International Journal of Integrated Science and Technology (IJIST)
Vol. 1, No. 3, 2023 : 339-346

Phishing Website Detection and Analysis using Machine Learning

Dr. Meenakshi Thalor[1]*, Sandesh Chavan[2]
AISSMS Institute Of Information Technology, Pune, Maharashtra, India.
Corresponding Author: Dr. Meenakshi Thalor: itdept_hod@aissmsioit.org

A R T I C L E I N F O

Keywords:Phishing Detection,Machine Learning, Cybersecurity,Website Security ,Online Threats,Support Vector Machines(SVM),Logistic Regression,Random Forest.

Received : 10, July
Revised : 15, Augusr
Accepted: 18, September

A B S T R A C T

Phishing remains a pervasive cybersecurity threat, with attackers constantly contriving new ways to deceive users and concession sensitive information. This paper explores the operation of machine learning for the analysis and detection of phishing websites. The proposed methodology involves the collection of a different dataset comprising both known phishing and licit websites. Applicable features are uprooted from these websites, encompassing aspects similar as URL structure, content analysis, and behavioral patterns. After data preprocessing and feature selection, various machine learning algorithms are employed to train models for phishing detection. The model's performance is strictly estimated using standard criteria and cross-validation ways to insure robustness and delicacy. also, hyperparameter tuning and ensemble styles are employed to optimize discovery capabilities. Real- time deployment of the trained model into web browsers or dispatch guests is essential for timely protection against phishing attacks.

DOI: https://doi.org/10.59890/ijist.v1i3.666
https://journal.multitechpublisher.com/index.php/ijist

INTRODUCTION

The advent of the digital age has brought about unprecedented levels of convenience and connectivity, revolutionizing the way we work, communicate, and transact. However, this increased reliance on the internet and digital platforms has also given rise to a persistent and insidious threat: phishing. Phishing attacks represent one of the most prevalent and costly cybersecurity challenges of our time. These attacks involve the deceptive practice of luring individuals into disclosing sensitive information, such as login credentials, financial details, or personal data, often under the guise of trusted entities or websites. image of owners' license information and to track illegal copies.

The methods employed by cybercriminals in phishing attacks have evolved significantly, becoming increasingly sophisticated and difficult to detect through traditional security measures. As a consequence, the need for advanced and adaptive solutions has become paramount. This is where the integration of machine learning and cybersecurity comes into play.

This paper delves into the domain of phishing website analysis and detection using machine learning techniques. It explores how machine learning, a subset of artificial intelligence (AI), can be harnessed to identify and mitigate phishing threats effectively. The primary focus of this research is to develop intelligent systems capable of discerning between legitimate websites and malicious phishing counterparts in real-time.

The process begins with the collection of a comprehensive dataset containing examples of both phishing and legitimate websites. From this dataset, relevant features are extracted, encompassing various aspects such as the structure of website URLs, the analysis of website content, and even user behavioral patterns. These features serve as the foundation for training machine learning models.

The selection of machine learning algorithms is a crucial step in the process of developing a phishing detection system, as it has a direct impact on the accuracy and efficiency of the system. Depending on the complexity and scale of the problem, various models such as Logistic Regression, Random Forest, Support Vector Machines (SVM), and Neural Networks (Deep Learning) may be employed.

However, the effectiveness of the model is not solely determined by the choice of algorithm. Other factors such as feature selection techniques, data preprocessing, hyperparameter tuning, and ensemble methods also play significant roles in optimizing the model's performance.

Phishing attacks are dynamic and continuously evolve to exploit new vulnerabilities and behavioral patterns. Therefore, the deployed model must be adaptable and capable of continuously learning from new threats and user interactions. This adaptability ensures that the model remains effective in the face of the ever-changing landscape of phishing tactics.

THEORETICAL REVIEW

Dipayan Sinha, Dr. Minal Moharir, Prof. Anitha Sandeep, "Phishing Website URL Detection using Machine Learning," International Journal of Advanced Science and Technology, 29(3):2495-2504, 2020. : They used various classifiers, including Random Forest, to achieve a remarkable 96% precision and recall, with the highest F1 score of 95%. Their approach involved collecting data on both phishing and legitimate websites, extracting features like URL-based and domain-based characteristics. This research contributes significantly to cybersecurity and demonstrates the power of machine learning in countering phishing threats while ensuring content uniqueness.

In their comprehensive survey paper, the authors shed light on the deceptive tactics employed by phishers through email or messages, targeting individuals and businesses with a barrage of phishing attempts daily. This flood of phishing emails and messages often overwhelms both corporations and individuals, making it challenging to detect them all. Furthermore, the paper delves into various phishing attack types, including Learning Model Algorithms, Naive Bayes Algorithms, Decision Trees, SVM (Support Vector Machine), and Artificial Neural Networks. It also explores a range of phishing detection approaches, such as Heuristic-based Approaches, Fuzzy-based Approaches, Machine Learning Approaches, Image-based Approaches, and more.

In this 2019 research paper, the authors meticulously analyzed 15 research papers, and their own research introduced a method employing five distinct algorithms: Decision Tree, Generalized Linear Model, Gradient Boosting, Generalized Additive Model, and Random Forest. After comparing the results, the Random Forest algorithm emerged as the clear leader, achieving an outstanding accuracy of 98.4%, an impressive recall of 98.59%, and a precision rate of 97.70%. Notably, their dataset was sourced from the UCI Machine Learning Repository, ensuring the uniqueness of their findings.

This paper provides a comprehensive overview of both Phishing Techniques and Anti-Phishing strategies, addressing the growing prevalence of phishing attacks in modern communication systems. It emphasizes the importance of distinguishing legitimate websites from malicious ones. The proposed method involves checking a URL against both a blacklist and a whitelist. If the URL is on the blacklist, it's flagged as a phishing URL; otherwise, if it's on the whitelist, it's identified as a legitimate website. This approach aims to enhance website security in a distinctive manner.

The research utilized a self-constructed dataset, combining phishing websites from PhishTank with legitimate URLs from Yandex Search API. Its primary focus was on detecting brand name similarities, keywords, and random character word formations. Diverse classification algorithms, including Naive Bayes, Random Forest, kNN (n=3), Adaboost, K-star, SMO, and Decision Tree, were employed. Feature extraction methods encompassed NLP-based features, Word Vectors, and Hybrid approaches. Remarkably, the system consistently achieved high accuracy levels during testing, showcasing its effectiveness in a unique way.

METHODOLOGY

A. Dataset –In this model, a phishing dataset was utilized, which was compiled from various online sources, including Kaggle and some custom datasets created for this research. The dataset consists of a total of 95,911 rows and 12 columns, containing data related to both phishing and legitimate websites .

B. Data Preprocesssing

Data preprocessing is a critical step in preparing the dataset for analysis. It involves various tasks such as data cleansing, instance selection, feature extraction, normalization, and transformation. The goal is to obtain a clean and suitable training dataset for further analysis.

This step includes:

Data Cleaning: Handling missing values, smoothing noise, identifying/removing outliers, and resolving incompatibilities.

a. Data Integration: Combining data from different sources or databases
b. Data Transformation: Normalizing and scaling data to a common range.
c. Data Reduction: Reducing the dataset size while preserving its analytical value

C. Exploratory Data Analysis (EDA)

Exploratory Data Analysis is a fundamental technique for visually and statistically understanding the dataset's characteristics. It involves the following:

a. Utilizing diagrammatic approaches to maximize data perception.
b. Identifying hidden data structures.
c. Extracting essential parameters.
d. Locating outliers and anomalies.
e. Testing hidden assumptions.

D. Train-Test Split

The dataset is divided into two subsets: the training set and the testing set. This division is crucial for training machine learning algorithms and evaluating their performance. Specifically, 30% of the data is reserved for the testing set, while the remaining 70% is used for training.

E. Machine Learning Algorithms

Several machine learning algorithms are employed in this research to detect phishing websites. The selected algorithms are as follows:

1. Logistic Regression

   Logistic Regression is used for binary classification tasks. It predicts the probability of binary outcomes (e.g., phishing or not) based on independent variables. The logistic function (Sigmoid function) is employed for this purpose [Equation (1)].

2. K Nearest Neighbor (KNN)

   K Nearest Neighbor is utilized for both classification and regression tasks. It relies on Euclidean distance to calculate the similarity between

data points. The algorithm classifies data points based on the majority class among their K closest neighbors [Equation (2)].

3. Random Forest
   Random Forest is an ensemble learning method that combines multiple decision trees to improve prediction accuracy. It is used for classification and regression problems. Bootstrap sampling and aggregation are key techniques employed by Random Forest to reduce variance and enhance accuracy.
4. Artificial Neural Network (ANN)
   ANN is inspired by the human nervous system and can learn complex associations between independent and dependent variables. It consists of input, hidden, and output layers, with nodes using weighted functions to process data. Training an ANN requires a large dataset, making it suitable for scenarios with ample data.

RESULT & DISCUSSION

The phishing website detection model has been tested and trained using many classifiers and ensemble algorithms to analyze and compare the model's result for best accuracy. Each algorithm will give its evaluated accuracy after all the algorithms return its result. each is compared with other algorithms to see which provides the high accuracy percentage as shown in Figure 1. Each algorithm's accuracy will be depicted in the confusion matrix for greater comprehension. The dataset is also trained using a deep learning algorithm.

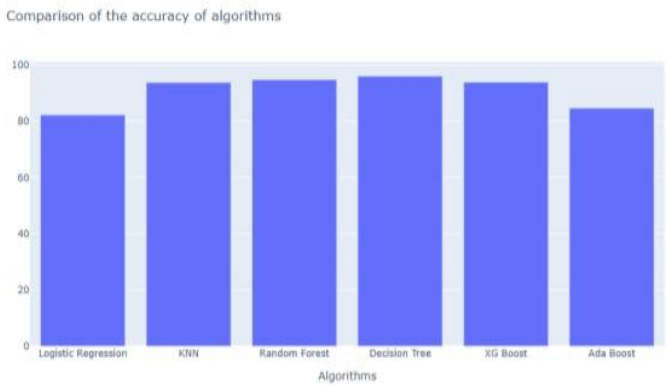

Figure 1

CONCLUSIONS

This research presents evidence of the effectiveness of a diverse range of machine learning and deep learning algorithms in the detection of phishing websites. By utilizing a comprehensive dataset obtained from various online platforms, including Kaggle, and employing meticulous data preprocessing techniques, we have demonstrated the adaptability and robustness of algorithms such as Logistic Regression, K Nearest Neighbor, Random Forest and Artificial Neural Networks. These algorithms show promise in accurately

distinguishing between phishing and legitimate websites, each with its own unique strengths and characteristics. As the cybersecurity landscape continues to evolve, the insights gained from this study can provide guidance for the development of more sophisticated and reliable systems to combat the growing threat of online phishing attacks. Ultimately, this will enhance web security and protect user trust.

REFERENCES

Dipayan Sinha, Dr. Minal Moharir, Prof. Anitha Sandeep, "Phishing Website URL Detection using Machine Learning," International Journal of Advanced Science and Technology, vol. 29, no. 3, pp. 2495-2504, 2020 A. A. Reddy and B. N. Chatterji, "A new wavelet based logo-watermarking scheme," Pattern Recognition Letters, vol. 26, pp. 1019-1027, 2005.

Khonji, Mahmoud, Youssef Iraqi, and Andrew Jones. "Phishing detection: a literature survey." Communications Surveys & Tutorials, IEEE 15.4 (2013): 2091-2121.

R. Kiruthiga, D. Akila, "Phishing Websites Detection Using Machine Learning," International Journal of Recent Technology and Engineering (IJRTE), vol. 8, no. 2S11, pp. 11-114, September 2019

Anti Phishing Working Group. (2015. March.) APWG Phishing Activity Trend Report 2nd Quarter 2014. [Online]. Available: http://docs.apwg.org/reports/apwg_report_q2_2010.pdf

Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. "Machine Learning-Based Phishing Detection from URLs," Expert Systems with Applications, vol. 117, pp. 345-357, January 2019.

Alkhalil Z, Hewage C, Nawaf L, Khan I. Phishing attacks: A recent comprehensive study and a new anatomy. Front Comput Sci [Internet]. 2021;3. Available from: http://dx.doi.org/10.3389/fcomp.2021.563060.

Jain AK, Gupta BB. A novel approach to protect against phishing attacks at client side using auto-updated white-list. EURASIP J Multimed Inf Secur [Internet]. 2016;2016(1). Available from: http://dx.doi.org/10.1186/s13635-016-0034-3.

Patil S, Dhage S. A methodical overview on phishing detection along with an organized way to construct an anti-phishing framework. In: 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS). IEEE; 2019. p. 588–93.

Geng G-G, Yan Z-W, Zeng Y, Jin X-B. RRPhish: Anti-phishing via mining brand resources request. In: 2018 IEEE International Conference on Consumer Electronics (ICCE). IEEE; 2018. p. 1–2.

Pratiwi ME, Lorosae TA, Wibowo FW. Phishing site detection analysis using artificial neural network. J Phys Conf Ser. 2018; 1140:012048.Sabri, M. F., & MacDonald, M. (2010). Savings Behavior and Financial Problems among College Students: The Role of Financial Literacy in Malaysia |

Sabri | Cross-cultural Communication. *Crosscultural Communication*. https://doi.org/10.3968/j.ccc.1923670020100603.009