IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 3, April 2023
Impact Factor: 7.301

Phishing Website Detection using Machine Learning Rules with Cryptography Technique

Balaji K[1], Iyappan K[2], Manoj Kumar T[3], Grace Mary S[4]
Students, Department of Computer Science Engineering[1,2,3]
Assistant Professor, Department of Computer Science Engineering[4]
Anjalai Ammal Mahalingam Engineering College, Thiruvarur, India

Abstract: *Phishing, a type of cybersecurity attack aimed at stealing personal information like passwords and credit card numbers, can be mitigated through the use of machine learning techniques for detecting phishing websites. In response, cybersecurity experts are actively seeking reliable and robust detection techniques to identify phishing websites. This research paper focuses on the implementation of machine learning technology for detecting phishing URLs by extracting and analyzing various features from both legitimate and phishing URLs. Support Vector Machine algorithms are employed for this purpose, to identify the most effective algorithm based on accuracy rate, false positive rate, and false negative rate, to enhance phishing detection measures. Cryptography Technique like the RSA algorithm is used to secure encryption method to encrypt the user search data before being stored on the server. The paper further discusses the reasons why ensemble machine learning methods are well-suited for binary phishing classification challenges in real-time detection scenarios, underscoring their efficacy in anti-phishing techniques.*

Keywords: Phishing, cybersecurity, Support Vector Machine algorithms, RSA algorithm rism

I. INTRODUCTION

Phishing attacks mimic the appearance and characteristics of legitimate emails, making them appear similar to the source. This can deceive users into visiting phishing websites via links provided in the emails, which are designed to look like authentic organization websites. Phishers use alarming or urgent messages to coerce users into providing personal information, which can be misused. In the training phase, it is crucial to use labeled data that includes samples of both phishing and legitimate URLs to develop an effective detection model. The dataset used for machine learning should consist of labeled samples that accurately represent phishing and legitimate URLs. There are various machine learning algorithms available, each with its mechanism, and using multiple algorithms can improve the accuracy of prediction in detecting phishing URLs. The existing system has good accuracy but there is still room for improvement to effectively combat phishing attacks, In this approach utilizes supervised machine learning, specifically SVM, to detect malware. It enhances the traditional signature-based detection system by incorporating behavior-monitoring techniques. Additionally, it fully supports selective aggregate functions for analyzing online user behavior while ensuring differential privacy.

II. LITERATURE SURVEY

Extensive research has been conducted in the field of phishing detection by numerous researchers. Through a comprehensive review of existing literature, we have gathered valuable insights and leveraged them to inspire our own methodologies, aimed at developing a more secure and accurate system.

Sahingoz, O. K., Buber, E., Demir, O., and Diri, B. ,[1] The dataset utilized in this study was created in-house, comprising phishing websites obtained from PhishTank, and legitimate URLs obtained from Yandex Search API. The focus was on detecting brand name similarity, keywords, and randomly generated character combinations. Multiple classification algorithms, including Naive Bayes, Random Forest, KNN (n=3), Adaboost, K-star, SMO, and Decision Tree, along with feature extraction techniques such as NLP-based features, Word Vectors, and Hybrid, were employed. The experimental results demonstrated consistently high accuracy levels during the testing phase.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 3, April 2023
Impact Factor: 7.301

J. James, Sandhya L., and C. Thomas [2] The proposed system employed a combination of lexical features, host properties, and page-related properties for phishing website detection. Data mining algorithms were used to gain insights into URL patterns, with classification algorithms such as Naïve Bayes, J48 Decision Tree, K-NN, and SVM being considered. Among these, Decision Tree exhibited the highest accuracy of 91.08% compared to other algorithms. Therefore, Tree-based classifiers were found to be the most suitable for phishing URL classification based on the experimental results.

Pradeepthi, K. V., and Kannan, A.,[3] The system employs classification algorithms to detect phishing URLs based on their URL structure, without accessing the actual content of the URLs. A dataset of 4500 URL records is used, consisting of 2500 genuine URLs collected from the DMOZ repository and 2000 phishing URLs from PHISHTANK. The dataset undergoes feature selection and classification during the training phase. Classification is performed using algorithms such as Naive Bayes, Random Forest, Random Tree, Multi-layer Perceptron, C-RT, J 48 Tree, LMT, C 4.5, ID 3, and K-Nearest Neighbor. The Random Forest Algorithm yields the highest classification accuracy

Dipayan Sinha, Dr. Minal Moharir, and Prof. AnithaSandeep[4] Machine learning techniques, such as Logistic Regression, Decision Tree, Random Forest, Adaboost, Gradient Boosting, Gaussian NB, and Fuzzy Pattern Tree Classifier, were utilized for detecting phishing websites. The data collection process involved gathering URLs from both phishing and legitimate websites. Relevant features were extracted from the URLs, including IP Address, presence of '@' symbol, dashes, long URL, unusual numbers, dot count, sub-domains, as well as domain-based features such as Page Rank, Domain age, and Website validity. Among the algorithms tested, the Random Forest algorithm demonstrated the highest precision, recall, and F1-score, achieving an impressive 96% precision and recall, and a F1-score of 95%.

R. Kiruthiga and D. Akila[5] This research paper reviewed a total of 15 previous studies, and proposed a method that employs five different algorithms, namely Decision Tree, Generalized Linear Model, Gradient Boosting, Generalized Additive Model, and Random Forest. Among these, the Random Forest algorithm demonstrated the highest accuracy of 98.4%, recall of 98.59%, and precision of 97.70%. The dataset utilized in this study was obtained from the UCI machine learning repository.

III. MACHINE LEARNING ALGORITHM

Support Vector Machine (SVM)

SVM is a widely used Supervised Learning algorithm in Machine Learning, known for its versatility in handling both Classification and Regression problems. SVM aims to create an optimal decision boundary or hyperplane in n-dimensional space, effectively segregating data points into different classes. This hyperplane allows for the accurate categorization of new data points in the future.

The process for detecting phishing websites involves several steps. First, a list of URLs to be checked for phishing is collected. For each URL, the RSA algorithm is used to securely access the website content. Features are then extracted from the website content, including the URL structure, content type, and presence of specific keywords or phrases. Subsequently, SVM is utilized to classify the website as either a phishing website or not, based on the extracted features. Finally, reports or notifications are generated and sent to the relevant parties based on the outcomes of the classification process.

IV. CRYPTOGRAPHY TECHNIQUE

4.1 RSA algorithm

The RSA algorithm, invented by Rivest, Shamir, and Adleman in 1978, is a widely recognized public key encryption technique known for its high level of security. As an asymmetric cryptography algorithm, RSA employs two distinct keys - the Public Key, which is shared with everyone, and the Private Key, which is kept confidential.

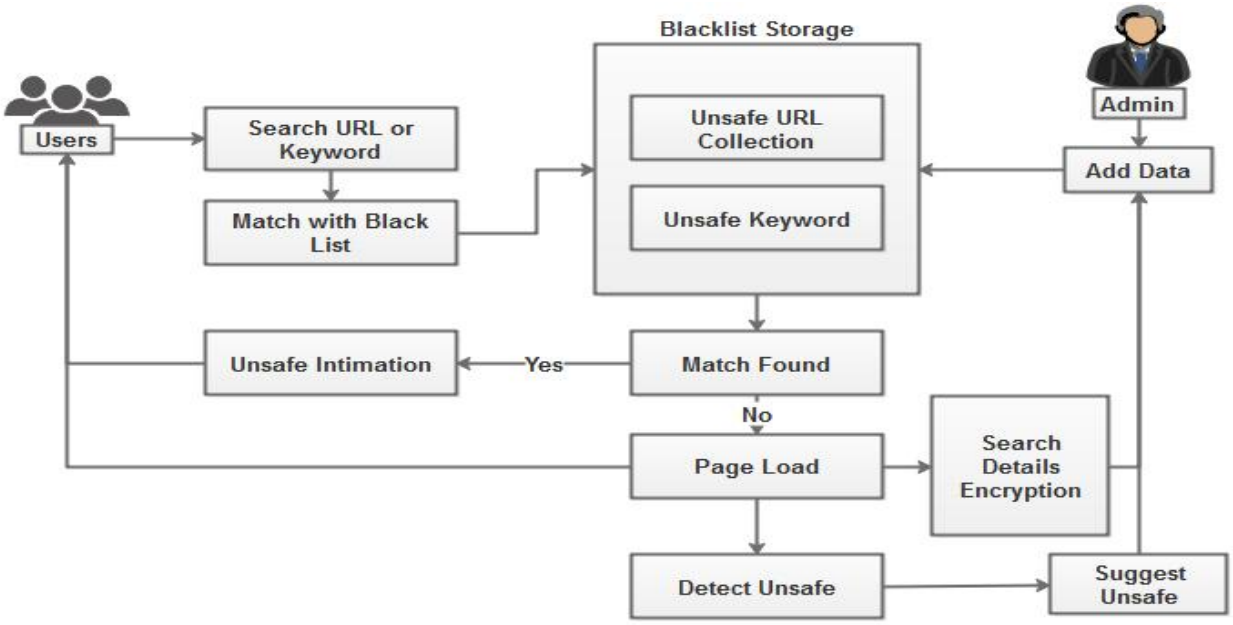
In RSA, the Public Key is used for encrypting data, while the Private Key is used for decryption. For instance, a client like a browser can send its public key to a server to request data. The server then encrypts the data using the client's public key and transmits the encrypted data. The client, in turn, uses its private key to decrypt the received data, ensuring secure communication.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 3, April 2023
Impact Factor: 7.301

V. PROPOSED SYSTEM

Traditional anti-phishing approaches rely on blacklist methods or feature-based machine learning techniques, which have limitations such as the inability to detect new phishing attacks and high false positive rates. Additionally, existing methods often extract features from third-party sources such as search engines, making them complex, slow, and unsuitable for real-time environments. To address these challenges, this paper introduces a novel machine learning-based anti-phishing approach that extracts features exclusively from the client side, offering a more efficient and effective solution.

First user search for the URL or keyword then the system match the URL with the blacklist, blacklist storage contain unsafe URL and keywords which already detected URLs, when a match has occurred it automatically sent unsafe Intimation to the user, which doesn't mean the page will be loaded and detected with the machine learning mechanism and it classifies with SVM mechanism when it looks like a phishing website it will automatically add in blacklist storage and the search details are encrypted safely with the RSA Algorithm



VI. CONCLUSION AND FUTURE SCOPE

In the future, with access to structured datasets of phishing URLs, we can further optimize the speed and accuracy of our detection technique. Additionally, combining multiple classifiers or exploring other phishing detection techniques, such as lexical features, network-based features, content-based features, web page-based features, and HTML and JavaScript features, may further improve system performance. Specifically, our approach focuses on extracting features from URLs and leveraging various classifiers for enhanced accuracy.

VII. ACKNOWLEDGMENT

I would like to take this opportunity to express my heartfelt gratitude to all those who have supported me throughout the research project report. I am truly thankful for their unwavering guidance, invaluable constructive criticism, and friendly advice during the course of my project work. Their honest and insightful views on various project-related matters have been immensely helpful. I am also grateful to Principal Dr. S.N. Ramaswamy and the management of AAMEC for their continuous support and encouragement. My sincere indebtedness goes to my Head of Department, Dr. K. Velmurugan, and my guide, Asst. Professor Mrs. S. GraceMary, for their unwavering guidance, constant supervision, and provision of necessary information throughout the project. I am also grateful to the review committee for their valuable suggestions and feedback. Furthermore, I extend my thanks to the laboratory staff for their valuable support. Last but not least, I sincerely appreciate the teaching and non-teaching staff from AAMEC who have contributed in various ways to my endeavor.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Impact Factor: 7.301
Volume 3, Issue 3, April 2023

REFERENCES

[1]. "Machine Learning-Based Phishing Detection from URLs" authored by Sahingoz, O. K., Buber, E., Demir, O., and Diri, B., published in the journal "Expert Systems with Applications" in January 2019, volume 117, pages 345-357.

[2]. "Detection of phishing URLs using machine learning techniques" authored by J. James, Sandhya L., and C. Thomas, published in the International Conference on Control Communication and Computing (ICCC) in December 2013.

[3]. Performance study of classification techniques for phishing URL detection" authored by Pradeepthi, K. V., and Kannan, A., published in the Sixth International Conference on Advanced Computing (IcoAC) in December 2014.

[4]. "Dipayan Sinha, Dr. Minal Moharir, and Prof. Anitha" Sandeep titled the journal 'Phishing Website URL Detection using Machine Learning' published in the International Journal of Advanced Science and Technology in 2020

[5]. "R. Kiruthiga and D. Akila" discusses the detection of phishing websites using machine learning techniques this study was published in the September 2019 issue of the International Journal of Recent Technology and Engineering (IJRTE), Volume 8, Issue 2011
"YounessMourtaji", Hybrid Rule-Based Solution for Phishing URL Detection Using Convolutional Neural Network,2021.

[6]. "SK Hasane Ahammad ", Phishing URL detection using machine learning methods,2022.

[7]. "Jianting Yuan", Malicious URL Detection Based on a Parallel Neural Joint Model,2020.

[8]. "Pingfan Xu", A Transformer-based Model to Detect Phishing URLs,2021.

[9]. "Andrei Butnaru ", Towards Lightweight URL-Based Phishing Detection,2021.