IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 2, May 2024

Detection of Phishing Website using XG – Boost Algorithm

Sridevi Malipatil[1], A S Kruthik[2], Eldad Nischal[3], G S Vinay Kumar[4], Ajay Kumar H A[5]
Assistant Professor, Computer Science and Engineering[1]
Students, Department of Computer Science and Engineering[2,3,4,5]
Rao Bahadur Y Mahabaleswarappa Engineering College, Ballari, India

Abstract: *Phishing, a cybercriminal's attempted attack, is a social web-engineering attack in which valuable data or personal information might be stolen from either email addresses or websites. There are many methods available to detect phishing, but new ones are being introduced in an attempt to increase detection accuracy and decrease phishing websites success to steal information. Phishing is generally detected using Machine Learning methods with different kinds of algorithms. In this study, our aim is to use Machine Learning to detect phishing websites. We used the data from Kaggle consisting of 86 features and 11,430 total URLs, half of them are phishing and half of them are legitimate. We trained our data using Decision Tree (DT), Random Forest (RF), XGBoost, Multilayer Perceptrons, K-Nearest Neighbors, Naive Bayes, AdaBoost, and Gradient Boosting and reached the highest accuracy of 96.6using X G Boost.*

Keywords: Machine learning algorithm, HTML, datasets, Decision tree, Random forest, XG Boost, Multilayer perceptrons, K-nearest neighbors, Naïve bayes, AdaBoost, Gradient boosting.

I. INTRODUCTION

The internet is now a necessary component of our everyday lives due to the continuously advancing technologies. The majority of our daily activities rely on using the internet. The number of social networking sites has skyrocketed in the past few years. Phishing is one of the frequent and harmful hazards that users of the internet face as a result of their frequent use. Phishing is the act of pretending to be a legitimate website in order to deceive people by obtaining their personal information, such as passwords, account numbers, national insurance numbers, and usernames. Phishing scams may be the most prevalent type of cybercrime in use today. Phishing attacks can occur in a wide range of industries, including online payment systems, cloud storage networks, webmail, finance, file hosting, and many more.More phishing attacks have targeted the webmail and online payment industries than any other. Users should be aware of the risks associated with spear phishing and email phishing schemes, which are two methods of phishing.

II. LITERATURE SURVEY

1. In the paper "Phishing Detection System Through Hybrid Machine Learning Based on URL",it covers a comprehensive range of topics related to phishing detection systems. The study utilizes a dataset from Kaggle and applies various machine learning models such as decision tree, linear regression, random forest, support vector machine, gradient boosting machine, K-Neighbor classifier, naive Bayes, and a hybrid model (LR+SVC+DT) with soft and hard voting to achieve high-performance results. Additionally, the research incorporates canopy feature selection, cross-fold validation, and Grid search hyperparameter optimization techniques with the LSD Ensemble model to enhance the phishing detection system's efficiency. The study also references previous works on classification techniques in data mining, machine learning solutions in sewer systems, pruning of random forest classifiers, and the concept of random forests. Furthermore, the paper acknowledges the contributions of researchers from various institutions and provides insights into the future of phishing detection systems, suggesting the integration of list-based and machine learning-based approaches for more efficient detection of phishing URLs.

2. The paper "Knowledge-Based Approach to Detect Potentially Risky Websites" presents a novel Knowledge-Based System (KBS) called DOCRIW for automating the detection of potentially risky websites. The KBS is built using information collected from websites that present illegal and malicious content, and includes a module to predict the risk

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 2, May 2024

of websites not found in this knowledge database using a binary classifier trained with supervised learning. The paper also discusses related work on KBSs, malware and fraud detection solutions, and machine learning methods applied to malware and fraud detection. It highlights the use of the domain name as the primary input for determining the risk level of a website, and the application of similarity measures and assembling methods for optimal classification.

3. The literature survey on the paper “Phishing URL detection: A real case scenario through login URLs” encompasses various approaches and methodologies employed by researchers in the field. Researchers have primarily focused on phishing detection using list-based methods, automatic detection techniques utilizing machine learning, and deep learning algorithms. Studies such as those by Sahingoz et al, Sadique et al, Jain and Gupta, and Banik and Sarma have explored different feature extraction methods, including handcrafted features, lexical features, WHOIS features, and GeoIP-based features, to enhance the accuracy of phishing detection models. Additionally, the use of machine learning classifiers like Random Forest, XGBoost, and LightGBM has shown promising results in accurately distinguishing between legitimate and phishing URLs. The incorporation of login page URLs in the dataset for training and testing models, as discussed in the paper, highlights the importance of considering real-world scenarios to improve the robustness and effectiveness of phishing detection systems.

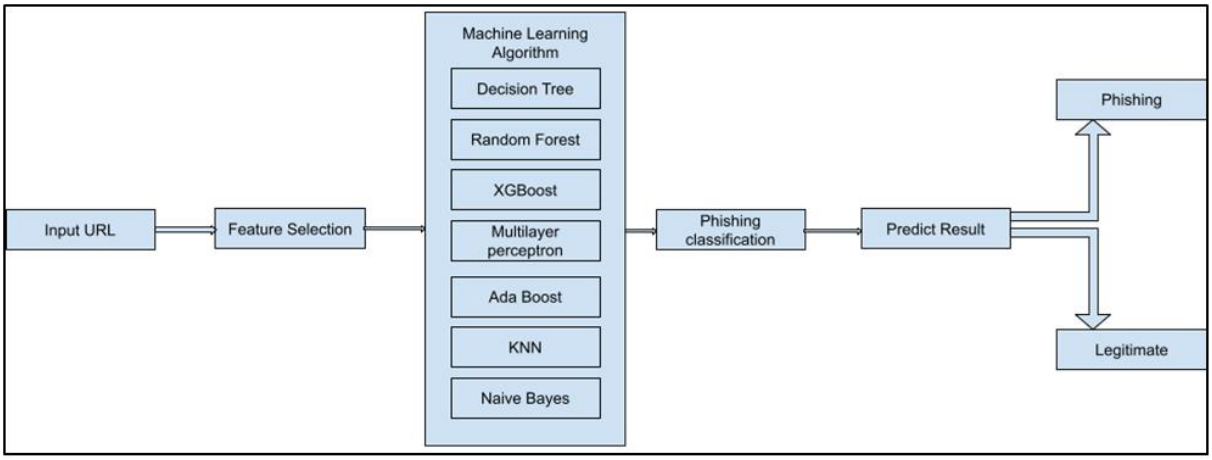
4. This paper, titled "Uncovering the Cloak: A Systematic Review of Techniques Used to Conceal Phishing Websites," provides a comprehensive systematic review of primary studies conducted between 2012 and 2022 on using cloaking techniques to evade detection by anti-phishing entities. The review focuses on different server-side and client-side detection strategies, phishing techniques and cloaking mechanisms, toolkits, blacklists, phishing or anti-phishing ecosystems, and other such concepts. The study identifies the research characteristics of present studies and extracts the most important thematic findings to understand the state-of-art topics in this domain. The limitations of the study are also outlined. This systematic literature review (SLR) is one of the first reviews to be conducted for analyzing the current cloaking or evasion techniques used by phishers.

III. PROPOSED SYSTEM

The purpose of phishing website URLs is to obtain personal information such as passwords, user names, and online banking activity. Phishers use websites that are grammatically and aesthetically similar to those authentic ones. The rapid progress of phishing strategies due to technological advancements must be stopped by employing anti-phishing tools to detect phishing. Machine learning is a powerful tool for preventing phishing attacks. Because it is easier to trick a victim into opening a malicious link that appears to be real than it is to try to circumvent a computer's security mechanisms, attackers commonly utilize phishing. The malicious links in the message body are designed to look as though they lead to the fake company by using its logos and other authentic details.The technique being described uses machine learning to develop a novel way of identifying phishing websites. In order to predict phishing websites, we will build a model in this research using machine learning algorithms. The dataset, which includes the specifics of websites that have been subjected to phishing attacks, is available for download as a.csv file from www.phishtank.com. The algorithms Random Forest and Decision Tree, which are widely used in current systems to anticipate phishing websites, need to have their accuracy, precision, and recall improved. The final user can see a graphical representation of the accuracy, precision, and recall.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 2, May 2024

IV. ARCHITECTURE OF WORKING MODEL



V. METHODOLOGY

- **Problem statement :**Phishing is a form of social engineering and scam where attackers deceive people into revealing sensitive information or installing malware such as ransomware. Phishing attacks have become increasingly sophisticated and often transparently mirror the site being targeted, allowing the attacker to observe everything while the victim is navigating the site, and transverse any additional security boundaries with the victim.
- **Data Collection :**we Collected a dataset from Kaggle website consisting of 86 features and 11,430 total URLs, half of them are phishing and half of them are legitimate.URL features such as URL length, having at ,HTTPS Domain , Tinyurl, Domain name, iframe, label etc .
- **Data Preprocessing:** We Preprocess the data by cleaning and normalizing the features. This might include removing irrelevant features ,handling missing values, and encoding categorical variables. We will split the data into training and testing set .ensuring that the two sets are representative of the overall population.
- **Model selection :** we considered several machine learning algorithms for phishing classification such as Decision Tree, Random Fores , XGBoost ,KNN ,Ada Boost , Naïve Bayes ,Multi layer perceptron. We will evaluate the performance of each model using metrics such as accuracy , we select XGboost model that performs best on the testing set and secured 91% of accuracy .
- **Model Training:** We will train the XGBoost model using the training set. We will tune the hyperparameters of the model to optimize its performance. We will also ensure that the model is not overfitting or underfitting the data.
- **Model deployment :**We Created a Web interface that allows user to interact user input URLs and Receive a Prediction of Whether the URL is Phishing or Legitimate.
- **Limitations and Future work:** we Acknowledge limitations of model such as its accuracy on certain types of URLs. And also future work such as improving model performance on specific types of phishing attacks.

VI. CONCLUSION

Anti-Phishing Extension has been proposed to handle the phishing contents. The proposed approach consists of three algorithms: 'Phishing detection, checking URL for IP address, and checking redirection of the user's information. This paper focuses on phishing's main consequences, such as stealing personal information from bank accounts, credit cards, social media, etc. as determined by protecting users from phishing attacks to deal with the human factor. The proposed Anti Phishing Extension approach helps detect phishing attacks efficiently and accurately.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 2, May 2024

VII. FUTURE RESEARCH DIRECTION

Future research directions could focus on enhancing the efficiency and scalability of phishing detection systems. One potential direction is to explore the integration of advanced machine learning techniques, such as deep learning algorithms, to further improve the accuracy of identifying sophisticated phishing websites. Additionally, investigating the utilization of real-time data streams and dynamic feature extraction methods could enhance the adaptability of detection systems to evolving phishing tactics. Furthermore, exploring the incorporation of user behavior analysis and anomaly detection techniques may provide valuable insights into detecting previously unseen phishing attacks. Collaborative research efforts that combine expertise in cybersecurity, machine learning, and data analytics could lead to the development of more robust and proactive solutions for combating phishing threats in the digital landscape.

REFERENCES

[1]. Mohammed Hazim Alkawaz, Stephanie Joanne Steven and Asif Iqbal Hajamydeen, "Detecting Phishing Website Using Machine Learning", 16th IEEE International Colloquium on Signal Processing its Applications (CSPA 2020), 28-29 Feb. 2020.

[2]. Sandeep Kumar Satapathy, Shruti Mishra, Pradeep Kumar Mallick, Lavanya Badiginchala, Ravali Reddy Gudur and Siri Chandana Guttha, "Classification of Features for detecting Phishing Web Sites based on Machine Learning Techniques", International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 8, no. 8S2, June 2019, ISSN 2278-3075

[3]. Vaibhav Patil and S. P. Godse, Detection and Prevention of Phishing Websites using Machine Learning Approach, ISBN 978-1-5386-5257-2018.

[4]. S Nandhini and V. Vasanthi, Extraction of Features and Classification on Phishing Websites using Web Mining Techniques, vol. 5, no. 4, 2017, ISSN 2321-9939.

[5]. Sagar Patil, Yogesh Shetye and Nilesh Shendage, "Detecting Phishing Websites", International Research Journal of Engineering and Technology, vol. 07, no. 02, Feb 2020.

[6]. Mahajan Mayuri Vilas, Kakade Prachi Ghansham, Sawant Purva Jaypralash and Pawar Shila, "Detection of Phishing Website Using Machine Learning Approach", International Conference on Electrical Electronics Communication Computer Technologies and Optimization Techniques(ICEECCOT), 2019.

[7]. Ankit Kumar Jain and B BGuptha, "A Machine Learning based approach for phishing detection using Hyperlinks information", Part of Springer Nature, 2018.

[8]. Basnet, R., Mukkamala, S., Sung, A.H. "Detection of phishing attacks: A machine learning approach." Studies in Fuzziness and Soft Computing, 226:373–383, 2014.