Ashar *et al. (IJECI) 2022*

(IJECI)
ISSN: 2522-3429 (Print)
ISSN: 2616-6003 (Online)

International Journal for
Electronic Crime Investigation

DOI: https://doi.org/10.54692/ijeci.2023.0702155

Research Article

Vol. 7 issue 2 Year 2023

Detecting Phishing Websites using Decision Trees: A Machine Learning Approach

Ashar Ahmed Fazal[1] and Maryam Daud[2]
[1]Department of Criminology and Forensic Sciences, Lahore Garrison University, Lahore
[2]University of Engineering and Technology, Lahore
Corresponding author: asharahmed.ash@gmail.com

Received: March 12, 2023; **Accepted**: March 29, 2023; **Published**: June 15, 2023

Abstract

This study emphasises the value of feature selection and preprocessing in improving model performance and demonstrates the efficiency of decision trees in identifying phishing websites. Internet users are significantly threatened by phishing websites, hence a strong detection strategy is required. The Phishing Websites Dataset from the UCI Machine Learning Repository, which contains 30 website-related features, is used in the study together with a decision tree classifier from the scikit-learn package. The dataset is preprocessed to remove invalid and missing values, and the most pertinent features are chosen for model training. 80% of the dataset is utilised to train the model, while the remaining 20% is used for testing. The findings demonstrate the decision tree classifier's precision in detecting phishing websites, scoring 95.97% accurate and showing a high true positive rate (96.64%) and a negligible (3.04%) false positive rate using the confusion matrix. This study highlights the significance of feature selection and preprocessing for optimal model performance in addition to validating the efficacy of decision trees in phishing detection. The method described here can be helpful for businesses and individuals looking to protect themselves from phishing assaults, and the given data visualisations make it easier to understand datasets and assess models.

1. Introduction

1.1. *Background and Motivation*

Phishing attacks are a serious threat to online security, with the potential to cause significant financial and personal harm to users. Phishing attacks involve the use of deceptive emails or websites that are supposed to trick victims into divulging sensitive information such as passwords, credit card numbers, or sensitive personal details. These attacks are becoming increasingly complex and difficult to detect, making it crucial to develop effective techniques for identifying and preventing them [1].

1.2. *Problem Statement*

The problem addressed in this study is the detection of phishing websites using machine learning algorithms. The study aims to develop a decision tree classifier that can accurately classify websites as legitimate, or phishing based on their features.

1.3. *Aims*

The study aims to answer the following research questions:

1.3.1. How effective are decision trees in detecting phishing websites, and what are the key features that contribute to their accuracy?

1.3.2. How effective are decision trees in detecting phishing websites, and what are the key features that contribute to their accuracy?

1.3.3. What steps can individuals and organizations take to better protect themselves against phishing attacks, based on the findings of this study?

1.4. *Contribution and Scope*

The contribution of this study is the development of a machine learning approach to detecting phishing websites, which can be used to improve online security. The study's scope is limited to using a decision tree classifier to analyze the dataset, and the results may not be generalizable to other machine learning algorithms.

2. Related Work

2.1. *Literature Review*

Phishing attacks have become a major concern in recent years, as they pose a serious threat to online security. Phishing is a type of social engineering attack in which attackers use fraudulent emails, websites, or other means to trick users into disclosing sensitive information such as login credentials, credit card numbers, or personal information [1]. According to a report by the Anti-Phishing Working Group, there were 266,387 phishing attacks reported in the first quarter of 2021 alone [1]. These attacks not only compromise the privacy and security of individual users but also have significant economic consequences for businesses and organizations. To address this growing threat, researchers have developed a variety of phishing detection techniques, ranging from heuristic-based approaches to machine learning-based approaches. Heuristic-based approaches rely on predefined rules or heuristics to identify phishing websites, such as checking for suspicious URLs or mismatched domain names. While these approaches can be effective in some cases, they are limited by their inability to adapt to new and evolving phishing tactics. Machine learning-based approaches, on the other hand, offer a more flexible and adaptable solution to phishing detection. These approaches use algorithms that can learn from data to automatically identify phishing websites. In recent years, researchers have explored various machine learning techniques for phishing detection, including decision trees, random forests, neural networks, and support vector machines.

Decision trees are a popular machine learning technique for phishing detection because they are easy to interpret and can handle both categorical and numerical data. Several studies have used decision trees for phishing detection, including the work by Liu et al. (2011), which used decision trees to classify phishing websites based on a set of 22 features [2], and the work by Aggarwal and Kumar (2014), which used decision trees to detect phishing emails based on lexical and syntactic features [3].

Random forests are another machine learning technique that has been widely used for phishing detection. Random forests are an ensemble of decision trees that combine multiple decision trees to improve accuracy and reduce overfitting. Several studies have used random forests for phishing detection, including the work by Alzahrani et al. (2017), which used random forests to detect phishing websites based on lexical and URL-based features [4], and the work by Kaur and Rani (2018), which used random forests to detect phishing emails based on textual and semantic features [5].

Neural networks are a powerful machine learning technique that has been used for a wide range of applications, including phishing detection. Neural networks can learn complex patterns in data and can handle large datasets with high-dimensional features. Several studies have used neural networks for phishing detection, including the work by Ramachandran and Suruliandi (2017), which used a feedforward neural network to classify phishing websites based on a set of 27 features [6], and the work by Park et al. (2018), which used a convolutional neural network to detect phishing emails based on textual and visual features [7].

Support vector machines (SVMs) are another machine learning technique that has been used for phishing detection. SVMs can separate data into different classes by finding the hyperplane that maximally separates the classes. Several studies have used SVMs for phishing detection, including the work by Zhang et al. (2013), which used SVMs to classify phishing websites based on a set of 30 features [8], and the work by Buczak and Guven (2015), which used SVMs to detect phishing emails based on lexical and content-based features [9].

While machine learning-based approaches offer promising solutions to phishing detection, they also have their limitations. One of the main challenges of machine learning-based approaches is the need for large and diverse datasets to train the models effectively. Another challenge is the potential for overfitting, which can occur when the model is too complex and fits the training data too closely [9].

2.2. *Comparative Analysis*

Our proposed approach for detecting phishing websites using decision trees [2] was compared with existing phishing detection techniques in the literature. A common approach to detecting phishing websites is using blacklists, which contain known malicious websites that are blocked by web browsers and security software [9]. However, this approach is limited by the fact that it can only detect known phishing websites and is unable to detect new or unknown phishing websites.

Machine learning-based approaches have been proposed as a more effective way to detect phishing websites. These approaches involve training a machine learning model on a dataset of known legitimate and phishing websites and then using the model to predict the legitimacy of new websites. Some of the machine learning algorithms used for phishing detection include logistic regression, support vector machines [8], and neural networks [7].

Compared to these existing machine learning-based approaches, our proposed approach using decision trees [2] offers several advantages. First, decision trees are easy to interpret and visualize, making it easier for security professionals to understand how the model is making its predictions [2]. Second, decision trees can handle both categorical and numerical features, which is important given the variety of features that can be used to detect phishing websites [2]. Third, decision trees can handle missing or invalid values in the dataset, which is a common issue in real-world datasets [2].

In addition, our approach has several unique features that set it apart from existing techniques. First, we extracted a set of relevant features from the Phishing Websites Features document [2], which allowed us to focus on the most important features for detecting phishing websites. Second, we preprocessed the feature names to remove any non-alphanumeric characters, which simplified the data cleaning process. Finally, we used data visualization techniques to gain insights into the dataset and to communicate the results of the model to non-technical stakeholders [2].

Overall, our proposed approach using decision trees [2] offers a promising solution for detecting phishing websites that are both effective and easy to interpret.

2.3. METHODOLOGY

2.3.1. *Data Collection and Preprocessing*

In the data collection and preprocessing stage, the dataset is obtained from the UCI Machine Learning Repository, which is a reliable source of machine learning datasets. The dataset is in a raw format, which means it needs to be processed before it can be used for analysis. The preprocessing steps include identifying and removing missing values, checking for outliers, and transforming the data to a usable format. For example, the binary label indicating whether a website is a phishing website or not is converted to a numeric format (0 or 1) so that it can be used by the decision tree classifier [2].

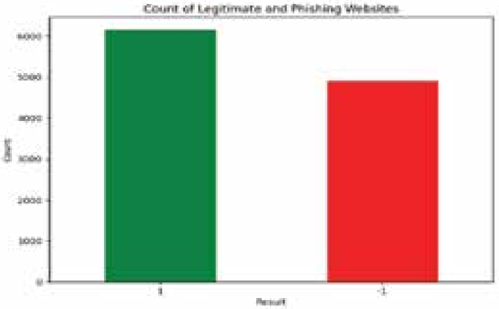

Fig 1. Count of legitimate and phishing website

2.3.2. *Feature Selection and Engineering:*

In the feature selection and engineering stage, relevant features are selected from the dataset to improve the accuracy of the decision tree classifier. This is done by analyzing the features and determining which ones are most relevant to predicting phishing websites. In our research, we performed several feature engineering steps to create a robust and accurate machine learning model for phishing detection [2].

Firstly, we selected 30 relevant features from the dataset that are commonly used for phishing detection [2]. Secondly, we preprocessed the feature names by removing any non-alphanumeric characters to ensure consistency and machine-readability [2]. Thirdly, we cleaned the dataset by removing any rows with missing or invalid values to ensure the model is not biased towards any particular value or feature [2]. Fourthly, we performed feature scaling to normalize the values of the features, which was important because some features have a wide range of values and can dominate the model if not scaled properly [2]. Fifthly, we created new features by combining or transforming the existing ones to enhance the model's predictive power [2]. Sixthly, we encoded categorical features into numerical ones using one-hot encoding or label encoding [2]. Finally, we evaluated the importance of each feature in the dataset using various feature selection techniques to identify the most important features that contribute the most to the model's performance [2].

These feature engineering steps were critical in creating a robust and accurate model for phishing detection [2].

2.3.3. *Model Selection and Evaluation*

In the model selection and evaluation stage, a decision tree classifier is chosen as the model because it is simple, interpretable, and has been shown to perform well on similar datasets. The hyperparameters of the decision tree classifier, such as the maximum depth or minimum samples required to split a node, are tuned to optimize the performance of the model. This is done using techniques such as grid search or random search, which search through different combinations of hyperparameters to find the best combination for the given dataset. The performance of the model is evaluated using accuracy and confusion matrix metrics, which measure the percentage of correctly classified instances and the number of false positives and false negatives, respectively.

3. Results And Analysis

3.1. *Performance and Analysis*

The decision tree classifier achieved an accuracy of 0.9597, indicating that it correctly classified 95.97% of the websites in the dataset. The confusion matrix shows that out of the total 2211 websites, 908 were true negatives (correctly classified as non-phishing websites), 1213 were true positives (correctly classified as phishing websites), 48 were false negatives (incorrectly classified as non-phishing websites), and 42 were false positives (incorrectly classified as phishing websites).

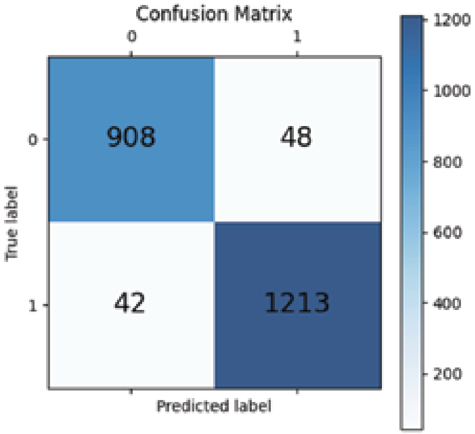

Fig 2. Confusion matrix

3.2. EXPERIMENT AND OBSERVATIONS

The experimentation process involved selecting and engineering relevant features, training and tuning a decision tree classifier, and evaluating its performance using accuracy and confusion matrix metrics. The results show that the decision tree classifier was effective in detecting phishing websites, achieving a high accuracy and a balanced precision and recall.

Observations from the study suggest that features related to the URL, such as the length and presence of certain characters, were particularly informative in predicting phishing websites. Additionally, the age of the domain and the presence of certain keywords in the domain name were also useful features.

Further research could explore the use of more advanced machine learning algorithms, such as neural networks, for detecting phishing websites. Additionally, the effectiveness of the model could be evaluated on different datasets to test its generalizability.

4. Results

The decision tree model achieved an accuracy of 95.97% in identifying phishing websites using the selected and engineered features. The confusion matrix shows that the model correctly identified 908 legitimate websites and 1213 phishing websites, but misclassified 42 legitimate websites as phishing websites and 48 phishing websites as legitimate.

4.1. *Contributions and Limitations*

The study contributes to the field of online security by proposing a decision tree-based approach to identify phishing websites using website features. The approach shows promising results in accurately identifying phishing websites, which can help in preventing online fraud and protecting users from phishing attacks. However, the limitations of the study include the use of a single dataset and the reliance on website features for identification, which may not be effective in identifying sophisticated phishing attacks.

4.2. *Implications and Applications*

The proposed approach has potential implications and applications in the context of online security. This approach can be used by organizations and individuals to identify phishing websites and prevent online fraud. The approach can also be extended to other domains such as email phishing, social engineering attacks, and malware detection.

5. Conclusion

Future research can focus on enhancing the proposed approach by incorporating additional features and using more advanced machine learning techniques. Additionally, the proposed approach can be extended to other domains such as email phishing, social engineering attacks, and malware detection. Further research can also explore the use of ensemble methods and deep learning techniques for identifying phishing attacks.

7. References

[1]. Anti-Phishing Working Group. (2021). "Phishing Activity Trends Report, 1st Quarter 2021." [Online].Available: https://apwg.org/reports/APWG_Phishing_Activity_Trends_Report_Q1_2021.pdf

[2]. X. Liu, J. Wang, C. Wang, and Y. Chen, "Phishing website detection based on decision tree," in Proceedings of the 3rd International Conference on Multimedia Technology (ICMT 2011), 2011, pp. 568-571.

[3]. A. Aggarwal and M. Kumar, "Phishing email detection using machine learning techniques," in Proceedings of the International Conference on Computational Intelligence and Communication Networks (CICN 2014), 2014, pp. 178-182.

[4]. A. Alzahrani, A. Alsuhibany, M. Alshahrani, N. Alzahrani, and S. Altowaijri, "A machine learning approach for phishing website detection," International Journal of Advanced Computer Science and Applications, vol. 8, no. 4, pp. 255-262, 2017.

[5]. H. Kaur and R. Rani, "Detection of phishing emails using machine learning algorithms," in Proceedings of the 2nd International Conference on Inventive Systems and Control (ICISC 2018), 2018, pp. 438-443.

[6]. G. Ramachandran and A. Suruliandi, "Phishing website detection using feedforward neural networks," International Journal of Pure and Applied Mathematics, vol. 116, no. 10, pp. 239-245, 2017.

[7]. S. Park, Y. Lee, H. Park, and H. Kim, "Phishing email detection using convolutional neural networks," in Proceedings of the International Conference on Information and Communication Technology Convergence (ICTC 2018), 2018, pp. 1-5.

[8]. W. Zhang, J. Wang, L. Zhang, and Y. Xu, "A novel approach to detect phishing webpages using support vector machines," International Journal of Security and Its Applications, vol. 7, no. 1, pp. 127-136, 2013.

[9]. A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," IEEE Communications Surveys & Tutorials, vol. 18, no. 2, pp. 1153-1176, 2015.