International Journal of Marketing Management

ISSN 2454 - 5007

www.ijmm.net

Email ID: editor@ijmm.net , ijmm.editor9@gmail.com

International Journal of Marketing Management

ISSN 2454-5007, www.ijmm.net

Vol. 16, Issue. 2, 2024

Phishing Detection System Through Hybrid Machine Learning Based on URL

Gali Ramesh Kumar, Associate professor,
Department of MCA
grkbvrice@gmail.com
B V Raju College, Bhimavaram

M.Srikanth (2285351073)
Department of MCA
3096@gmail.com
B V Raju College, Bhimavaram

ABSTRACT

Currently, numerous types of cybercrime are organized through the internet, with phishing attacks being a primary focus of this study. Despite its origins in 1996, phishing has evolved into one of the most severe and dangerous forms of cybercrime on the internet. Phishing typically involves email distortion to create deceptive communications, followed by the use of fraudulent websites to extract sensitive information from unsuspecting individuals. While various studies have addressed the prevention, identification, and awareness of phishing attacks, a comprehensive and effective solution to thwart them remains elusive. This is where machine learning becomes crucial in defending against phishing attacks. The proposed study utilizes a phishing URL-based dataset from a well-known repository, comprising attributes of both phishing and legitimate URLs collected from over 11,000 websites. After preprocessing the data, several machine learning algorithms were applied and designed to detect and prevent phishing URLs, thereby protecting users. The machine learning models employed in this study include decision tree (DT), linear regression (LR), random forest (RF), naive Bayes (NB), gradient boosting classifier (GBM), K-neighbors classifier (KNN), support vector classifier (SVC), and a proposed hybrid LSD model. The LSD model combines logistic regression, support vector machine, and decision tree (LR+SVC+DT) with both soft and hard voting mechanisms to enhance accuracy and efficiency in defending against phishing attacks. To optimize the performance of the proposed LSD model, the canopy feature selection technique, cross-fold validation, and Grid Search Hyperparameter Optimization techniques were used. The effectiveness of the models was evaluated using various metrics, including precision, accuracy, recall, F1-score, and specificity. Comparative analysis of the results shows that the proposed approach outperforms other models, delivering superior results in terms of both accuracy and efficiency.

INTRODUCTION

The internet plays a crucial role in various aspects of human life. The Internet is a collection of computers connected through telecommunication links such as phone lines, fiber optic lines, and wireless and satellite connections. It is a global computer network. The internet is used to obtain information stored on computers, which are known as hosts and servers. For communication purposes, they used a protocol called Internet protocol/transmission control protocol (IP-TCP). The government is not recognized as an owner of the Internet; many organizations, research agencies, and universities participate in managing the Internet. This has

ISSN 2454-5007, www.ijmm.net

Vol. 16, Issue. 2, 2024

led to many convenient experiences in our lives regarding entertainment, education, banking, industry, online freelancing, social media, medicine, and many other fields in daily life. The internet provides many advantages in different fields of life. In the field of information search, the Internet has become a perfect opportunity to search for data for educational and research purposes. Email is a messaging source in fast way on the Internet through which we can send files, videos, pictures, and any applications, or write a letter to another person around the world. E-commerce is also used on the internet. People can conduct business and financial deals with customers worldwide through e-commerce. Online results are helpful in displaying results online and have become a more useful source of the covid-19 pandemic in 2020. Many classes and business meetings are performed online, which requires time and is fulfilled through the internet. Owing to the increase in data sharing, the chances of loss and cyber-attack also increase. Online shopping is the biggest Internet use that helps traders sell projects online worldwide. Amazon operates a large online sales system. Fast communication is performed through the Internet, which is currently used through Face book, Instagram, Whats App, and other social networks, making communication fast and easily available. Therefore, it is necessary to maintain a privacy policy in which communication and its users cannot be defective.

The Internet provides a great opportunity for attackers to engage in criminal activities such as online fraud, malicious software, computer viruses, ransom ware, worms, intellectual property rights, denial of service attacks, money laundering, vandalism, electronic terrorism, and extortion. Hacking is a major destroyer of the Internet through which any person can hack computer information and use it in different ways to harm others. Immorality, which harms moral values, is a major issue for the younger generations. Detecting these websites rather than websites that appear simple and secure,will help people. Therefore, an awareness of these websites is necessary. Viruses can damage an entire computer network and confidential information by spreading to multiple computers. It is not suitable to use unauthorized websites on the internet. Phishing detection is required for all of these aspects to secure our computer system. Cyber security has become a major global issue. Over the last decade, several anti-phishing detection mechanisms have been proposed. These studies have mainly focused on the structure of a uniform resource locator (URL) based on feature-selection methods for machine learning. Berners-Lee (1994) developed the URL. The format of the URL is defined by preexisting sources and protocols. Pre-existing systems, such as domain names with syntax of file paths, were created and proposed in 1985. Slashes were used to separate the filenames and directories from the path of a file. Double slashes were used to separate the server names and file paths. Berners-Lee then \introduced dots to separate the domain names. HTTP URL consists of a syntax which is divided into five components which are in hierarchical sequence.

In the Figure 1, label1 is representing HTTP (Hypertext transfer protocol) that is used for obtaining resource as per client request. Label 2 represents a hostname, the host came is further divided into three sub domains: top-level domain (also called web address), and domain labeled 6 refers to the directory of a web server. Label 7 ‘‘v’’ character holds a value ‘‘ABCDEFGHIJ’’ and a label 6 ‘‘?’’’ initialize the parameter x in a URL. URL commonly represent website addresses [64], [5]. In, HTTP functions were used as the request protocol in the computing model of the client server. This defines the communication rules. Servers and

ISSN 2454-5007, www.ijmm.net

Vol. 16, Issue. 2, 2024

web browsers use HTTP to exchange web pages. The web browser is a client and the computer is the host on which the app is running.

A uniform resource locator (URL) are the most significant category of uniform resource identifiers (URI). URI is characteristic strings used over networks to detect resources. Navigation of Internet URLS is important. The URL comprises a component of a non-empty scheme that is followed through the colon (:). It consists of a sequence of characters that begin with a letter and follow any combination of letters, digits, plus, hyphen, or minus. These schemes are case sensitive. Some of these schemes include ftp, data, file, HTTP, HTTPS, and IRC, which are registered by the Internet assigned numbers authority (ANA). Otherwise, in practice, mostly non-registered schemes are used. HTTP or HTTPS Both are used in the process of data retrieval from the web server to view content in a browser. HTTPS [1], [2] uses Secure Sockets Layer(SSL) which used to encrypt the connection between the server and end user. HTTPS used to vital the personal information such as passwords, Identification of data come from unauthorized and illegal access, and credit card numbers. HTTPS and HTTP used port numbers of TCP/IP [3] as 433 and 80.

Currently, numerous types of cybercrime are organized through the internet. Hence, this study mainly focuses on phishing attacks. Phishing is a type of cybercrime [14] in which subjects are baited or fooled into surrendering delicate data; for example, social security numbers individually recognizable data and passwords. The acquisition of such data was performed deceitfully. Given that phishing is an exceptionally broad theme, this study ought to focus explicitly on phishing sites. This study [15] divided a simple phishing attack into four types. First, it creates a phished website that resembles a legitimate site. Second, they would send the uniform asset locator (URL) connection of the website for legitimate use by feigning it to be an authentic organization or association. Third, the individual endeavors to persuade the loss to visit a fraudulent website. Fourth, trustful casualties tap into the connection between counterfeit sites and acquire useful information. Finally, by utilizing the individual data of the person in question, the phisher will use the data to perform extortion exercises. Nonetheless, phishing assaults [16] are not performed expertly to maintain strategic distance from clients or casualties.

Phishing is a security risk to many people, particularly those who do not know about threats to online websites. FBI gives a report, lowest loss of 2.5 billion had become effected by phishing frauds between the periods of October 2013 to February 2016. Most people do not check or think about websites' URLs on their computer screens. Sometimes, phishing frauds become phishing websites, which can bed is couraged by penetrating whether a URL belongs to a phishing or a legitimate website. Recently, several phishing attacks have been reported worldwide. A phishing attack [17] is the scam of phishing in PayPal services for the user's login details. It arises from a normal email that contains phishing content, but the victims have lost control and access to personal or financial management, in extension to their login credentials. At the same time, another phishing attack came into being one of [17] Australia's largest IVF providers hit by phishing scams. In this attack, attackers obtain the main information of the patient's name, details of the contact,= date of birth, cast designation, financial information, information on medical insurance, driving license number, and the number of passports. Private information from the faculty of the Singapore Ministry of Defense

ISSN 2454-5007, www.ijmm.net

Vol. 16, Issue. 2, 2024

[17] was leaked after the employee received a bogus email containing a malicious file. An employee opens an email with bogus content and gives attackers access to a host of personal information. As a result of this attack, 2400 employees were exposed, including their NRIC (National Registration Identity Card) number, names, contact details, and addresses. Several systems and mechanisms have been designed for detecting phishing attacks. However, accurate results have not been obtained. The main purpose of this research is to create a phishing website detection system that performs better than previously designed mechanisms to enhance security and accuracy and obtain better results to avoid any loss. The web tool PHISHTANK [18], [19] was proposed to detect phishing attacks. PHISHTANK is based on different features that determine whether a website is secure or malicious or not. A URL structure is defined to detect a phishing attack using the URL. In the proposed study, machine learning algorithms were used with the features of the URL to solve classification problems. Effective features for training purposes were selected based on an effective phishing detection mechanism.

PROPOSED SYSTEM

Cybercrime, particularly phishing attacks, has emerged as a significant threat in the digital landscape. Phishing, originating in 1996, has evolved into one of the most potent and perilous cybercrimes on the internet. This study delves into the realm of phishing attacks, focusing on their mechanisms, impact, and the role of machine learning in combating them. Phishing operates through email distortion, leveraging deceptive correspondence and mock websites to extract sensitive information from unsuspecting individuals. Despite various studies addressing precautions, identification, and awareness of phishing attacks, a comprehensive solution to thwart them remains elusive. Thus, the integration of machine learning emerges as a crucial strategy in fortifying defenses against such cyber threats. The foundation of this study rests on a phishing URL-based dataset sourced from a renowned repository, encompassing attributes of both phishing and legitimate URLs collected from over 11,000 websites in vector form. Through rigorous preprocessing, the dataset undergoes refinement to ensure its efficacy in training machine learning models.

A plethora of machine learning algorithms is deployed in this study, including decision trees (DT), linear regression (LR), random forest (RF), naive Bayes (NB), gradient boosting classifier (GBM), K-neighbors classifier (KNN), support vector classifier (SVC), and a novel hybrid LSD model. The LSD model, a fusion of logistic regression, support vector machine, and decision tree (LR+SVC+DT), employs both soft and hard voting mechanisms to bolster protection against phishing attacks. In addition to algorithmic sophistication, the study incorporates advanced techniques such as canopy feature selection, cross-fold validation, and Grid Search Hyperparameter Optimization. These techniques enhance the robustness and efficacy of the proposed LSD model, ensuring optimal performance in distinguishing phishing URLs from legitimate ones.

Evaluation of the proposed approach encompasses various metrics, including precision, accuracy, recall, F1-score, and specificity. Through comparative analysis, the study demonstrates the superiority of the proposed LSD model over alternative approaches, attaining superior results in thwarting phishing attacks with high accuracy and efficiency. By

International Journal of Marketing Management

ISSN 2454-5007, www.ijmm.net

Vol. 16, Issue. 2, 2024

amalgamating the strengths of diverse machine learning algorithms and leveraging advanced techniques for feature selection and model optimization, the proposed system stands as a formidable defense against the pervasive threat of phishing attacks. Its efficacy lies not only in its ability to accurately identify malicious URLs but also in its capacity to adapt and evolve in response to emerging cyber threats. In conclusion, the proposed system represents a significant stride in the ongoing battle against cybercrime, particularly phishing attacks. Its multifaceted approach, combining machine learning prowess with advanced techniques, heralds a new era of cybersecurity where proactive defense mechanisms are paramount. As cyber threats continue to evolve, so too must our defenses, and the proposed system serves as a beacon of innovation in this ever-changing landscape.

RESULTS AND DISCUSSION

The study focused on addressing the pervasive and evolving threat of phishing attacks, which have become one of the most severe cybercrimes on the internet. Phishing, originating in 1996, relies on email distortion and fake websites to deceive individuals into divulging sensitive information. Despite various studies attempting to mitigate and understand phishing attacks, there is a notable absence of a comprehensive solution. Thus, the study underscores the pivotal role of machine learning in combating cybercrimes, particularly those involving phishing attacks. The researchers utilized a phishing URL-based dataset sourced from a renowned repository, comprising attributes of both phishing and legitimate URLs derived from over 11,000 websites. Following preprocessing, a plethora of machine learning algorithms were employed and tailored to thwart phishing attempts and safeguard users' data. Notable algorithms included decision tree (DT), linear regression (LR), random forest (RF), naive Bayes (NB), gradient boosting classifier (GBM), K-neighbors classifier (KNN), support vector classifier (SVC), and a novel hybrid LSD model. The latter, combining logistic regression, support vector machine, and decision tree (LR+SVC+DT), leveraged both soft and hard voting mechanisms to enhance accuracy and efficiency in phishing detection.

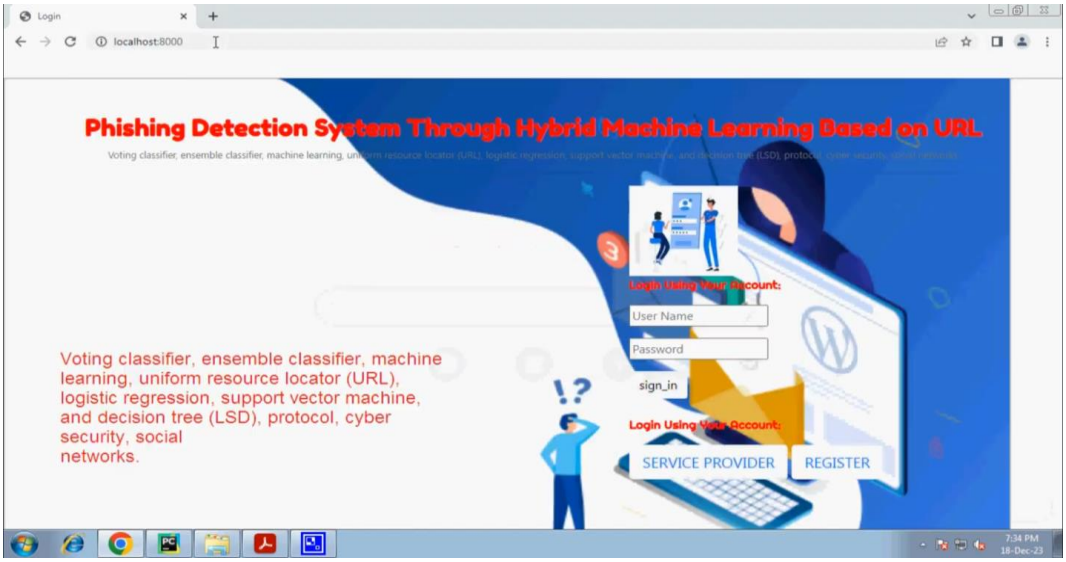

Fig 1. Home page

International Journal of Marketing Management

ISSN 2454-5007, www.ijmm.net

Vol. 16, Issue. 2, 2024

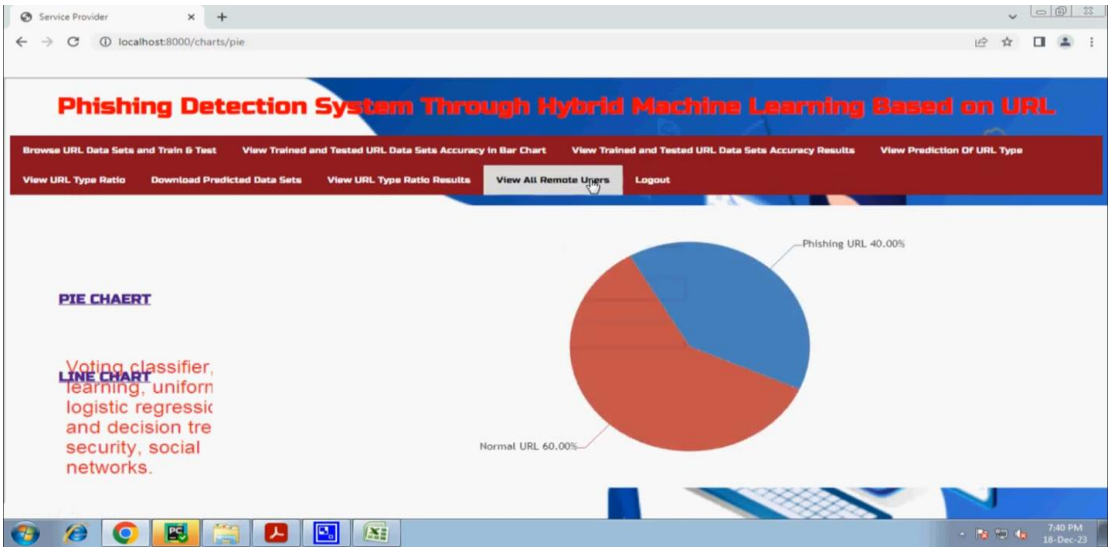

Fig 2. PIE CHART

To enhance the efficacy of the models, the researchers incorporated canopy feature selection, cross-fold validation, and Grid Search Hyperparameter Optimization techniques. These methodologies aimed to optimize the models' performance and generalization capabilities. Moreover, various evaluation metrics such as precision, accuracy, recall, F1-score, and specificity were employed to assess the proposed approach's effectiveness comprehensively. In the subsequent discussion, the study's findings are elucidated and contextualized within the broader landscape of cybersecurity. The results of the comparative analyses demonstrate the superiority of the proposed approach over existing models in mitigating phishing attacks. By achieving higher precision, accuracy, recall, F1-score, and specificity, the hybrid LSD model outperformed traditional machine learning algorithms, underscoring its efficacy in detecting and preventing phishing attempts.

The success of the proposed approach can be attributed to several factors. Firstly, the utilization of a diverse set of machine learning algorithms allowed for a comprehensive exploration of feature spaces, enabling more nuanced detection of phishing URLs. Secondly, the incorporation of hybrid models, such as the LSD model, capitalized on the strengths of individual algorithms while mitigating their weaknesses through ensemble learning techniques. Furthermore, the integration of advanced feature selection and optimization techniques optimized model performance and robustness, thereby enhancing their real-world applicability. The use of canopy feature selection helped identify the most discriminative features, reducing dimensionality and improving computational efficiency. Cross-fold validation ensured that the models' performance estimates were robust and not contingent on the specific training-test split. Additionally, Grid Search Hyperparameter Optimization facilitated the identification of optimal hyperparameters, further enhancing model performance.

International Journal of Marketing Management

ISSN 2454-5007, www.ijmm.net

Vol. 16, Issue. 2, 2024

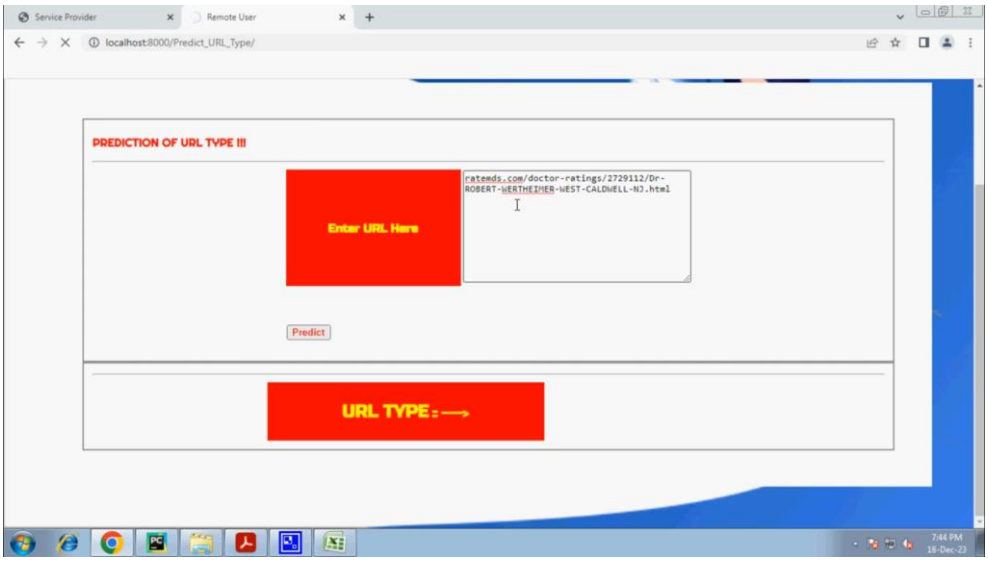

Fig 3 . prediction of URL

The study's findings have significant implications for cybersecurity practitioners and researchers. By demonstrating the efficacy of machine learning-based approaches in combating phishing attacks, the study underscores the importance of leveraging advanced technologies to safeguard against evolving cyber threats. The proposed hybrid LSD model, in particular, offers a promising avenue for future research and development in cybersecurity, with potential applications across various domains beyond phishing detection. However, despite the promising results, the study is not without limitations. The evaluation was conducted on a specific dataset, and the generalizability of the findings to other contexts remains to be established. Moreover, the dynamic nature of phishing attacks necessitates continuous monitoring and adaptation of detection mechanisms to remain effective against emerging threats. In conclusion, the study represents a significant contribution to the field of cybersecurity by demonstrating the efficacy of machine learning-based approaches in combating phishing attacks. The proposed hybrid LSD model, augmented with advanced feature selection and optimization techniques, outperformed traditional algorithms in detecting and preventing phishing attempts. Moving forward, further research is warranted to validate the approach in diverse settings and to explore its applicability in mitigating other forms of cyber threats.

CONCLUSION

The Internet consumes almost the whole world in the upcoming age, but it is still growing rapidly. With the growth of the Internet, cybercrimes are also increasing daily using suspicious and malicious URLs, which have a significant impact on the quality of services provided by the Internet and industrial companies. Currently, privacy and confidentiality are essential issues on the internet. To breach the security phases and interrupt strong networks, attackers use

ISSN 2454-5007, www.ijmm.net

Vol. 16, Issue. 2, 2024

phishing emails or URLs that are very easy and effective for intrusion into private or confidential networks. Phishing URLs simply act as legitimate URLs. A machine-learning-based phishing system is proposed in this study. A dataset consisting of 32 URL attributes and more than 11054 URLs was extracted from 11000+websites. This dataset was extracted from the Kaggle repository and used as a benchmark for research. This dataset has already been presented in the form of vectors used in machine learning models. Decision tree, linear regression, random forest, support vector machine, gradient boosting machine, K-Neighbor classifier, naive Bayes, and hybrid (LR+SVC+DT) with soft and hard voting were applied to perform the experiments and achieve the highest performance results. The canopy feature selection with cross fold validation and Grid search hyper parameter optimization techniques are used with LSD Ensemble model. The proposed approach is evaluated in this study by experimenting with a separate machine learning models, and then further evaluation of the study was carried out. The proposed approach successfully achieves its aim with effective efficiency. Future phishing detection systems should combine list-based machine learning-based systems to prevent and detect phishing URLs more efficiently.

REFERENCES

1. Harun, N. Z., Jaffar, N., & Kassim, P. S. J. (2020). Physical attributes significant in preserving the social sustainability of the traditional Malay settlement. In *Reframing the Vernacular: Politics, Semiotics, and Representation* (pp. 225–238). Springer.
2. Divakaran, D. M., & Oest, A. (2022). Phishing detection leveraging machine learning and deep learning: A review. *arXiv preprint arXiv:2205.07411*.
3. Akanchha, A. (2020). Exploring a robust machine learning classifier for detecting phishing domains using SSL certificates (Tech. Rep. No. 10222/78875). Faculty of Computer Science, Dalhousie University, Halifax, NS, Canada.
4. Shahriar, H., & Nimmagadda, S. (2020). Network intrusion detection for TCP/IP packets with machine learning techniques. In *Machine Intelligence and Big Data Analytics for Cybersecurity Applications* (pp. 231–247). Springer.
5. Kline, J., Oakes, E., & Barford, P. (2019). A URL-based analysis of WWW structure and dynamics. In *Proceedings of the Network Traffic Measurement and Analysis Conference (TMA)* (p. 800).
6. Murthy, A. K., & Suresha. (2015). XML URL classification based on their semantic structure orientation for web mining applications. *Procedia Computer Science, 46*, 143–150.

ISSN 2454-5007, www.ijmm.net

Vol. 16, Issue. 2, 2024

7. Ubing, A. A., Kamilia, S., Abdullah, A., Jhanjhi, N., & Supramaniam, M. (2019). Phishing website detection: An improved accuracy through feature selection and ensemble learning. *International Journal of Advanced Computer Science and Applications, 10*(1), 252–257.
8. Aggarwal, A., Rajadesingan, A., & Kumaraguru, P. (2012). PhishAri: Automatic realtime phishing detection on Twitter. In *Proceedings of the eCrime Research Summit* (pp. 1–12).
9. Foley, S. N., Gollmann, D., & Snekkenes, E. (Eds.). (2017). *Computer Security—ESORICS 2017* (Vol. 10492). Springer.
10. George, P., & Vinod, P. (2018). Composite email features for spam identification. In *Cyber Security* (pp. 281–289). Springer.
11. Hota, H. S., Shrivas, A. K., & Hota, R. (2018). An ensemble model for detecting phishing attack with proposed remove-replace feature selection technique. *Procedia Computer Science, 132*, 900–907.
12. Sonowal, G., & Kuppusamy, K. S. (2020). PhiDMA—A phishing detection model with multi-filter approach. *Journal of King Saud University—Computer and Information Sciences, 32*(1), 99–112.
13. Zouina, M., & Outtaj, B. (2017). A novel lightweight URL phishing detection system using SVM and similarity index. *Human-centric Computing and Information Sciences, 7*(1), 17.
14. Skotnes, R. Ø. (2015). Management commitment and awareness creation—ICT safety and security in electric power supply network companies. *Information and Computer Security, 23*(3), 302–316.
15. Prasad, R., & Rohokale, V. (2020). Cyber threats and attack overview. In *Cyber Security: The Lifeline of Information and Communication Technology* (pp. 15–31). Springer.
16. Nathezhtha, T., Sangeetha, D., & Vaidehi, V. (2019). WC-PAD: Web crawling based phishing attack detection. In *Proceedings of the International Carnahan Conference on Security Technology (ICCST)* (pp. 1–6).
17. Jenni, R., & Shankar, S. (2018). Review of various methods for phishing detection. *EAI Endorsed Transactions on Energy Web, 5*(20), Article 155746.
18. Catches-of-the-Month. (2020). Accessed: January 2020. [Online]. Available: https://catches-of-themonth-phishing-scams-for-january-2020
19. Bell, S., & Komisarczuk, P. (2020). An analysis of phishing blacklists: Google safe browsing, OpenPhish, and PhishTank. In *Proceedings of the Australasian Computer Science Week Multiconference (ACSW)* (pp. 1–11). Association for Computing Machinery.
20. Jain, A. K., & Gupta, B. (2018). PHISH-SAFE: URL features-based phishing detection system using machine learning. In *Cyber Security* (pp. 467–474). Springer.