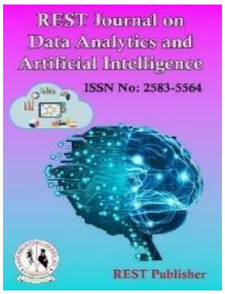
REST Journal on Data Analytics and Artificial Intelligence

Vol: 3(3), September 2024

REST Publisher; ISSN: 2583-5564

Website: http: //restpublisher.com/journals/jdaai/

DOI: https://doi.org/10.46632/jdaai/3/3/15



Malicious Website Prediction Using Machine Learning Methodologies

R. Sravani, M.V. Subba Reddy, M. Nagaprarthana Devi, *G. Ramasubba reddy

Sai Rajeswari Institute of Technology, Proddatur, Andhra Pradesh, India

*Corresponding author: subbareddy1227@gmail.com

Abstract: *Nowadays, digital technology has advanced faster than any previous invention, leading to widespread use of machine learning algorithms for generating predictions or decisions without explicit programming. These algorithms rely on a sample set of data, known as training data, to function effectively. However, the absence of high-quality data poses a significant challenge in machine learning, as data quality is crucial for the algorithm's performance. In phishing detection, the ultimate accuracy depends on various key features, including the URL, domain identity, security, and encryption criteria. To extract and verify these criteria from phishing data sets, we utilize regression techniques and classification algorithms. Specifically, we employ decision tree and logistic regression methods as two machine learning techniques. Logistic regression, a standard approach for binary classification problems, originates from the statistical discipline and achieves a 95% accuracy rate on trained data sets. Decision trees, a form of supervised machine learning, continuously split data based on specific parameters and consist of decision nodes and leaves, representing choices and outcomes, respectively. Decision trees achieve an 85% accuracy rate on trained data.*

Keywords: *Phishing Detection, regression technique, Decision Tree Method, Logistic Regression.*

1. INTRODUCTION

Artificial intelligence encompasses a subfield called machine learning, which describes the ability of IT systems to solve problems autonomously by identifying patterns within databases. In essence, machine learning enables IT systems to detect patterns using pre-existing algorithms and data sets, subsequently generating suitable solution concepts. Through this process, machine learning creates artificial knowledge from experience. The primary objective of machine learning is to understand data structures and incorporate them into models that people can use and comprehend. Machine learning, while a subset of computer science, operates differently from traditional computational methods. Traditional computing involves using predefined instructions, or algorithms, to address problems. On the other hand, machine learning techniques enable systems to learn from input data and perform statistical analyses to generate values within a set range. This approach allows computers to create models from sampled data, which supports automated decision-making based on these data inputs. Identifying phishing websites is crucial to protecting legitimate websites and their users from various malicious activities. Adversaries often disguise harmful URLs as legitimate ones to deceive unsuspecting users, leading to unethical activities such as stealing private and personal data from user devices, resulting in substantial global losses annually. In this study, we employ a machine learning algorithm to classify URLs based on their characteristics and behavior. To identify malicious URLs, we collect both malicious and benign URLs, labeling the former as dishonest and the latter as honest. These datasets, stored in a CSV file, include numerous annotated URLs. The process begins with tokenizing the URLs, followed by loading and storing the data in a list. This list is then used to vectorize the URLs using tf-idf scores for classification, rather than a bag of words. Next, we apply logistic regression in conjunction with the decision tree method, dividing the data into training and testing sets. The results are evaluated based on the accuracy of the test data, and predictions are made accordingly. Many customers use various websites to pay for online purchases. Given the vast amount of private information disclosed on social networking sites, incidents of user privacy violations are common. By using this website, we can determine which URLs are legitimate. This helps users make confident online purchases. Additionally, this website can be utilized by various e-commerce businesses to secure entire transactions, ensuring a safer online shopping experience.

2. MATERIALS AND METHOD

Phishing is a cyber-attack that uses advanced social engineering tactics and other methods to extract personal information from website visitors. According to the Anti-Phishing Working Group, the number of unique phishing websites identified has increased at an average annual rate of 36.29% over the past six years and 97.36% in the past two years. This rise has intensified the focus of the cybersecurity sector on mitigating phishing attacks. Significant research and development efforts have targeted phishing attempts, analysing their unique content, network, and URL features. The intuitions, data analysis techniques, and evaluation processes differ significantly among existing methods, necessitating a comprehensive systematization to compare the advantages and disadvantages of each approach and their applicability in various contexts. This approach provides an in-depth analysis of phishing detection techniques, particularly software-based ones, by examining evaluation datasets, detection features, detection methods, and evaluation metrics, beginning with a taxonomy of phishing detection. Ultimately, this method offers insights aimed at developing more effective and efficient phishing detection systems. Social networks have become widely popular platforms for user interaction, making the protection of user privacy on these sites a significant area of research due to the extensive amount of personal information shared. Phishing attacks continue to be a common technique for stealing information, resulting in numerous privacy breaches. In these web-based phishing attacks, hackers design counterfeit web pages that resemble prominent websites, such as social media platforms, in order to deceive users into providing sensitive information like passwords, credit card numbers, and social security numbers. To address these threats, an effective phishing alarm system should incorporate elements that are challenging for attackers to circumvent. One specific algorithm evaluates the level of suspicion of web pages based on their visual similarity. This approach uses cascading style sheets (CSS) as a foundation to accurately assess the visual similarity of each page element since CSS governs page layout across different browser implementations. The grading system is based on the weighted page-component similarity, recognizing that not all page elements impact pages equally. This strategy was prototyped in the Chrome browser. Hackers continually develop innovative techniques to defeat existing defences. Since phishing attempts are constantly evolving, it is essential to develop new and effective responses to these ever-changing threats, as their effects can be devastating. Artificial intelligence techniques have become a cornerstone of contemporary countermeasures used to mitigate phishing assaults. However, AI-based phishing defences face drawbacks, including a high false alarm rate and difficulty in understanding how most phishing tactics operate. Various methods have chosen important website features based on user experience or frequency analysis. To enhance the identification of phishing websites, a particle swarm optimization (PSO)-based feature weighting approach is introduced for effective phishing detection. This method employs PSO to assign appropriate weights to different website characteristics, ultimately improving the accuracy of phishing website detection.

3. PROPOSED WORK

The growing prevalence of cyber threats necessitates the development of robust systems capable of predicting and identifying malicious websites. This proposed work aims to utilize machine learning methodologies to accurately predict malicious websites by creating a model that can distinguish between benign and harmful sites based on various features extracted from URLs and website content. The primary steps in this approach include data collection, feature extraction, model selection, training, evaluation, and implementation. This multi-step process is designed to ensure the creation of a reliable and efficient predictive system.

Data collection is the foundational step in this project, involving the compilation of a comprehensive dataset that includes both malicious and benign websites. The data will be sourced from publicly available datasets from repositories such as Kaggle and the Open Threat Exchange (OTX), web scraping techniques to gather URLs and associated metadata, and crowdsourced platforms where users report suspicious or malicious websites. This diverse data collection strategy ensures a rich dataset that enhances the model’s ability to learn and generalize across various types of websites. Feature extraction is crucial for preparing the dataset for machine learning algorithms. Features will be derived from URLs, website content, and network-based information. URL-based features include the length of the URL, the number of special characters, the presence of an IP address, the number of subdomains, and the use of suspicious words or domains. Content-based features involve the presence of embedded malicious scripts, analysis of HTML tags and structure, frequency and type of external links, use of obfuscation techniques, and text analysis for phishing keywords. Network-based features include WHOIS information, hosting server details, SSL certificate validity, and IP reputation. These features collectively provide a comprehensive view of the website's characteristics, enhancing the model's predictive power. Model selection and training involve assessing various machine learning models to determine their effectiveness in predicting malicious websites. We will consider supervised learning models, including logistic regression, decision trees, random forests, support vector machines (SVM), and gradient boosting machines (GBM). Additionally, we will examine ensemble methods that combine predictions from multiple models to enhance overall performance and minimize overfitting. The dataset will be divided into training and testing sets, and cross-validation will be utilized to ensure the models are robust. Feature selection techniques, such as Recursive Feature Elimination (RFE) and feature importance scores from tree-based models, will be employed to identify the most relevant features, thereby optimizing the model's performance.

The trained models will be assessed using a range of metrics to evaluate their performance. Accuracy will indicate the overall correctness of the model, while precision and recall will measure the model's effectiveness in identifying malicious websites in terms of relevance and completeness. The F1-score, which represents the harmonic mean of precision and recall, offers a consolidated performance measure. Furthermore, the ROC-AUC (Receiver Operating Characteristic - Area Under Curve) will be used to assess the model's capability to differentiate between classes. Confusion matrices will help visualize true positives, true negatives, false positives, and false negatives, providing valuable insights into the model's strengths and weaknesses. Upon selecting the optimal model, it will be implemented in a real-world scenario. Implementation steps include developing browser plugins or extensions to analyze URLs in real-time and warn users about potential threats, creating a RESTful API that allows external applications to submit URLs for analysis and receive predictions, and ensuring periodic updates to the model with new data to adapt to evolving cyber threats. This real-time and integrated approach ensures that users and systems remain protected against malicious websites.

The deployed model will be continuously monitored to ensure its effectiveness in predicting malicious websites. Performance monitoring will track the model’s metrics over time, while a feedback loop will incorporate user feedback to improve accuracy. Additionally, the model will be regularly retrained with new data to keep up with emerging cyber threats, ensuring it remains a robust and reliable tool for cybersecurity. This proposed work aims to develop a machine learning-based system for predicting malicious websites, leveraging comprehensive data collection, advanced feature extraction, and robust model selection and evaluation processes. By implementing and continuously updating this system, it will effectively differentiate between benign and harmful websites, contributing to a safer online environment and enhancing overall cybersecurity measures.

4. IMPLEMENTATION

Method 1: Collecting Datasets

- **Code:** url_data = pd. read_csv("urldata.csv")
- **Method:** Collecting Datasets
- **Input:** URL
- **Output:** URL_DATA
- **Process:** The URL is collected from Google and stored in our system database.

Importance: The initial stage of any machine learning project involves collecting high-quality data. For phishing detection, it is essential to compile a thorough dataset that encompasses both legitimate and malicious URLs. This dataset forms the basis for training and testing the machine learning models. By using a diverse and representative set of URLs, we can ensure that our models learn to identify phishing attempts accurately and generalize well to new, unseen data.

Method 2: Tokenization

- **Code:** tkns_BySlash = str (f. encode('utf-8')). split ('/')
- **Method:** Tokenization
- **Input Parameter:** URL
- **Output Parameter:** Splitted Data
- **Process:** This tokenization method is used to split the URL by characters like '/', '#', '%', '@', '&', etc., making it easier to identify whether the URL is good or bad.

Importance: Tokenization is a reprocessing step that breaks down the URL into smaller components, or tokens, which can be analysed individually. This process helps in identifying patterns and features within the URL that may indicate whether it is legitimate or malicious. By splitting the URL into meaningful segments, we can extract valuable information that contributes to the detection process, making the subsequent analysis more effective.

Method 3: Vectorization

- **Code:** vectorizer = TfidfVectorizer (tokenizer=make Tokens)
- **Method:** Vectorizer
- **Input Parameter:** URL
- **Output Parameter:** URL List
- **Process:** This method counts the number of times a token appears in the document and uses this value as its weight. TF-IDF means the weight allocated to each token depends not only on its frequency in a document but also on how persistent that term is in the entire corpus.

Importance: Vectorization converts the tokenized URLs into numerical formats that machine learning models can understand. The TF-IDF (Term Frequency-Inverse Document Frequency) vectorizer assigns weights to tokens according to their frequency and significance, which improves the model's ability to distinguish between relevant and irrelevant features. This process is crucial for transforming textual data into a format that is compatible with machine learning algorithms, thereby enhancing the accuracy and efficiency of the detection process.

Method 4: Decision Tree Algorithm

- **Code:** DecisionTreeClassifier (class weight=None, criterion='gini', max_depth=3)
- **Method:** Decision Tree Algorithm
- **Input Parameter:** Trained Dataset
- **Output Parameter:** Accuracy
- **Process:** The decision tree algorithm is used to give the accuracy of the URL and check whether it is a good URL or a phishing URL.

Importance: Decision trees are a widely used supervised learning technique for classification tasks. They operate by recursively dividing the dataset according to certain features, forming a tree-like arrangement of decision nodes and leaves. Each node signifies a decision made based on a feature, while each leaf corresponds to a class label (such as legitimate or phishing URL). Decision trees are straightforward to interpret and visualize, which helps in understanding the model's decision-making process.

Method 5: Logistic Regression Algorithm

- **Code:** LogisticRegression (C=1.0, class weight=None, dual=False, fit intercept=True)
- **Method:** Logistic Regression Algorithm
- **Input Parameter:** Trained Datasets
- **Output Parameter:** Accuracy
- **Process:** This algorithm is also used to find the accuracy of the trained dataset and determine whether the URL is good to use in the environment.

Importance: Logistic regression is a statistical technique utilized for binary classification tasks. It calculates the likelihood that a specific input falls into a particular category (such as legitimate or phishing URL) using a logistic function. This method is advantageous because it offers a probabilistic approach to classification, enabling the assessment of confidence levels in predictions. Additionally, logistic regression is computationally efficient, making it appropriate for handling large datasets, and it tends to perform well when there is a linear relationship between the input features and the class labels. In the context of phishing detection, logistic regression can effectively distinguish between legitimate and malicious URLs based on their features.

5. RESULTS AND DISCUSSION

In the context of phishing detection, the performance of classification algorithms is critical to ensuring accurate and reliable identification of malicious URLs. Two prominent machine learning techniques, logistic regression and decision tree algorithms, have been evaluated based on several key metrics: True Positive Rate (TPR), True Negative Rate (TNR), overall accuracy, and precision. Logistic regression achieved a True Positive Rate of 91.23%, a True Negative Rate of 92.21%, an overall accuracy of 92.85%, and a precision of 93.50%. These results indicate that logistic regression is highly effective in identifying legitimate URLs while maintaining a high level of precision, minimizing false positives. On the other hand, the decision tree algorithm demonstrated a True Positive Rate of 90.92%, a True Negative Rate of 95.02%, an overall accuracy of 93.95%, and a precision of 95.56%. The decision tree outperformed logistic regression in terms of True Negative Rate, overall accuracy, and precision, showcasing its strength in correctly identifying phishing URLs and providing reliable classification results. Both algorithms offer robust solutions for phishing detection, with the decision tree algorithm slightly outperforming logistic regression in key areas. This comparison highlights the importance of using multiple metrics to evaluate the effectiveness of machine learning models in cybersecurity applications.

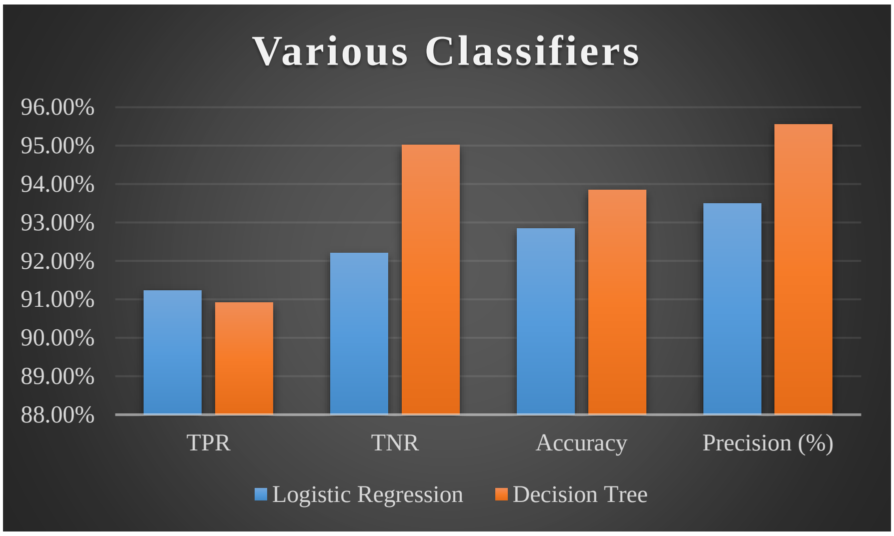


FIGURE. 1

6. CONCLUSION

Phishing is a method of obtaining users' private information via email or website. As internet usage is increasing quickly, virtually everything is now available online. Since phishing attacks utilise new tactics every day as technology advances, it is now possible to identify phishing websites online for detection in order to prevent the problem of privacy. In conclusion, the implementation of machine learning techniques, specifically logistic regression and decision tree algorithms, for phishing detection has demonstrated promising results in accurately identifying malicious URLs. Logistic regression achieved commendable performance metrics, with a True Positive Rate of 91.23%, a True Negative Rate of 92.21%, an overall accuracy of 92.85%, and a precision of 93.50%. Meanwhile, the decision tree algorithm showed superior performance in several key areas, achieving a True Positive Rate of 90.92%, a True Negative Rate of 95.02%, an overall accuracy of 93.95%, and a precision of 95.56%. These findings underscore the effectiveness of both algorithms in enhancing cybersecurity measures by accurately distinguishing between legitimate and phishing URLs. The decision tree algorithm's higher True Negative Rate, overall accuracy, and precision highlight its robustness in phishing detection scenarios. Consequently, incorporating these machine learning models into cybersecurity frameworks can significantly improve the detection and prevention of phishing attacks, safeguarding users and their sensitive information from cyber threats. The study reinforces the critical role of machine learning in developing advanced, reliable, and efficient cybersecurity solutions.

REFERENCES

[1]. Niruban, R., Deepa, R., Vignesh, G. D., (2020), "A novel iterative demosaicing algorithm using fuzzy based dual tree wavelet transform," Journal of Critical Reviews, vol. 7, pp. 141-145. doi:10.31838/jcr.07.09.27.

[2]. Rajesh G., Mercilin Raajini X., Ashoka Rajan R., Gokuldhev M., Swetha C., (2020), "A Multi-Objective Routing Optimization Using Swarm Intelligence in IoT Networks," Lecture Notes in Networks and Systems, vol. 118, no., pp. 603-613. doi:10.1007/978-981-15-3284-9_69.

[3]. A. Belabed, E. Aimeur, and A. Chikh, “A personalized whitelist approach for phishing webpage detection,” in Proc. 7th Int. Conf. Availability, Rel. Security (ARES), Aug. 2012, pp. 249–254.

[4]. T.-C. Chen, S. Dick, and J. Miller, “Detecting visually similar Web pages: Application to phishing detection,” ACM Trans. Internet Technology, vol. 10, no. 2, pp. 1–38, May 2010.

[5]. N. Chou, R. Ledesma, Y. Teraguchi, D. Boneh, and J. C. Mitchell, “Client-side defense against Web-based identity theft,” in Proc. 11th Annu. Network Distribution System Security Symp. (NDSS), 2004,

pp. 1–16.

[6]. Z. Dong, K. Kane, and L. J. Camp, “Phishing in smooth waters: The state of banking certificates in the US,” in Proc. Res. Conf. Common., Inf. Internet Policy (TPRC), 2014, p. 16.

[7]. J. Corbett, L. Invernizzi, C. Kruegel, and G. Vigna, “Eyes of a human, eyes of a program: Leveraging different views of the Web for analysis and detection,” in Proceedings of Research in Attacks, Intrusions and Defenses (RAID), Gothenburg, Sweden: Springer, 2014.

[8]. Rajesh G., Mercilin Raajini X., Ashoka Rajan R., Gokuldhev M., Swetha C., (2020), "A Multi-Objective Routing Optimization Using Swarm Intelligence in IoT Networks," Lecture Notes in Networks and Systems, vol. 118, no., pp. 603-613. doi:10.1007/978-981-15-3284-9_69.

[9]. Mythili V., Kaliyappan M., Hariharan S., Dhanasekar S., (2018), "A new approach for solving travelling salesman problem with fuzzy numbers using dynamic programming," International Journal of Mechanical Engineering and Technology, vol. 9, no. 11, pp. 954-966.

[10]. Kohila S., Malliga G. S., (2017), "Classification of the Thyroiditis based on characteristic sonographic textural features and correlated histopathology results," 2016 IEEE International Conference on Signal and Image Processing, ICSIP 2016, vol., no., pp. 305-309. doi:10.1109/SIPROCESS.2016.7888273.

[11]. Anbuchezhiyan M., Arputhalatha A., Ponnusamy S., Syed Suresh Babu K., (2015), "Effect of phosphorous on the growth, optical, mechanical and thermal properties of L-alanine crystals," Photonics Letters of Poland, vol. 7, no. 2, pp. 44-46. doi:10.4302/plp.2015.2.05.

[12]. Umapathy K., Balaji V., Duraisamy V., Saravanakumar S. S., (2015), "Performance of wavelet based medical image fusion on FPGA using high level language C," Jurnal Teknologi, vol. 76, no. 12, pp. 105-109. doi:10.11113/jt. v76.5888.

[13]. Shanmugaraj M., Vishal J., Rahul G., (2014), "Analysis of oxygen enriched combustion technology in a single cylinder DI diesel engine," Applied Mechanics and Materials, vol. 592-594, no., pp. 1433-1437. doi:10.4028/www.scientific.net/AMM.592-594.1433.

[14]. Kavitha K. V. N., Ashok S., Imoize A. L., Ojo S., Selvan K. S., Ahanger T. A., Alhassan M., (2022), "On the Use of Wavelet Domain and Machine Learning for the Analysis of Epileptic Seizure Detection from EEG Signals," Journal of Healthcare Engineering, vol. 2022, no., pp.-. doi:10.1155/2022/8928021.

[15]. Sowmya, S., Kannan, K. N., Anbu, S., Veeralakshmi, P., Kapilavani, R. K., (2022), "Preventing collaborative attacks against on demand routing using recommendation based trust framework in MANET," AIP Conference Proceedings, vol. 2393, pp.-. doi:10.1063/5.0079725.

[16]. Veeralakshmi, P., Sowmya, S., Kannan, K. N., Anbu, S., Ayyappan, G., (2022), "An efficient and smart IoT based pisciculture for developing countries," AIP Conference Proceedings, vol. 2393, pp.-. doi:10.1063/5.0074418.

[17]. Benila, A., Priyadharshini, R. I., Slacer, P. P., Jacob, J. J., Theresa, M. M., (2022), "Plan and development of efficient branch predictor for in-order RISC-V processor," AIP Conference Proceedings, vol. 2393, pp.-. doi:10.1063/5.0074195.

[18]. Chandra, K. Ram, M. Ramachandran, Sathiyaraj Chinnasamy, and Manjula Selvam. "Recent trends in Workplace Learning Methodology." *Contemporaneity of Language and Literature in the Robotized Millennium* 4, no. 1 (2022): 28-36.

[19]. Kannan, K. N., Anbu, S., Veeralakshmi, P., Sowmya, S., Reena, R., (2022), "Garden nurture using internet of things (IoT)," AIP Conference Proceedings, vol. 2393, pp.-. doi:10.1063/5.0074517.

[20]. Theresa, M. M., Rajam, P. P., Benila, A., Priyadharshini, R. I., Jacob, J. J., (2022), "An intelligent bike safety and surveillance system," AIP Conference Proceedings, vol. 2393, pp.-. doi:10.1063/5.0074199.

[21]. Selvapandian, D., Jacob, J. J., Kannamma, R., Dhanapal, R., Immanuel, J. D., (2020), "An efficient bidirectional broadcasting using signal initiation and data aggregation for WSN," Proceedings of the 3rd International Conference on Intelligent Sustainable Systems, ICISS 2020, vol., pp. 1367-1371. doi:10.1109/ICISS49785.2020.9315941.