Monitoring and Management of Phishing and Malicious Websites Using Machine Learning

Miss. Tejal S. Gite[1], Prof. Sunil H. Sangale[2], Miss. Sanjana K. Jagtap[3], Miss. Sakshi Y. Amrutkar[4], Miss. Komal G. Godse[5]

[1,3,4,5]Students, Computer Technology, K K Wagh Polytechnic, Nashik, India
[2]Senior Lecturer, Computer Technology, K K Wagh Polytechnic, Nashik, India

Abstract

Phishing internet sites are one of the internet protections issues that focus on human vulnerabilities rather than software program vulnerabilities. It can be defined because of the process of attracting online users to gain their touchy facts which include usernames and passwords.

In this paper, we provide a sensible machine for detecting phishing websites. The gadget acts as a further functionality to an internet browser as an extension that routinely notifies the consumer whilst it detects a phishing internet site. The system is based on a device gaining knowledge of approach, particularly supervised mastering. We have decided on the XGBoost method because of its true overall performance in classification. Our focus is to pursue a better overall performance classifier by analyzing the features of phishing websites and choose the better aggregate of them to train the classifier. As a result, we finished our paper with accuracy of 99% accuracy with 48 features.

Keywords: XGBoost (Extreme Gradient Boosting), Classifier, Features, Phishing, Train, Accuracy.

Introduction

Internet and cloud technology improvements in recent years have significantly increased electronic trade, or consumer to-consumer online transactions. The resources of an enterprise are harmed by this growth, which permits unauthorized access to sensitive information about users. One well-known attack that manipulates users into accessing dangerous content and giving up their information is phishing. Most phishing websites use the same website interface and universal resource location (URL) as the legitimate websites. In recent years, with the increasing use of mobile devices, there is a growing trend to move almost all real-world operations to the cyber world. Although this makes easy our daily lives, it also brings many security breaches due to the anonymous structure of the Internet. Used antivirus programs and firewall systems can prevent most of the attacks. However, experienced attackers target on the weakness of the computer users by trying to phish them with bogus webpages.

These pages imitate some popular banking, social media, e-commerce, etc. sites to steal some sensitive information such as, user-ids, passwords, bank account, credit card numbers, etc. Phishing detection is a challenging problem, and many different solutions are proposed in the market as a blacklist, rule-based detection, anomaly-based detection, etc. In the literature, it is seen that current works tend on the use of machine learning-based anomaly detection due to its dynamic structure, especially for catching the “zeroday" attacks. In this paper, we proposed a machine learning-based phishing detection system to analyze the URLs. Phishing is the most commonly used social engineering and cyber attack. Through such attacks, the phisher targets naïve online users by tricking them into revealing confidential information, with the purpose of using it fraudulently. In order to avoid getting phished, users should have awareness of phishing websites. Have a blacklist of phishing websites which requires the knowledge of website being detected as phishing. Detect them in their early appearance, using machine learning and deep neural network algorithms of the above three, the machine learning based method is proven to be most effective than the other methods. Even then, online users are still being trapped into revealing sensitive information in phishing websites. A phishing website is a common social engineering method that mimics trustful uniform resource locators (URLs) and webpages.

The objective of this project is to train machine learning models and deep neural nets on the dataset created to predict phishing websites. Both phishing and benign URLs of websites are gathered to form a dataset and from them required URL and website content- based features are extracted. The performance level of each model is measures and compared. The phishing website has evolved as a major cybersecurity threat in recent times. The phishing websites host spam, malware, ransomware, drive-by exploits, etc. A phishing website many a time lookalike a very popular website and lure an unsuspecting user to fall victim to the trap. The victim of the scams incurs a monetary loss, loss of private information and loss of reputation. Hence, it is imperative to find a solution that could mitigate such security threats in a timely manner. Traditionally, the detection of phishing websites is done using blacklists. There are many popular websites which host a list of blacklisted websites, e. g. Phis Tank. The blacklisting technique lack in two aspects, blacklists might not be exhaustive and do not detect a newly generated phishing website. In recent times machine learning techniques have been used in the classification and detection of phishing websites. In, this paper we have compared different machine learning techniques for the phishing website. In our daily life, we carry out most of our work on digital platforms. Using a computer and the internet in many areas facilitates our business and private life.

It allows us to complete our transaction and operations quickly in areas such as trade, health, education, communication, banking, aviation, research, engineering, entertainment, and public services. The users who need to access a local network have been able to easily connect to the Internet anywhere and anytime with the development of mobile and wireless technologies. Although this situation provides great convenience, it has revealed serious deficits in terms of information security. Thus, the need for users in cyberspace to take measures against possible cyber- attacks has emerged. Attacks can be carried out by people such as cyber criminals, pirates, or non- malicious (white-capped) attackers and hacktivists. The aim is to reach the computer or the information it contains or to capture personal information in different ways. The attacks, as internet worms (Morris Worm), started in 1988, and they have been carried out until today.

Problem Definition

The project " Monitoring and Management of Phishing and Malicious Websites using Machine Learning" involves developing techniques and systems to detect, classify, and mitigate the threat posed by these deceptive and harmful online entities. Phishing and malicious websites are created with the intent to deceive users, steal sensitive information, or distribute malware, posing significant risks to individuals, organizations, and the overall cybersecurity landscape.

Literature Survey

This chapter comprises of the literature review and theoretical background of the project. The literature review deals basically with related project written by other researchers, the difficulties they encountered, limitations and modifications that should be made.

In their paper "Detection of Phishing Website Using Machine Learning Approach" , the goal of the study is to carry out ELM employing 30 different primary components that are characterized using ML. To prevent being discovered, most phishing URLs use HTTPS. Website phishing can be identified in three different ways. The first method evaluates several URL components; the second method assesses a website's authority, determines if it has been introduced or not, and determines who is in charge of it; the third method verifies a website's veracity.[1]

In paper "Phishing Attacks Detection using Machine Learning and Deep Learning Models" In this study, the highest correlated features from two distinct datasets were chosen. These features combined content-based, URL and domainbased features. Then, a comparison of the performance of a number of ML models was carried out. The results also sought to pinpoint the top characteristics that aid the algorithm in spotting phishing websites. The Random Forest (RF) method produced the best classification results for both datasets.[2]

The user-received URLs will be entered to the machine learning model, which will then process the input and report the results, indicating whether the URLs are phishing or not. SVM, Neural Networks, Random Forest, Decision Tree, XG boost, and other machine learning algorithms can all be used to categorize these URLs. The suggested method uses the Random Forest and Decision Tree classifiers. With an accuracy of 87.0% and 82.4% for Random Forest and decision tree classifiers, respectively, the suggested technique successfully distinguished between Phishing and Legitimate URLs.[3]

Methodology

Internet consumers lose billions of dollars each year as a consequence of website phishing. Phishers prey on people's online security by stealing usernames, passwords, and financial account information. Due to the use of URL obfuscation to shorten the URL, link redirections, modifying links to make them appear trustworthy, and a long list of other techniques, detecting phishing websites is difficult. This made it necessary to convert from conventional programming methods to an approach based on machine learning. When new phishing strategies are launched, phishing detection solutions do suffer from low detection accuracy and high false alarm rates. Additionally, since registering new domains has gotten simpler, the most popular methodology, the blacklist-based method, is ineffective at responding to phishing assaults that are on the rise. No comprehensive blacklist can guarantee a flawlessly up-todate database.

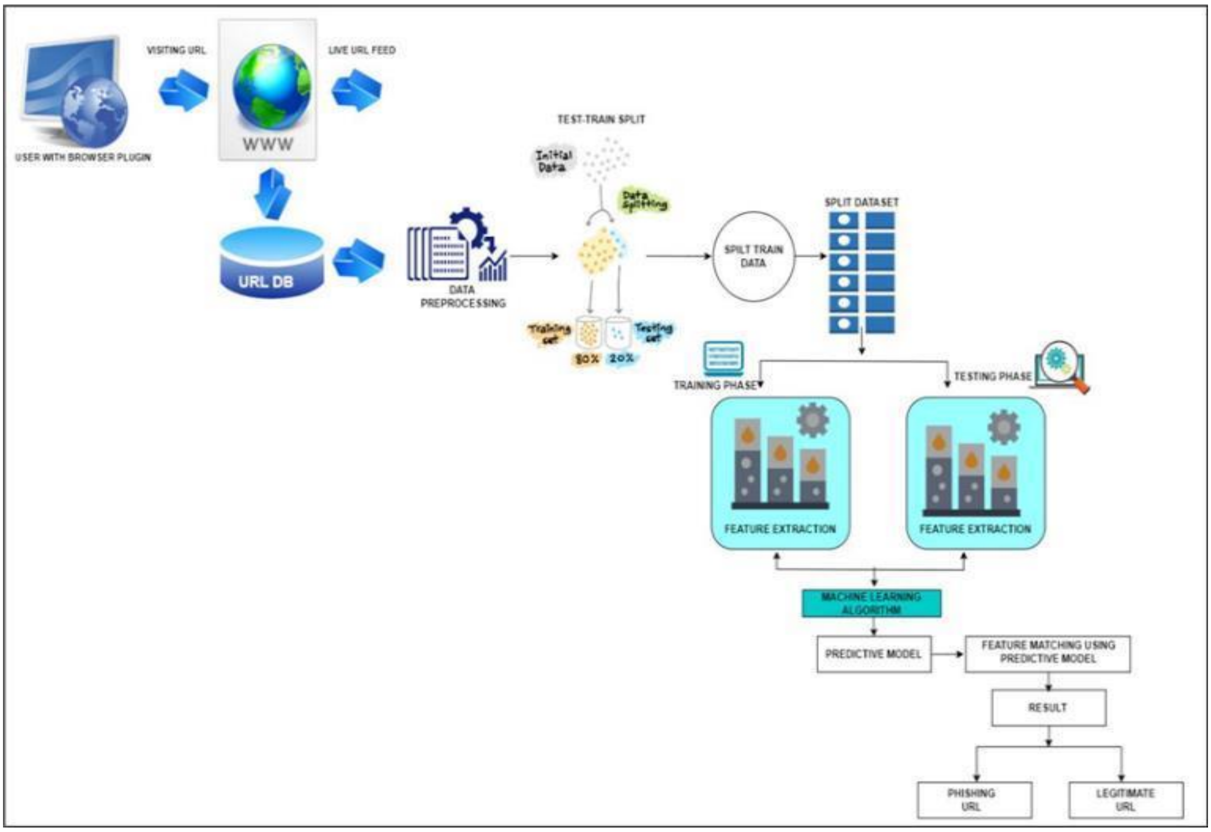


Figure 1 System Architecture

Future Scope

ML algorithms can be further developed to analyze various characteristics of URLs and websites, such as domain reputation, content analysis, linguistic patterns, SSL certificates, and HTML code. Advanced ML models can be trained to identify subtle and complex phishing techniques employed by attackers.It has the potential to enhance the overall security landscape, protect users from sophisticated phishing attacks, and minimize the financial and reputational damages caused by such threats.

Features:

Some of the features used in phishing URL and website detection using machine learning (ML) include:

1. **URL Components Analysis**: ML algorithms analyze different components of a URL such as domain name, subdomain, top-level domain, path, and parameters to identify suspicious patterns or irregularities**.**
2. **Domain Reputation:** ML models leverage historical data to evaluate the reputation of a domain. They assess factors like domain age, registration details, and previous malicious activities associated with the domain.
3. **SSL Certificate Analysis:** ML algorithms examine the validity and authenticity of the SSL certificate used by a website. They check for expiration dates, issuer information, and whether the certificate matches the domain.
4. **Page Content Analysis:** ML algorithms analyze the content of a webpage to detect potential phishing attempts. They look for suspicious keywords, HTML tags, pop-ups, redirects, and hidden forms that could indicate fraudulent behavior.

5. **Visual Similarity:** ML models use computer vision techniques to compare the visual similarity between a suspicious website and legitimate websites. They analyze the layout, logos, colors, and overall design to identify potential clones or replicas.
6. **URL Blacklisting:** ML algorithms compare the queried URL against known blacklists or databases of known phishing URLs. They check for matches or similarities to previously identified phishing attacks.
7. **7.Behavior Analysis:** ML models analyze the behavior of a website or URL, looking for anomalous patterns. They consider factors like the website's response time, hosting location, IP address, and user interactions to determine if it aligns with normal behavior or if it exhibits phishing characteristics.
8. **Machine Learning Classification:** ML models are trained on large datasets of legitimate and phishing URLs or websites. They learn patterns, features, and characteristics that distinguish between the two, allowing them to classify new unknown URLs or websites.

Applications:

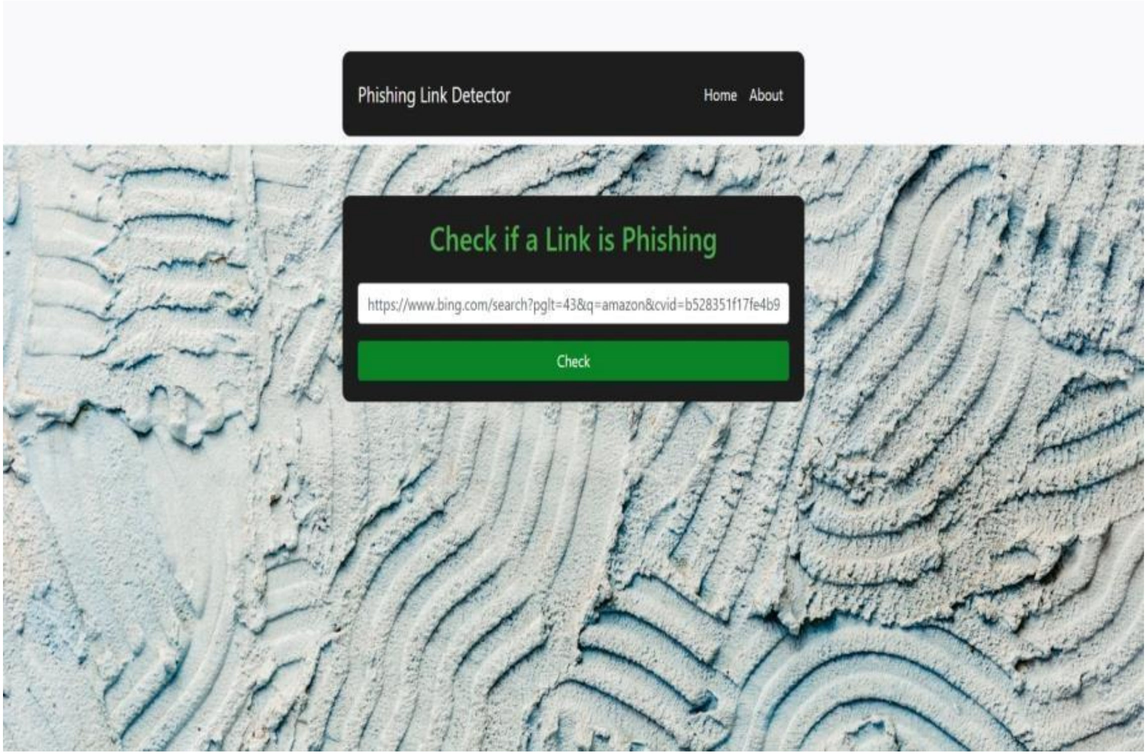
1. **Fraud Prevention:** Machine learning algorithms can be used to detect phishing attacks aimed at fraudulent activities such as identity theft, fraudulent transactions, and data breaches. This comprehensive platform combines machine learning with threat intelligence feeds to monitor and manage phishing and malicious websites. It continuously scans the web, identifies fraudulent sites, and gathers threat data. The platform enables organizations to correlate these findings with other security incidents, improving their overall fraud prevention efforts.
2. **Email Security:** Machine learning algorithms can be integrated into email security systems to detect phishing attempts, spam messages, and other harmful emails. Implement robust email filtering solutions to detect and block phishing emails. This application integrates with email security solutions to enhance anti-phishing capabilities. Using ML it scans inbound emails for phishing links and malicious content. When a phishing attempt is detected, it provides realtime alerts, quarantines malicious messages, and allows for user reporting to prevent further attacks.
3. **Web Security:** Machine learning algorithms can be used to monitor web traffic for malicious activity, which can be used to detect phishing attempts and other cyber threats in real-time. This web security application uses machine learning to bolster its threat detection capabilities. It monitors web traffic in real-time, identifying and blocking access to malicious websites and phishing attempts.
4. **Endpoint Protection:** Machine learning algorithms can be used to protect endpoints such as laptops, mobile devices, and desktop computers from phishing attacks, malware, and other cyber threats. Within an endpoint protection suite, this application employs machine learning to analyze URLs accessed by endpoints. It checks URLs against known databases of malicious websites and detects potential phishing sites. When a malicious link is identified, it blocks access and provides alerts to endpoint users.
5. **Anti-Phishing Solutions:** Machine learning algorithms can be used to develop anti-phishing solutions that can help businesses and organizations detect and prevent phishing attacks. This dedicated anti-phishing solution combines machine learning with AI algorithms to monitor, detect, and manage phishing attempts. It analyzes URLs, content, and user behavior to identify fraudulent sites and phishing emails. It offers real time alerts, reporting, and automated response actions.

Acknowledgement:

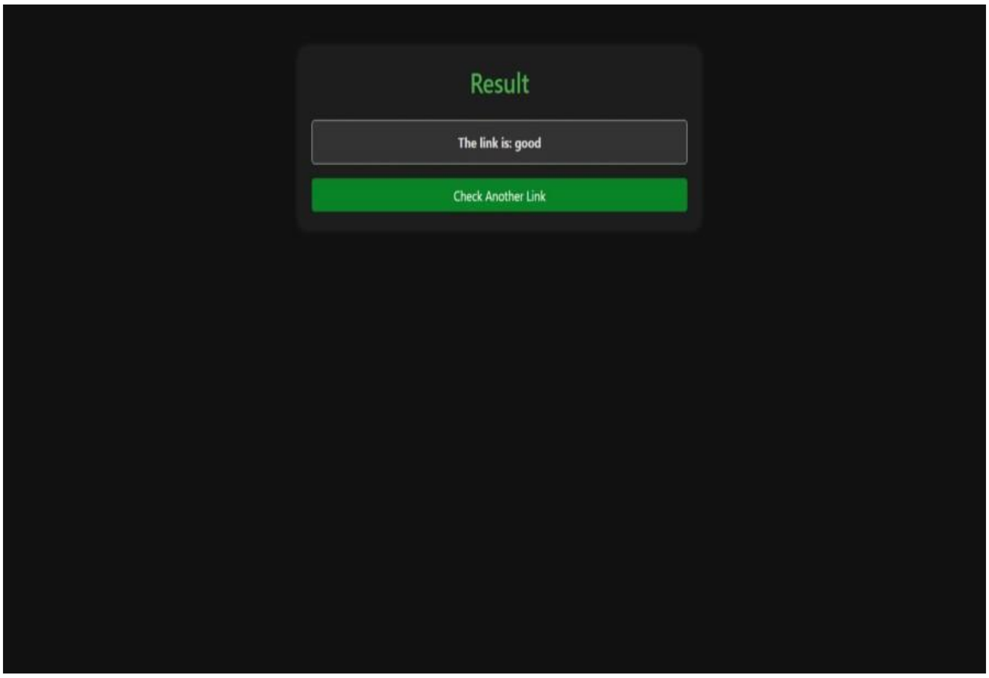
I would like to show my sincere gratitude towards Mr. G. B. Katkade HOD, Department of Computer Technology, K. K. Wagh Polytechnic, Nashik and Mr. S. H. Sangale, Sr. Lecturer, Department of Computer Technology, K. K. Wagh Polytechnic, Nashik for their valuable guidance and encouragement.

Result

Checking Amazon Shopping Link

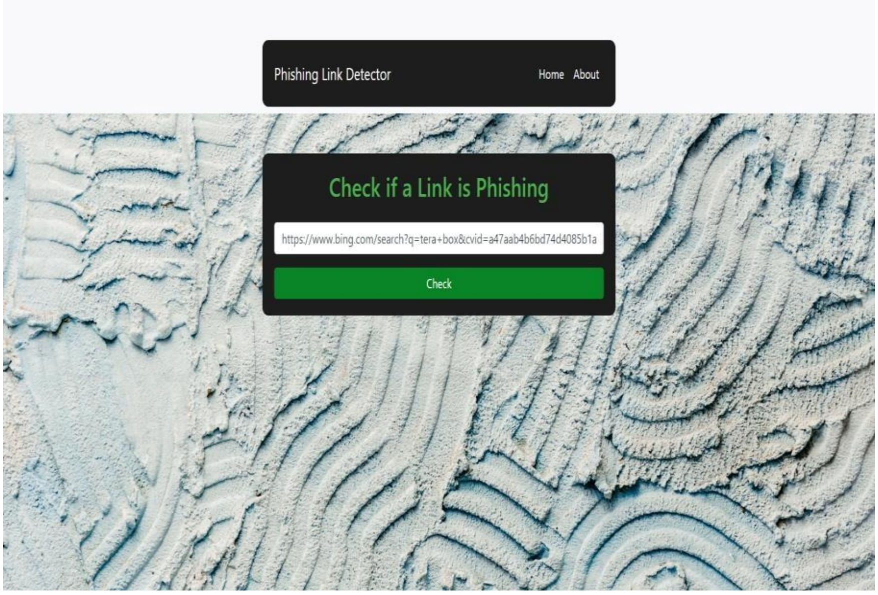


https://www.bing.com/search?pglt=43&q=amazon&cvid=b528351f17fe4b9



The link is: good

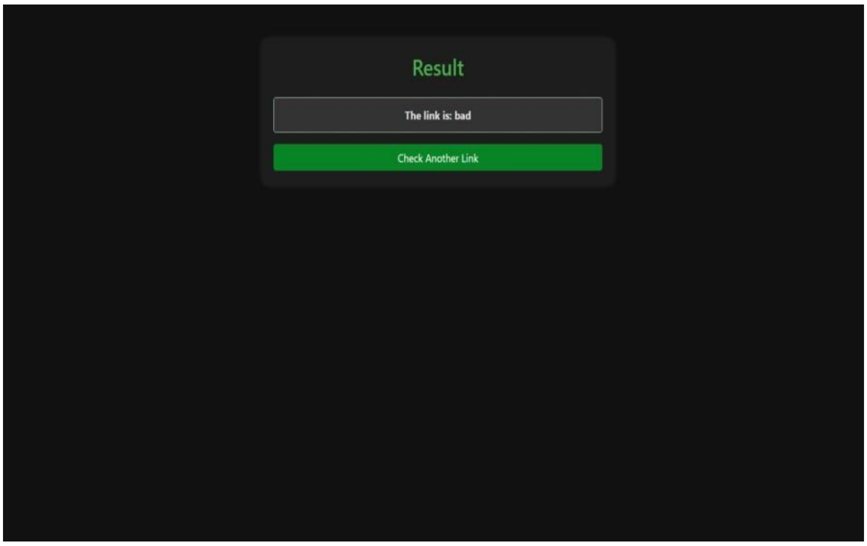
Checking Tera-Box Link



Phishing Link Detector

Check if a Link is Phishing

https://www.bing.com/search?q=tera+box&cvid=a47aab4b6bd74d4085b1a



Result

The link is: bad

Conclusion

In this project, we have explored how well to classify phishing URLs from the given set of URLs containing benign and phishing URLs. We have also discussed the randomization of the dataset, feature engineering, feature extraction using lexical analysis host-based features and statistical analysis. We have also used different classifiers for the comparative study and found that the findings are almost consistent across the different classifiers. We also observed dataset randomization yielded a great optimization and the accuracy of the classifier improved significantly. We have adopted a simple approach to extract the features from the URLs using simple regular expressions. There could be more features that can be experimented and that might lead to improving further the accuracy of the system. The dataset used in this paper contains the URLs list which may be a little old, hence regular continuous-training along with a new dataset would enhance the model accuracy and performance significantly. In our experiment we have not used the content based features as the main problem with the contentbased strategy for detecting phishing URLs is the non-availability of phishing web-sites and the life span of the phishing website is small, and it is difficult to train an ML classifier based on its content-based features. In the future, we would like to incorporate a rule-based prediction based on the content analysis of a URL. Hence, the combination of classification based lexicalanalyzer along with a rule-based URL content analyzer for phishing URL detection would provide a comprehensive solution.

References

Book References:

1. M. M. Vilas, K. P. Ghansham, S. P. Jaypralash & P. Shila, "Detection of Phishing Website Using Machine Learning Approach," 2019 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT), Mysuru, India, 2019, pp. 384-389, doi: 10.1109/ICEECCOT46775.2019.9114695.
2. M. Aljabri and S. Mirza, "Phishing Attacks Detection using Machine Learning and Deep Learning Models," 2022 7th International Conference on Data Science and Machine Learning Applications (CDMA), Riyadh, Saudi Arabia, 2022, pp. 175-180, doi: 10.1109/CDMA54072.2022.00034.
3. A. Mandadi, S. Boppana, V. Ravella and R. Kavitha, "Phishing Website Detection Using MachineLearning," 2022 IEEE 7th International conference for Convergence in Technology (I2CT), Mumbai, India, 2022, pp. 1-4, doi: 10.1109/I2CT54291.2022.9824801.
4. Hemali Sampat, Manisha Saharkar, Ajay Pandey and Hezal Lopes, “Detection of Phishing Website Using Machine Learning,” 2018 International Research Journal of Engineering and Technology (IRJET),2018, e-ISSN: 2395- 0056, p-ISSN: 2395- 0072.

Bibliography:

1. https://www.**ibm**.com/developerworks/opensource/t op-**projects**/php/
2. www.research.**ibm**.com/labs/africa/**project**- lucy.shtml
3. www.idc.**iit**b.ac.in/**projects**/student/**project**-areas.htm