International Journal for Innovative Engineering and Management Research

PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL

www.ijiemr.org

COPY RIGHT

ELSEVIER

SSRN

IJIEMR Transactions, online available on 05th Apr 2023. Link :http://www.ijiemr.org/downloads.php?vol=Volume-12&issue=Issue 04

10.48047/IJIEMR/V12/ISSUE 04/10

Title **MACHINE LEARNING MODEL FOR PREVENTING CYBER THREATS IN PHISHING URL DETECTION**

Volume 12, ISSUE 04, Pages: 67-75

Paper Authors

Mr.Ch.Vijayananda Ratnam, Ch.Sai Durga, G.Pravallika, B.Ganesh, Ch.Hema Chandana

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

MACHINE LEARNING MODEL FOR PREVENTING CYBER THREATS IN PHISHING URL DETECTION

[1]**Mr.Ch.Vijayananda Ratnam**, Assistant Professor, Department of CSE,
Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dt., Andhra Pradesh.

[2]**Ch.Sai Durga**, [3]**G.Pravallika**, [4]**B.Ganesh**, [5]**Ch.Hema Chandana**
[2,3,4,5] UG Students, Department of CSE,
Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dt., Andhra Pradesh.
[2,3,4,5] chsaidurga1111@gmail.com, ganji.pravallika2002@gmail.com,
ganeshbalam629@gmail.com, chinnihemachandana1@gmail.com

Abstract

Phishing, which involves tricking unsuspecting online users into revealing confidential information for fraudulent purposes, is the most commonly used social engineering and cyber attack. To avoid falling victim to these attacks, users should be aware of phishing websites and maintain a blacklist of known phishing websites. Early detection of phishing websites can be achieved through the use of machine learning and deep neural network algorithms. Among these methods, machine learning has proven to be the most effective in detecting phishing websites. However, despite these efforts, online users still fall prey to phishing websites, which mimic legitimate URLs and webpages. The objective of this project is to train machine learning models and deep neural networks on a dataset of phishing and benign website URLs to predict phishing websites. Relevant URL and website content-based features are extracted from the dataset to form a classification problem, where input URLs are classified as either phishing (1) or legitimate (0). The performance of each model, including Decision Tree, Random Forest, Multilayer Perceptrons, XGBoost, Autoencoder Neural Network, and Support Vector Machines, will be measured and compared.

Keywords: Decision Tree, Random Forest, Multilayer Perceptrons, XGBoost, Autoencoder Neural Network, Support Vector Machines, Phishing attacks.

Introduction

The Internet has become an essential aspect of our daily lives, but it also provides opportunities for anonymous and malicious activities like phishing. Phishers employ social engineering techniques or create fake websites to steal sensitive information, such as account IDs, usernames, and passwords from both individuals and organizations. Although many methods have been developed to detect phishing websites, phishers have adapted their techniques to evade detection.

Machine learning has emerged as one of the most effective methods for detecting these malicious activities. This is because phishing attacks share common characteristics that can be identified by machine learning algorithms. Phishing attacks based on URLs involve sending malicious links to users that appear legitimate, tricking them into clicking on them. In phishing detection, incoming URLs are analyzed based on different features to determine if they are phishing or legitimate, and are classified accordingly. Various machine learning algorithms are trained on datasets of URL features to classify a given URL as phishing or legitimate.

Phishing Attacks

Phishing attacks involve sending false communications that seem to originate from a trustworthy source, typically via email, with the intention of stealing sensitive information such as login credentials or credit card details, or installing malware on the victim's device. Phishing is a prevalent form of cyber attack that everyone should be familiar with in order to safeguard themselves.

Different types of phishing attacks include:

A. Spear Phishing

Spear phishing is a more targeted approach to phishing, as opposed to a widespread attempt to deceive a large group of people. Attackers typically research their victims on various platforms, such as social media, to gather information and customize their communications to appear more legitimate. Spear phishing is frequently used as the initial step to infiltrate a company's security defenses and carry out a deliberate attack.

B. Deceptive Phishing

The most prevalent form of phishing is deceptive phishing, which involves attackers attempting to acquire sensitive information from their victims. Attackers may use this information to launch further attacks or to steal money. An instance of deceptive phishing is a fraudulent email from a bank that prompts the recipient to click on a link and verify their account information.

C. Whaling

When attackers set their sights on high-profile targets like CEOs, it's known as whaling. These attackers typically invest a significant amount of time researching their target to identify the ideal opportunity and method of stealing login credentials. Whaling particularly concerning because top-level executives have access to a vast amount of sensitive company information.

D. Pharming

Pharming is a cyber attack that shares similarities with phishing, as it involves directing users to a fake website that appears to be genuine. Unlike phishing, victims do not need to click on any malicious links to be redirected to the fraudulent site. Instead, attackers can infect either the user's computer or the website's DNS server and redirect the user to the bogus site, even if the correct URL is manually entered. This makes pharming an even more insidious type of attack that can easily deceive even the most cautious internet users.

Literature Review

In a paper authored by Rishikesh Mahajan and Irfan Siddavatam [10], three classification algorithms - Decision Tree, Random Forest, and Support Vector Machine - were selected. Their dataset comprised 17,058 benign URLs and 19,653 phishing URLs, each with 16 features, collected from Alexa and PhishTank websites, respectively. The dataset was divided into training and testing sets in the ratios of 50:50, 70:30, and 90:10. The performance evaluation metrics included accuracy score, false negative rate, and false positive rate. The authors achieved a 77.14% accuracy score for the Random Forest algorithm, with the lowest false negative rate. The paper concluded that the accuracy of the classification algorithms increases with an increase in the amount of training data used. Bahrami, A., & Asghari, M conducted a study in [9], where they trained various classifiers such as Logistic Regression, Naive Bayes Classifier, Random Forest, Decision Tree, and K-Nearest Neighbor, using features extracted from the lexical structure of the URL. To address issues such as data imbalance, biased training, variance, and overfitting, they created a dataset of URLs that contained an equal number of labeled phishing and legitimate URLs. They further split the dataset into a 7:3 ratio for training and testing. The Naive Bayes Classifier achieved the highest accuracy score of 82%, with a precision of 1, recall of 0.80, and F1-score of 0.81. Kumar, J., & Kamboj, S. proposed a machine-learning-based phishing detection system in [7]. For three distinct datasets, they applied the following techniques: Logistic Regression (LR), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Decision Tree (DT), Naive Bayes (NB), XGBoost, Random Forest (RF), and Artificial Neural Network (ANN). Their findings indicated that models using LR, SVM, and NB had a low accuracy rate. They concluded that the RF or ANN algorithm might be preferable because they require less training time while maintaining a high accuracy rate. Singh, A. K., Singh, M., & Joshi, R. C.proposed an intelligent phishing detection system using the UCI dataset in [3]. RF was also found to be faster, more robust, and accurate than the other classifiers.

Problem Identification

After conducting a thorough observation and study on the classification of phishing websites using machine learning techniques, we have identified a problem. We need to develop a system that can accurately and efficiently classify websites as either legitimate or phishing, with minimal time consumption and cost.

Methodology

Feature Extraction:

1) Presence of IP address in URL: A feature is set to 1 if an IP address is present in the URL, and 0 otherwise. Typically, benign websites do not use IP addresses in their URLs to download webpages. However, if an attacker includes an IP address in the URL, it may

indicate an attempt to steal sensitive information.

2) Presence of @ symbol in URL: A feature is assigned a value of 1 if the URL contains an "@" symbol, and 0 otherwise. Phishers often add the "@" symbol to a URL to deceive users. This symbol causes the browser to ignore everything preceding it, and the actual address usually follows the "@" symbol.

3) Number of dots in Hostname: If a URL contains more than three dots, the feature is assigned a value of 1, otherwise it is assigned a value of 0. Phishing URLs often have many dots, such as in the example "http://shop.fun.myntra.phishing.com," where "phishing.com" is the actual domain name and the word "myntra" is used to deceive users. On average, benign URLs contain three dots or fewer.

4) Prefix or Suffix separated by (-) to domain: If a URL contains a dash symbol ("-") separating the domain name, the feature is assigned a value of 1, otherwise it is assigned a value of 0. Legitimate URLs rarely use dash symbols in their domain names. Phishers may add a dash symbol to a domain name to make it appear legitimate to users, as in the example where the actual website is "http://www.onlinemyntra.com" but the phisher creates a fake website like "http://www.online-myntra.com" to deceive innocent users.

5) URL redirection: If a URL path contains "//", the feature is assigned a value of 1, otherwise it is assigned a value of 0. The presence of "//" in the URL path indicates that the user will be redirected to another website.

6) HTTPS token in URL: If a URL contains the "HTTPS" token in the domain part, the feature is assigned a value of 1, otherwise it is assigned a value of 0. Phishers may add the "HTTPS" token to a URL to deceive users, as in the example "http://https-wwwpaypal-it-mpp-home.soft-hair.com".

7) URL Shortening Services “TinyURL”: If a URL is generated using a shortening service (such as bit.ly), the feature is assigned a value of 1, otherwise it is assigned a value of 0. Phishers may use shortening services to hide long phishing URLs and redirect users to their malicious websites.

8) Length of Host name: If a URL's length exceeds 25 characters, the feature is assigned a value of 1, otherwise it is assigned a value of 0. On average, benign URLs have a length of 25 characters.

9) Age of SSL Certificate: he presence of HTTPS is crucial in conveying the legitimacy of a website. However, the SSL certificates of benign websites typically have a minimum age of one to two years.

10) IFRAME: We extracted this feature by crawling the source code of the URL. The "iframe" tag is used to embed another webpage within the existing main webpage. Phishers may use the "iframe" tag and make it invisible, without any frame borders, so that the inserted webpage appears to be part of the main webpage. Users may then enter sensitive information, believing that they are on a legitimate website.

Implementation

Decision Tree

Decision trees are commonly used in phishing URL detection as a machine learning algorithm. The decision tree algorithm works by building a tree-like model of decisions based on the features of URLs that are classified as either legitimate or phishing. The algorithm starts by examining the entire dataset and finding the feature that best splits the data into two groups with the highest purity (i.e., the groups with the highest number of URLs classified as either legitimate or phishing). It then recursively applies this process to each of the resulting groups until a stopping criterion is met, such as reaching a certain depth or minimum number of samples in a leaf node[8]. In the case of phishing URL detection, the features used in the decision tree might include the presence of an IP address in the URL, the length of the domain name, the use of special characters or numbers in the domain name, and the presence of certain keywords or phrases in the URL. These features are selected based on their ability to distinguish between legitimate and phishing URLs in the training dataset. Once the decision tree has been built, new URLs can be classified by traversing the tree starting at the root node and following the decision paths based on the features of the URL being evaluated. The final classification is based on the leaf node reached by the URL, which is either a phishing or legitimate label.

Random Forest

Random forest in phishing URL detection. Random forest is an ensemble method that builds multiple decision trees and combines their results to make predictions[2]. The first step is to collect a dataset of URLs labelled as either legitimate or phishing. This dataset is typically split into training and testing sets. Next, features are extracted from each URL in the dataset. These features might include the length of the domain name, the presence of an IP address, the use of special characters or numbers in the domain name, and the presence of certain keywords or phrases in the URL. The random forest algorithm then builds multiple decision trees using different subsets of the training data and different subsets of the features. This helps to reduce overfitting and improve generalization performance. To classify a new URL, the random forest algorithm evaluates the URL against each of the decision trees in the forest and combines their results to make a final prediction. This approach helps to reduce the impact of individual decision trees that may be biased or overfit to the training data. Finally, the accuracy of the random forest algorithm is evaluated on a holdout test dataset to ensure that it is correctly identifying both legitimate and phishing URLs. RF can help security analysts to identify the most important features for phishing URL detection and improve the algorithm over time.

Multilayer Perceptrons

Multilayer perceptrons (MLPs) are a type of artificial neural network (ANN) that can also be used for phishing URL detection. The first step is to collect a dataset of URLs labelled as either legitimate or phishing. This dataset is typically split into training and testing sets. Next, features are extracted from each URL in the dataset. These features might include the length of the domain name, the presence of an IP address, the use of special characters or numbers in the domain name, and the presence of certain keywords or phrases in the URL. The input layer of the MLP takes in the features, and the output layer produces a prediction of whether the URL is legitimate or phishing. The hidden layers between the input and output layers can include any number of nodes and are used to learn complex relationships between the features and the target variable[1]. To classify a new URL, the MLP algorithm feeds the extracted features into the input layer of the trained neural network and generates a prediction at the output layer. Finally, the accuracy of the MLP algorithm is evaluated on a holdout test dataset to ensure that it is correctly identifying both legitimate and phishing URLs. Overall, MLPs are a powerful and flexible algorithm for phishing URL detection.

XGBoost

XGBoost is a popular machine learning technique that has been applied to various domains, including phishing URL detection. Phishing URLs are malicious links that can trick users into revealing sensitive information or downloading malware. Detecting these URLs is essential to prevent users from falling victim to phishing attacks[4]. XGBoost in phishing URL detection is its ability to handle imbalanced datasets, where the number of legitimate URLs far exceeds the number of phishing URLs. XGBoost can assign different weights to each class to ensure that the model does not overfit to the majority class. Another advantage is its interpretability, which allows security analysts to understand how the model makes its predictions. This transparency can aid in the detection of previously unseen phishing attacks and help improve the model's accuracy over time. Overall, XGBoost has proven to be a powerful tool in phishing URL detection and is widely used in various security applications.

Autoencoder Neural Network

In the context of phishing URL detection, ANN can be trained on a dataset of legitimate and phishing URLs to automatically learn the underlying patterns and features that distinguish between the two[5]. The basic idea behind ANN is to compress the input data (URLs in this case) into a lower-dimensional representation and then reconstruct the original data from this compressed representation. The compression and reconstruction process is achieved through a series of neural network layers that learn to extract and combine different features of the input data. Once the ANN is trained on a dataset of legitimate and phishing URLs, it can be used to detect phishing URLs by comparing the reconstructed output of a given URL with the original input. If the reconstructed output is significantly different from the original input, then the URL is flagged as a phishing URL.

Support Vector Machine

Support Vector Machines (SVM) is another popular machine learning technique that has been applied to phishing URL detection[6]. SVM is a binary classification algorithm that aims to find the hyperplane that separates the data into different classes while maximizing the margin between them. In phishing URL detection, SVM can be trained on a dataset of known phishing and legitimate URLs, where each URL is represented by a set of features, such as the URL's length, the presence of suspicious keywords, and the domain's age. The SVM model learns to distinguish between the two types of URLs based on these features. One advantage of SVM in phishing URL detection is its ability to handle high-dimensional feature spaces, where the number of features can be much larger than the number of training samples. SVM achieves this by mapping the feature space to a higher-dimensional space, where the data can be separated more easily. Another advantage is its ability to handle non-linear decision boundaries, which can be useful in detecting more sophisticated phishing attacks that use obfuscation techniques to hide the malicious intent of the URL.

Results & Conclusion

Phishing attacks are a serious concern in today's world and can cause significant damage to individuals and organizations alike. Therefore, the detection of phishing URLs has become crucial in protecting users against such attacks. Machine learning algorithms have shown great potential in detecting phishing URLs due to their ability to analyze large amounts of data and identify patterns. Several studies have been conducted to detect phishing URLs using various machine learning algorithms. These studies have used different datasets, algorithms, and performance evaluation metrics. However, most of these studies have achieved high accuracy in detecting phishing URLs, indicating the effectiveness of machine learning algorithms in this task. The graph for decision tree algorithm in phishing url detection

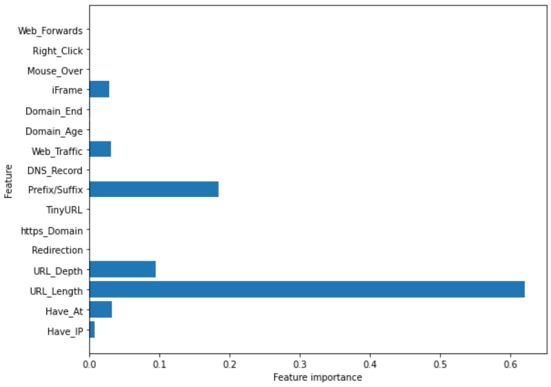

Fig .1.Decision Tree

The graph for random forest algorithm in phishing url detection

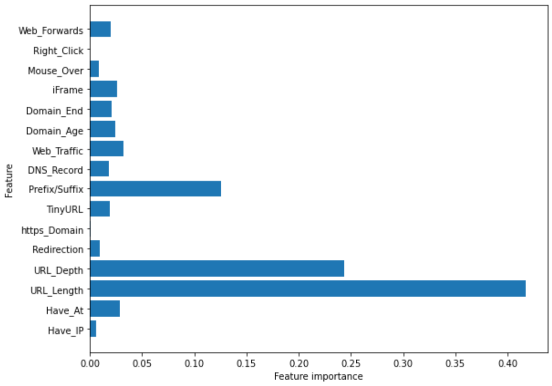

Fig.2.Random Forest

The results we get using various Machine Learning Algorithms for phishing url detection

Algorithm	Accuracy on training Data	Accuracy on test Data
Decision Tree	0.810	0.826
Random Forest	0.814	0.834
Multilayer Perceptrons	0.859	0.863
XGBoost Classifier	0.866	0.864
Autoencoder Neural Network	0.819	0.818
Support Vector Machine	0.798	0.818

In summary, the use of XGBoost Classifier Machine Learning algorithm in detecting phishing URLs is a promising approach. Further research and development in this area could lead to the creation of more accurate and effective tools to detect and prevent phishing attacks, ultimately improving the security of individuals and organizations.

Future Scope

Most current model focuses on binary classification of phishing and legitimate URLs. However, in the future types, we may explore multi-class classification, which can detect different of phishing attacks. Real-time detection of phishing URLs can help organizations quickly identify and respond to attacks. We can focus on developing algorithms that can detect phishing URLs in real-time and integrate them into existing security systems.

References

[1] Alazab, M., Hobbs, M., Abawajy, J., & Alazab, M. (2016). Detection of phishing attacks: A machine learning approach. Expert Systems with Applications

[2] Almehmadi, S., & Sloan, R. H. (2017). Detecting phishing websites using machine learning. Computers & Security.

[3] Singh, A. K., Singh, M., & Joshi, R. C. (2018). A review on phishing detection techniques. Journal of Network and Computer Applications.

[4] Arshad, R., Iqbal, F., & Iqbal, W. (2019). An efficient approach for phishing URL detection using machine learning techniques. Computers & Electrical Engineering.

[5] Garg, A., Sharma, A., & Singh, A. (2019). A review on phishing detection techniques using machine learning.

[6] Khan, M. F., & Akhtar, N. (2019). Detection of phishing URLs using machine learning and feature selection techniques. Arabian Journal for Science and Engineering.

[7] Kumar, J., & Kamboj, S. (2019). A survey on phishing detection techniques using machine learning. International

Journal of Emerging Technologies in Engineering Research.

[8] Gupta, R., & Kaur, P. (2020). Detection of phishing websites using machine learning algorithms. International Journal of Computer Applications.

[9] Bahrami, A., & Asghari, M. (2020). A survey on phishing detection using machine learning techniques. Journal of Ambient Intelligence and Humanized Computing.

[10] Mahajan, R., & Siddavatam, I. (2021). Comparative study of decision tree, random forest and support vector machine for phishing URL classification. International Journal of Machine Learning and Cybernetics.