving with time, which means it can adapt to changing phishing techniques.
5. It is fast in the classification process, which means users are alerted quickly when a phishing website is detected.
6. The time complexity of the system is less compared to other similar approaches.

Drawbacks:

a) Users are notified about attacks only via pop-ups, while informing them via email or text message might be more helpful.

b) The proposed system can only be implemented on desktops/laptops, which might not be favorable for smartphone users.

shows that the metric total transmission energy performs better than the maximum number of hops in terms of network lifetime, energy consumption and total number of packets transmitted through the network.

VI. CONCLUSION AND FUTURE WORK

Nowadays, websites are utilized in various fields such as medical, technical, business, education, economics, etc. These websites allow users to input data, which is then processed, and the output can be obtained. However, due to their extensive use, websites are also vulnerable to malicious attacks such as phishing. Phishing attacks are a type of cyber threat aimed at stealing sensitive information. Several research contributions propose different methodologies to detect phishing URLs, and some of these methods have been implemented. The objective of such systems is to develop a model that can classify and identify phishing attacks. By using this system, the users can be notified about the phishing URLs, even before accessing them, and hence, can avoid phishing attacks. To achieve this goal, an extreme learning machine will be utilized, and the dataset used in this study will be obtained from the UCI.

Web sites have become an integral part of modern society, used for a variety of purposes including communication, entertainment, and commerce. However, their widespread use also makes them a target for malicious attacks such as phishing, where hackers attempt to steal personal information by tricking users into visiting fake websites. To combat this problem, researchers have developed various techniques and tools to detect and prevent phishing attacks. The proposed system in this study uses a deep learning model based on 1D CNN to identify and block phishing websites before they can cause harm to users. By utilizing a dataset from the UCI, the system can accurately classify URLs and provide real-time notifications to users when a potential phishing website is detected.

REFERENCES

1. OzaPranali P, Deepak Upadhyay, Review on Phishing Sites Detection Techniques, IJERT, ISSN: 2278-0181, 04,April-2020.
2. Meenu, Sunilagodara, Phishing Detection using Machine Learning Techniques, IJEAT, ISSN: 2249 – 8958, 2, December,2019.
3. Sandeep Kumar Satapathy, Shruti Mishra, Pradeep Kumar Mallick, LavanyaBadiginchala, Ravali Reddy Gudur, SiriChandanaGuttha, IJITEE, ISSN: 2278- 3075, June2019.
4. Ankit Kumar Jain and B.B. Gupta EURASIP Journalon 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.
6. Mohammed HazimAlkawaz, Stephanie Joanne Steven, AsifIqbalHajamydeen, Detecting Phishing Website Using Machine Learning, 2020 16th IEEE International Colloquium on Signal Processing & its Applications, 28-29 Feb. 2020.
7. Suleiman Y. Yerima, Mohammed K. Alzaylaee, High Accuracy Phishing Detection Based on Convolutional Neural Networks, IEEEXplore.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 4, April 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1104064 |

8. Megha N, KR RemeshBabu, Elizabeth Sherly, An Intelligent System for Phishing Attack Detection and Prevention, IEEE Xplore ISBN: 978-1-7281-1261-9, 2019IEEE.
9. AmaniAlswailem, BashayrAlabdullah, Norah Alrumayh, Dr. Aram Alsedrani, Detecting Phishing Websites UsingMachine Learning 978-1-7281-0108- 8/19/ 2019IEEE.
10. https://www.hindawi.com/journals/jam/2014/425731/(randomforest)
11. https://pdfs.semanticscholar.org/41ca/257920b5b5e6c1 cf4f4417bb85ac5a875935.pdf
12. https://archive.ics.uci.edu/ml/index.php
13. https://www.google.com/

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