URL BASED PHISHING DETECTION

1st Mr. Gautam Sharma
Dept. of Computer Science(RMDSSOE)
Savitribai Phule Pune University
Pune, India
sharmagautam09092001@gmail.com

2nd Mrs. Pradnya Kasture
Dept. of Computer Science(RMDSSOE)
Savitribai Phule Pune University
Pune, India
PradnyaKasture.rmdssoe@sinhgad .edu

3rd Mr. Sanket Shendge
Dept. of Computer Science(RMDSSOE)
Savitribai Phule Pune University
Pune, India
sanketshendge.rmdssoe.comp@gmail .com

4th Mr. Yadnyesh Chaudhari
Dept. of Computer Science(RMDSSOE)
Savitribai Phule Pune University
Pune, India
yadnyeshchaudhari875@gmail.com

Abstract –

Phishing attacks pose a significant threat to online security, targeting individuals and organizations with deceptive emails and websites designed to steal sensitive information. Traditional methods of detecting phishing websites have become inadequate against the evolving tactics employed by cybercriminals. This study explores the application of machine learning techniques in phishing website detection, aiming to enhance accuracy and real-time response capabilities. The results highlight the significance of this research in strengthening online security. By leveraging machine learning techniques, the proposed system provides a proactive defense against phishing attacks, safeguarding users, businesses, and organizations from financial losses, identity theft, and reputational damage. Furthermore, the study underscores the importance of continuous research and collaboration in the ever-changing landscape of cybersecurity, ensuring a safer digital environment for all.

Keywords: ***Feature Extraction, SVM, Classification, Model-training, Random Forest, Naive Bayes, Decision Tree***

1.INTRODUCTION

Phishing websites pose a significant threat to online users, aiming to deceive them into revealing sensitive information such as usernames, passwords, and financial details. Detecting these malicious websites is crucial for ensuring online security. In this context, the use of advanced techniques such as Feature Extraction, Support Vector Machines (SVM), Random Forest (RF), and Naive Bayes (NB) can significantly enhance the effectiveness of phishing website detection. Phishing is a cyber-attack method where attackers create deceptive websites that mimic legitimate ones, tricking users into divulging confidential information. Phishing attacks often rely on social engineering techniques to exploit human psychology. Detecting phishing websites is challenging due to their dynamic and evolving nature. Therefore, employing sophisticated techniques becomes imperative to stay ahead of cyber threats.

Phishing Website Detection Using Feature Extraction:

Feature extraction involves selecting and transforming relevant information from the raw data to create a feature set that can be used for analysis. In the context of phishing website detection, extracting relevant features from URLs, HTML, JavaScript, and content is crucial.

Support Vector Machines (SVM):

SVM is a powerful classification algorithm that works well in high-dimensional spaces. It separates data into different classes by finding the hyperplane that maximizes the margin between classes. SVM can effectively classify phishing and legitimate websites based on extracted features.

Random Forest (RF):

RF is an ensemble learning algorithm that builds multiple decision trees and merges their predictions. It is robust, handles high-dimensional data well, and is less prone to overfitting. Random Forest can provide a reliable classification model for phishing website detection.

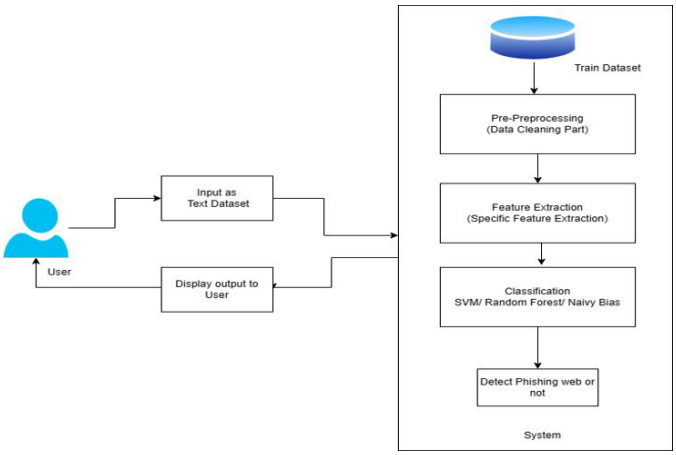
Naive Bayes (NB):

NB is a probabilistic algorithm based on Bayes' theorem. Despite its simplicity, NB can perform well in certain scenarios, especially with a large number of features. It is computationally efficient and can handle real-time classification tasks.

Decision Tree (DT):

A Decision Tree is a versatile machine learning algorithm used for both classification and regression tasks. It models decisions and their possible consequences as a tree-like structure of nodes and branches. Using a Decision Tree for phishing detection based on URLs involves analyzing various features of URLs to distinguish between legitimate and phishing websites.

Here is illustrates diagram that show the process of a phishing detection system that leverages machine learning techniques. Here's a step-by-step explanation of each component and their interactions:



1. **User Interaction**:

Input as Text Dataset: The user provides input, which could be URLs, emails, or text data that needs to be analyzed for phishing.

Display Output to User: After processing, the result is displayed back to the user, indicating whether the input is classified as phishing or not.

2. **System Workflow**:

Train Dataset: This is the initial dataset used to train the machine learning models. It contains labeled examples of phishing and legitimate data.

Pre-Preprocessing (Data Cleaning Part): This step involves cleaning the input data to remove noise and irrelevant information. Common tasks include removing duplicates, handling missing values, and normalizing text.

Feature Extraction (Specific Feature Extraction): Relevant features are extracted from the cleaned data. This might include characteristics such as URL length, presence of certain keywords, domain age, etc., which are indicative of phishing attempts.

Classification: The extracted features are fed into machine learning classifiers. The diagram mentions three possible algorithms:

SVM (Support Vector Machine): A supervised learning model used for classification tasks.

Random Forest: An ensemble learning method that uses multiple decision trees to improve predictive performance.

Naive Bayes: A probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions between features.

Decision Tree (DT): A Decision Tree is a versatile machine learning algorithm used for both classification and regression tasks.

Detect Phishing Web or Not: The classifier's output determines whether the input data is classified as phishing or legitimate.

This system combines data cleaning, feature extraction, and machine learning classification to automatically detect and identify phishing attempts, providing real-time feedback to the user.

2. RELATED WORK

In this section, we review prior research relevant to our proposed system. We categorize related work into several key areas that have direct bearing on our research. 1."TrustQR: A New Technique for the Detection of Phishing Attacks on QR Code" Graphic black and white squares, known as Quick Response (QR) code is a matrix barcode, which allows easy interaction between mobile and websites or printed material by doing away with the necessity of manually typing a URL or contact information. From the pages of magazines to the sides of buses and billboards, QR code technology is being used increasingly in smartphones. Unfortunately, Phishers have started using QR code for phishing attacks by using some features of QR code. This paper introduces a new approach called "TrustQR" which detects URL phishing on QR code. It uses QR code specific features and URL features to detect if the QR code content has a phishing URL. Some of the QR code specific features use QR code

content and its characteristics like length, type, and level of error correction to generate the cryptography key. This technique uses the machine learning classification technique.

2.“Setting priorities in behavioral interventions: An application to reducing phishing risk,”

Phishing risk is a growing area of concern for corporations, governments, and individuals. Given the evidence that users vary widely in their vulnerability to phishing attacks, we demonstrate an approach for assessing the benefits and costs of interventions that target the most vulnerable users. Our approach uses Monte Carlo simulation to (1) identify which users were most vulnerable, in signal detection theory terms; (2) assess the proportion of system-level risk attributable to the most vulnerable users; (3) estimate the monetary benefit and cost of behavioral interventions targeting different vulnerability levels; and (4) evaluate the sensitivity of these results to whether the attacks involve random or spear phishing. Using parameter estimates from previous research, we find that the most vulnerable users were less cautious and less able to distinguish between phishing and legitimate emails (positive response bias and low sensitivity, in signal detection theory terms). They also accounted for a large share of phishing risk for both random and spear phishing attacks. Under these conditions, our analysis estimates much greater net benefit for behavioral interventions that target these vulnerable users. Within the range of the model's assumptions, there was generally net benefit even for the least vulnerable users. However, the differences in the return on investment for interventions with users with different degrees of vulnerability indicate the importance of measuring that performance, and letting it guide interventions. This study suggests that interventions to reduce response bias, rather than to increase sensitivity, have greater net benefit.

3."Intelligent phishing detection system for e-banking using fuzzy data mining"

Detecting and identifying any phishing websites in real-time, particularly for e-banking, is really a complex and dynamic problem involving many factors and criteria. Because of the subjective considerations and the ambiguities involved in the detection, fuzzy data mining techniques can be an effective tool in assessing and identifying phishing websites for e-banking since it offers a more natural way of dealing with quality factors rather than exact values. In this paper, we present novel approach to overcome the ‘fuzziness’ in the e-banking phishing website assessment and propose an intelligent resilient and effective model for detecting e-banking phishing websites. The proposed model is based on fuzzy logic combined with data mining algorithms to characterize the e-banking phishing website factors and to investigate its techniques by classifying the phishing types and defining six e-banking phishing website attack criteria’s with a layer structure. Our experimental results showed the significance and importance of the e-banking phishing website criteria (URL & Domain Identity) represented by layer one and the various influence of the phishing characteristic on the final e-banking phishing website rate

4."A layout-similarity-based approach for detecting phishing pages"

Phishing is a current social engineering attack that results in online identity theft. In a phishing attack, the attacker persuades the victim to reveal confidential information by using web site spoofing techniques. Typically, the captured information is then used to make an illegal economic profit by purchasing goods or undertaking online banking transactions. Although simple in nature, because of their effectiveness, phishing attacks still remain a great source of concern for organizations with online customer services. In previous work, we have developed AntiPhish, a phishing protection system that prevents sensitive user information from being entered on phishing sites. The drawback is that this system requires cooperation from the user and occasionally raises false alarms. In this paper, we present an extension of our system (called DOMAntiPhish) that mitigates the shortcomings of our previous system. In particular, our novel approach leverages layout similarity information to distinguish between malicious and benign web pages. This makes it possible to reduce the involvement of the user and significantly reduces the false alarm rate. Our experimental evaluation demonstrates that our solution is feasible in practice.

5."Improving Email Security with Fuzzy Rules"

Phishing and other malicious email messages are increasingly serious security threats. An important tool for countering such email threats is the automated or semiautomated detection of malicious email. This paper reports work on using fuzzy rules to classify email for such purposes. The effectiveness of a fuzzy rule-based classifier is studied experimentally on a real dataset and compared with results for other classifiers, including those based on crisp rules and decision trees. The human-readability and editability of the classifiers produced by these methods is also studied.

6.”PhishAri: Automatic realtime phishing detection on twitter,”

With the advent of online social media, phishers have started using social networks like Twitter, Facebook, and

Foursquare to spread phishing scams. Twitter is an immensely popular micro-blogging network where people post short messages of 140 characters called tweets. It has over 100 million active users who post about 200 million tweets everyday. Phishers have started using Twitter as a medium to spread phishing because of this vast information dissemination. Further, it is difficult to detect phishing on Twitter unlike emails because of the quick spread of phishing links in the network, short size of the content, and use of URL obfuscation to shorten the URL. Our technique, PhishAri, detects phishing on Twitter in realtime. We use Twitter specific features along with URL features to detect whether a tweet posted with a URL is phishing or not. Some of the Twitter specific features we use are tweet content and its characteristics like length, hashtags, and mentions. Other Twitter features used are the characteristics of the Twitter user posting the tweet such as age of the account, number of tweets, and the follower-followee ratio. These Twitter specific features coupled with URL based features prove to be a strong mechanism to detect phishing tweets. We use machine learning classification techniques and detect phishing tweets with an accuracy of 92.52%. We have deployed our system for end-users by providing an easy to use Chrome browser extension which works in realtime and classifies a tweet as phishing or safe. We show that we are able to detect phishing tweets at zero hour with high accuracy which is much faster than public blacklists and as well as Twitter's own defense mechanism to detect malicious content. To the best of our knowledge, this is the first realtime, comprehensive and usable system to detect phishing on Twitter

7."Phishing websites detection through supervised learning networks" Phishing is an unlawful activity of making gullible people to reveal their insightful information into fake websites. The Aim of these phishing websites is to acquire confidential information such as usernames, passwords, banking credentials and some other personal information. Phishing website looks similar to legitimate website therefore people cannot make difference among them. Today users are heavily relying on the internet for online purchasing, ticket booking, bill payments, etc. As technology advances, the phishing approaches being used are also getting progressed and hence it stimulates anti-phishing methods to be upgraded. In this paper, we have implemented two algorithms named Adaline and Backpropion along with the support vector machine to enhance the detection rate and classification.

8.“Design and Evaluation of a Real-Time URL Spam Filtering Service”

On the heels of the widespread adoption of web services such as social networks and URL shorteners, scams, phishing, and malware have become regular threats. Despite extensive research, email-based spam filtering techniques generally fall short for protecting other web services. To better address this need, we present Monarch, a real-time system that crawls URLs as they are submitted to web services and determines whether the URLs direct to spam. We evaluate the viability of Monarch and the fundamental challenges that arise due to the diversity of web service spam. We show that Monarch can provide accurate, real-time protection, but that the underlying characteristics of spam do not generalize across web services. In particular, we find that spam targeting email qualitatively differs in significant ways from spam campaigns targeting Twitter. We explore the distinctions between email and Twitter spam, including the abuse of public web hosting and redirector services. Finally, we demonstrate Monarch's scalability, showing our system could protect a service such as Twitter -- which needs to process 15 million URLs/day -- for a bit under $800/day.

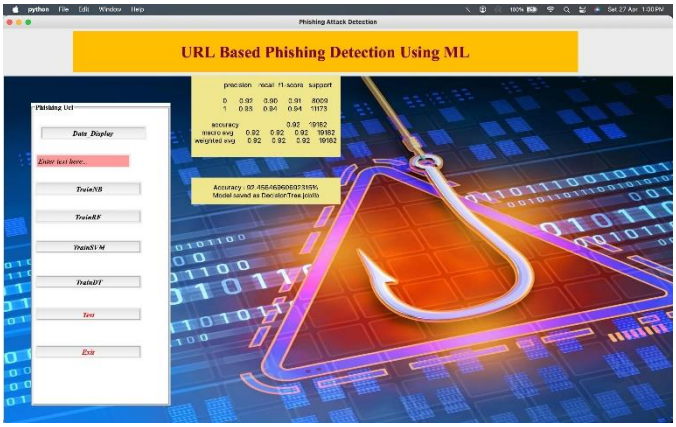
9."Trust based communication in unstructured P2P networks using reputation management and self-certification mechanism"

In unstructured P2P networks there is a possibility of malicious codes and false transactions. It generates the false identities in order to perform false transactions with other identities. The proposed method uses the concept of DHT and reputation management which provides efficient file searching. The self certification (RSA and MD5) is used for ensuring secure and timely availability of the reputation data of a peer to other peers. The reputations of the peers are used to determine whether a peer is a malicious peer or a good peer. Once the malicious peer is detected the transaction is aborted. The reputation of a given peer is attached to its identity. The identity certificates are generated using self-certification and all peers maintain their own (and hence trusted) certificate authority which issues the identity certificate(s) and digital signature to the peer.

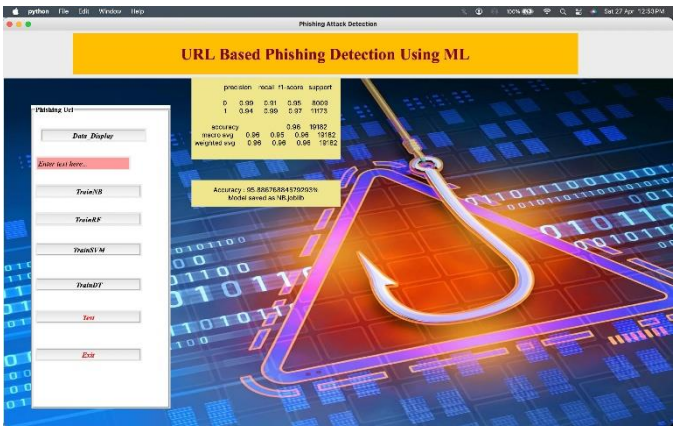
10."Information Source-Based Classification of Automatic Phishing Website Detectors" Phishing attacks allure users to submit their personal information to fake websites that mimic legitimate websites. Many anti-phishing techniques have emerged in recent years. However, the numbers of phishing attacks are still increasing. Two reasons can be blamed for this situation. First, users have too much trust and confidence on existing anti-phishing tools in general. Second, most users believe that they are foolproof against phishing attacks when anti-phishing tools are deployed. We believe that understanding of anti-phishing tools based on their common features can be the beginning step to address these issues. However, there is no extensive analysis of existing anti-phishing techniques. This paper

attempts to classify existing works based on information sources. The classification would not only provide useful information to develop new anti-phishing techniques or improve existing techniques, but also enable our understanding on the limitations of the existing techniques

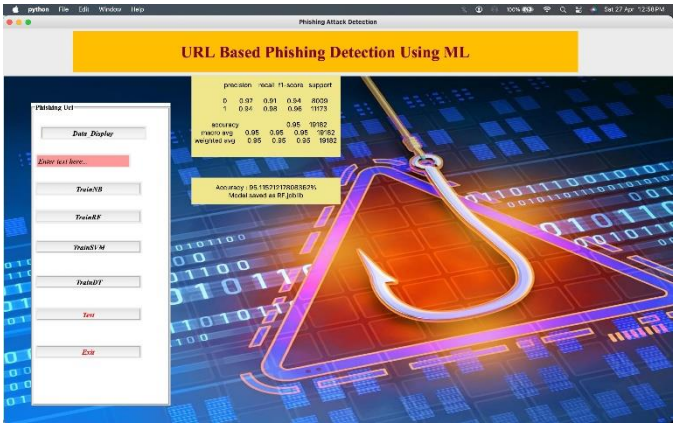
3. RESULTS



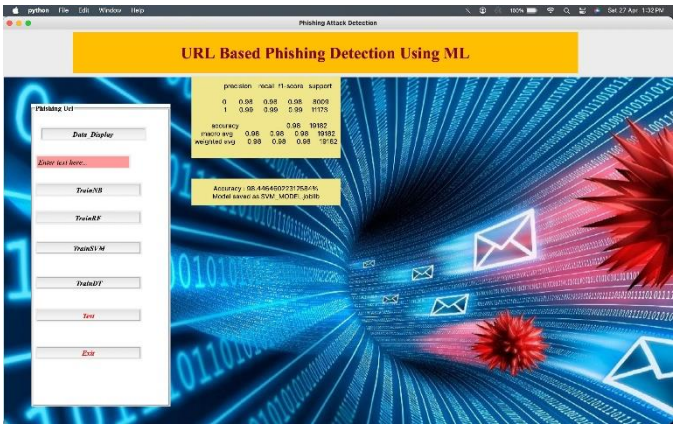
Accuracy using Decision Tree



Accuracy using NB



Accuracy using RF



Accuracy using SVM

4. CONCLUSIONS

Combining advanced feature extraction techniques with machine learning algorithms like SVM, RF, DT and NB provides a potent solution for detecting phishing websites. This multi-faceted approach allows for a comprehensive analysis of various aspects of phishing attacks, enabling more accurate and robust detection in the ever-evolving landscape of cyber threats. As cybercriminals continue to refine their tactics, ongoing research and adaptation of detection techniques are essential to stay one step ahead in the ongoing battle for online security.

ACKNOWLEDGEMENT

We are grateful to our project guides, Mrs. Vina Lomte, HOD, RMDSSOE, and Mrs. Pradnya Kasture, Assistant Professor, RMDSSOE, for their unwavering support, tolerance, and inspiration as well as for their insightful advice and insightful throughout the Research process.

REFERENCES

[1]. A.Y. Ahmad, M. Selvakumar, A. Mohammed, and A.-S. Samer, "TrustQR: A new technique for the detection of phishing attacks on QR code," Adv. Sci. Lett., vol. 22, no. 10, pp. 2905-2909, Oct.2021.

[2].C.C.Inez and F.Baruch,"Setting priorities in behavioral interventions: Anapplication to reducing phishing risk," Risk Anal., vol. 38, no. 4, pp. 826-838, Apr. 2021.

[3]. Aburrous, Maher Hossain, Mohammed Dahal, Keshav Thabtah, Fadi. (2020). Intelligent phishing detection system for ebanking using fuzzy datamining. Expert Systems with Applications. 37. 7913-7921. 10.1016/ j.eswa.2020.04.044.

[4].Rosiello, Angelo Kirda, Engin Kruegel, Ferrandi, Fabrizio. (2007). A layout-similarity-based approach for detecting phishing pages. Proceedings of the 3rd International Conference on Security and Privacy in Communication

[5].Chawathe, Sudarshan. Improving Email Security with Fuzzy Rules. 1864-1869. 10.1109/TrustCom/ BigDataSE.2018.00282. 2021

[6]. A. Aggarwal, A. Rajadesingan and P. Kumaraguru, "PhishAri: Automatic realtime phishing detection on twitter," eCrime Researchers Summit, Las Croabas, 2012, pp. 1-12, doi: 10.1109/eCrime.6489521 2022.

[7].P. Singh, Y. P. S. Maravi and S. Sharma, "Phishing Websites Detection through Supervised Learning Networks", 2020 International Conference on Computing and Communications Technologies (ICCCT), Chennai, 2020, pp. 61-65.

[8].K. Thomas, C. Grier, J. Ma, V. Paxson and D. Song, "Design and Evaluation of a Real-Time URL Spam Filtering Service", IEEE Symposium on Security and Privacy, Berkeley, CA, , pp. 447-462. 2021

[9].C.V. Arulkumar et al., "Secure Communication in Unstructured P2P Networks based on Reputation Management and Self certification", International Journal of Computer Applications, vol. 15, pp. 1-3,

[10]. H. Shahriar and M. Zulkernine, "Information Source-based Classification of Automatic PhishingWebsite Detectors", 2022 IEEE/IPSJ International Symposium on Applications and the Internet, Munich, Bavaria, pp. 190-195. 2022