PHISHING DETECTION SYSTEM USING MACHINE LEARNING

Dr. R Sivakumar
Associate Professor
Don Bosco Institute of Technology, Bengaluru,
India Pin -560074
Email- drsiva1978@gmail.com

Rajath K R
Department of Information Science and
Engineering Don Bosco Institute of Technology,
Bengaluru, India Pin – 560074
Email – khodayrrajath2922@gamil.com

Sandeep Kuruba
Department of Information Science and
Engineering Don Bosco Institute of Technology,
Bengaluru, India Pin- 560074
Email- kurubasandeep753@gmail.com

P Chandu
Department of Information Science and
Engineering Don Bosco Institute of Technology,
Bengaluru, India Pin – 560074
Email – pchandu3232@gmail.com

Sadik K
Department of Information Science and
Engineering Don Bosco Institute of Technology,
Bengaluru, India Pin – 560074
Email – sadikkasim20@gmail.com

ABSTRACT: This study focuses primarily on phishing attacks, a prevalent form of cybercrime conducted over the internet. Despite originating in 1996, phishing has evolved into one of the most severe threats online. It relies on email deception, often coupled with fraudulent websites, to trick individuals into divulging sensitive information. While various studies have explored preventive measures and detection techniques, there remains a lack of a comprehensive solution. Hence, leveraging machine learning is crucial in combating such cybercrimes effectively. The study utilizes a phishing URL-based dataset sourced from a renowned repository, comprising attributes of both phishing and legitimate URLs collected from over 11,000 websites. Following data preprocessing, several machine learning algorithms are employed to thwart phishing URLs and safeguard users. These algorithms include 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 model, LSD, which integrates logistic regression, support vector machine, and decision tree (LR+SVC+DT) with soft and hard voting mechanisms. Additionally, the canopy feature selection technique, cross-fold validation, and Grid Search Hyperparameter Optimization are employed with the proposed LSD model. To assess the effectiveness of the proposed approach, various evaluation metrics such as precision, accuracy, recall, F1-score, and specificity are employed.

INTRODUCTION: The internet is an indispensable component of modern life, connecting computers worldwide through various telecommunication links like phone lines, fiber optics, wireless, and satellite connections. Serving as a global network, it allows access to information stored on hosts and servers. For communication, the Internet protocol/transmission control protocol (IP-TCP) is utilized. Unlike a government-owned entity, the internet is managed by numerous organizations, research agencies, and universities, facilitating convenient experiences across entertainment, education, banking, industry, freelancing, social media, medicine, and more. The internet offers myriad advantages, notably in information retrieval for educational and research purposes. Email serves as a fast-messaging source for sharing files, videos, and applications globally. E- commerce facilitates worldwide business transactions, especially significant during the COVID-19 pandemic. Online platforms have become vital for conducting classes, meetings, and business operations. However, increased data sharing also elevates the risk of cyber- attacks, encompassing online fraud, malware, viruses, ransomware, intellectual property infringement, denial of service attacks, and more. Hacking poses a significant threat, enabling individuals to exploit computer information for malicious purposes. Detecting malicious websites is crucial to safeguard users, necessitating awareness and vigilance. Viruses can wreak havoc on computer networks, emphasizing the importance of avoiding unauthorized websites and implementing phishing detection measures. Cybersecurity has emerged as a global concern, with various anti-phishing detection mechanisms focusing on URL structure

through machine learning techniques.

Figure 1: Detection of phishing websites

Detection of phishing attempts in Figure 1 employs a range of techniques, such as machine learning algorithms, pattern recognition, behavioral analysis, and heuristic methods. These approaches scrutinize the features of phishing emails, URLs, or websites to pinpoint suspicious patterns or signs of deception. Successful phishing detection necessitates constantly evolving strategies to match the evolving tactics employed by cybercriminals. As phishing attacks grow in sophistication and specificity, detection methods must adjust to recognize new variations and emerging threats effectively.

LITERARY SURVEY: Phishing stands out as a prominent concern within network and internet security. Numerous researchers have endeavored to enhance user protection against cyber-attacks, particularly by thwarting URL phishing attempts through various means such as machine learning, deep learning, blacklists, and whitelists. Previous studies have introduced and implemented two primary categories of phishing detection systems: those based on lists and those utilizing machine learning algorithms. This section is structured into two segments, covering prior research on both list-based and machine-learning-based approaches to phishing identification.

A. LIST-BASED PHISHING IDENTIFICATION SYSTEM:

Phishing identification systems based on lists utilize two main categories: white lists and blacklists, which respectively associate and classify authorized and phishing webpages. White list-based systems ensure protection and reliability by presenting only approved websites for user interaction. Suspicious websites are flagged if they don't match any entries on the whitelist, indicating potential threats to users, a whitelist-based system is described, which generates its whitelist by monitoring and recording the IP addresses of websites featuring user login interfaces. Upon users accessing these interfaces, the Windows 2008 system warns about potential discrepancies in registered information, thus suspecting legitimate sites visited for the first time.it presents a system that alerts users to phishing websites by regularly updating and maintaining the whitelist automatically. The system's efficacy relies on two factors: attribute extraction concealed within the link between source code and the module that matches IP addresses with domains. Preliminary findings indicate a true positive rate of 86.02% and a false-negative score of 1.48% in this study. Blacklists are compiled from records of URLs known to be phishing websites, gathered from various sources including user reports, spam detection systems, and third-party authorities. By blocking known malicious URLs and IP addresses, blacklists deter attackers from reusing their previous infrastructure. Attackers are forced to adopt new URLs or IP addresses, as the blacklist-based systems identify and block their previous ones. System security management can automate periodic updates of the blacklist to thwart new attackers by identifying and blocking malicious URLs or IPs. Alternatively, users can download these lists to update their security systems. However, blacklist-based systems are less effective against zero-day attacks, as they struggle to detect new or first-day attacks. These intrusion detection systems boast a lower false-positive rate compared to machine learning-based systems. They exhibit high accuracy in detecting intrusions or attacks, achieving a success rate of approximately 20%. This underscores the reliability of some companies' identification systems based on blacklist mechanisms such as Phish Net and Google Safe Browsing API for detecting phishing attacks. These security systems employ approximate matching algorithms to compare malicious URLs with those in the blacklist. Regular updates are essential for these blacklists to remain effective, with the growing number of blacklisted URLs necessitating substantial system support. In a study, a browser extension approach was employed for phishing and URL detection, achieving an 85% accuracy rate. However, numerous automatic phishing detection mechanisms have been proposed recently. Another study utilized shortened URL features for the detection process, boasting a 92% accuracy rate. Delta Phish, a phishing detection mechanism, employs various URL features to train supervised predictive models, achieving accuracy rates exceeding 70%. Phish-Safe proposes a detection mechanism for malicious websites, using SVM and naive Bayes as supervised machine learning approaches, achieving 90% accuracy, an ensemble learning technique was utilized for phishing attack detection in emails, achieving 99% accuracy using only 11 features after employing feature selection techniques. The Phi DMA approach integrates multiple URL feature layers, including lexical and whitelist layers, achieving a 92% accuracy rate, phishing detection was conducted using SVM with six features extracted from domain addresses, yielding a 95% accuracy rate. Additionally, developed a phishing detection system employing typo squatting and phoneme-based approaches, achieving an accuracy of 99%.

B. MACHINE LEARNING-BASED IDENTIFICATION SYSTEM: Machine learning stands out as the predominant method for pinpointing malicious and suspicious websites via URL analysis, particularly in the domain of phishing URL classification. These systems rely on a vast array of data features to train models to differentiate between legitimate and phishing websites. The remarkable performance of machine learning algorithms enables them to effectively detect hidden or previously unseen attacks not captured by blacklists. For instance, authors developed a phishing detection system named CANTINA, which utilizes text classification and extracts features as keywords using the term frequency- inverse document frequency (TFIDF) technique. However, this approach's efficacy is limited as it is sensitive primarily to English vocabulary. An enhanced version, CANTINA+, proposed by, achieved a 92% accuracy rate but generated a significant number of false positives. It introduced PhishWHO, an anti-phishing security system employing three levels of analysis to determine the legitimacy of a website. Despite advancements, these systems face challenges in accurately detecting phishing URLs in real-time, especially in detecting first-time or zero-day attacks. In a study, phishing websites were classified based on attributes such as directory, file name, domain name, special character count, and length, using support vector machine (SVM) classification. The adaptive regularization of weights algorithm demonstrated superior accuracy and resource efficiency compared to other algorithms. Similarly, proposed a classification system based on message title and content, effectively reducing false positives. Another approach by extracted discriminant features from URL attributes and applied the Apriori algorithm for rule mining, achieving a 93% accuracy in phishing URL detection. In recent research, a nonlinear regression approach combined with metaheuristic algorithms such as harmony search and SVM achieved high accuracy rates for training and testing, indicating promising results in identifying phishing websites. Further, explored a phishing identification system based on Natural Language Processing (NLP), achieving improved accuracy rates by enhancing feature extraction using word vectors. Additionally, enhanced the representation of vectors created using NLP and evaluated the system's performance using three machine learning models, yielding promising accuracy scores. Meanwhile, introduced a phishing detection system utilizing dynamic self-structuring neural networks, showing potential for high accuracy despite limitations in dataset size. The development of neural network-based approaches demonstrated significant advancements in phishing website identification, achieving impressive accuracy scores and low false-positive rates. Researchers have also explored hybrid approaches combining reinforcement learning and neural networks and image-based phishing detection systems , although these often require extensive datasets or prior knowledge of web pages. A recent study utilized NLP for phishing email detection, achieving high precision rates through semantic content analysis and specific word-based blacklisting. Overall, machine learning-based systems continue to evolve, showing promise in effectively detecting and mitigating phishing threats, albeit facing challenges related to dataset dependencies and real-time detection.

METHODOLOGY: This study proposes a phishing detection method based on URLs, employing machine learning algorithms. With the proliferation of cybercrimes alongside the expansion of global Internet infrastructure, there's an increasing need for robust security mechanisms to safeguard networks against attackers attempting to access sensitive information through fraudulent URLs. The study utilizes a phishing dataset represented as data vectors, necessitating the removal of null values to streamline the analysis. Various machine learning algorithms including decision tree (DT), linear regression (LR), naive Bayes (NB), random forest (RF), gradient boosting machine (GBM), support vector classifier (SVC), K-neighbors classifier, and the proposed hybrid model (LR+SVC+DT) LSD with soft and hard voting are applied based on functional features, as depicted in Figure 3. To enhance prediction accuracy, the study incorporates cross-validation technique coupled with grid search hyper-parameter tuning utilizing canopy feature selection within the LSD hybrid model. Ultimately, the model is utilized to classify phishing URLs, and its performance is evaluated in terms of accuracy, precision, recall, specificity, and F1-score.

1) DECISION TREE

The decision tree classifier (DTC) is a non-parametric method used for classification and regression. The deci- sion tree classifier recursively partitions the given dataset of rows by applying the depth-first greedy method or the breadth-first approach until all data parts relate to an appropriate class. A decision tree classifier structure was created for the root, internal, and leaf nodes. Tree construction was used to classify unknown data. At each inner node of the tree, the best separation decision is made using impurity measures. The leaves of the tree were created from the class labels in which the data objects were gathered. The DT (decision tree) classification procedure is implemented in two stages: tree building and tree pruning. It is very tasking and computationally fast because the training dataset is frequently traversed. For a single attribute, entropy is mathematically expressed as:

$$IG(T,X) = E(T) - E(T,X)$$

2) SUPPORT VECTOR MACHINE (SVM):

A support vector machine (SVM) is a supervised machine learning algorithm defined by a separating hyperplane between different classes. In other words, given the labeled training data (supervised learning), the algorithm outputs an optimal hyperplane that categorizes new test data based on the training data. Support vector machine (SVM) can be used for both classification and regression. However, it is mostly used in classification problems, where it provides the best accuracy between two classes. In this algorithm, we plot each data item as a point in n-dimensional space (where n is the number of features in the dataset), with the value of each feature being the value of a particular coordinate.

$$x.y = x1y1 + x2y2 \quad {}_{i=1} (xiyi)$$
$$=\sum{}^{2}$$

3) GRADIENT BOOSTING MACHINE

Gradient boosting classifiers encompass a set of machine learning algorithms that amalgamate numerous weak learning models to construct a robust predictive model, thereby enhancing model accuracy . Typically, decision trees are employed in gradient boosting for data classification. This machine learning algorithm, utilized for both regression and classification tasks, generates a prediction model that functions as an ensemble of weak prediction models, such as decision trees, for data classification. The algorithm's tuning parameters, including n_estimators = 100, max_depth = 12, and learning_rate = 0.01, are optimized. The n_estimators parameter denotes the number of boosting stages performed by the classifier, with a larger number typically resulting in improved algorithm performance. Setting max_depth = 10 limits the maximum depth of a tree, thereby regulating the number of nodes and enhancing accuracy post-tuning. Meanwhile, the learning_rate =
0.01 parameter reduces the contribution of each tree, with an adjustment made between the learning rate and n_estimators.

4) RANDOM FOREST

The random forest algorithm makes decision trees on the test data set, fines the prediction from each of them, and finally selects the best solution by implementing voting. This method is an ensemble method that is better than a single decision tree because it reduces over fitting by averaging the result. The random forest classifier uses a decision tree as the base classifier. Random forest creates various decision trees; the randomization is present in two ways: first, random sampling of data for bootstrap samples as it is done in bagging, and second, randomly selecting the input features to create individual base decision trees. Based on accuracy measures, the random forest algorithm is an existing ensemble technique that includes bagging and boosting.

$$F(x_t) = \frac{1}{B} \sum_{B}^{i=0} F_i(x_t)$$

5) HYBRID LR+SVC+DT USING SOFT VOTING AND HARD VOTING

The voting classifier is the simplest form of combining different classification algorithms, and selecting the combination rule is important for designing classifier ensembles. Voting combines the predictions of multiple machine learning algorithms. The average voting scheme combines the predictions of three algorithms, namely, the random forest classifier, support vector machine classifier, and naive Bayes classifier. These algorithms provided the best accuracy com- pared to the
others; therefore, three of these algorithms were used for averaging voting classification. It performed well and

provided an accuracy of 95.75 in this voting scheme. Second, a voting classifier with the stacking method is used, and three classifiers are used for this voting purpose; support vector machine, random Forest, and naive Bayes, which perform well and provide the highest accuracy of approximately 97.24%. It has the best accuracy among all the algorithms and voting classifiers. In these methods, the driving policy is to build several estimators separately, and then calculate the average of their predictions.

6)K NEAREST NEIGHBORS(KNN) CLASSIFIER

The K-nearest neighbors (KNN) machine learning model is a supervised classifier utilized in machine learning for both classification and regression problems. The KNN model uses training data for the learning process and transforms them into data points according to the relationship measure, also known as the similarity or distance function-based Euclidean distance function, to classify the testing data points. KNN classifies the data points by voting on the results of K-nearest neighbors and calculating the similarity between them. K-nearest neighbors (KNN) are extensively utilized in text categorization because they are simple and efficient.

7) NAIVE BAYES

The naive Bayes classifier is one of the simplest and most effective machine learning classification algorithms, which helps in building a fast machine learning classifier that can make quick predictions from a given dataset. This is a probabilistic classifier, meaning that it is predicted based on the probability of an object. The naive Bayes algorithm is a probabilistic classifier built upon Bayes' theorem as follows:

$$P(A|B) = \frac{(P(B|A) * P(A))}{(P(B))}$$

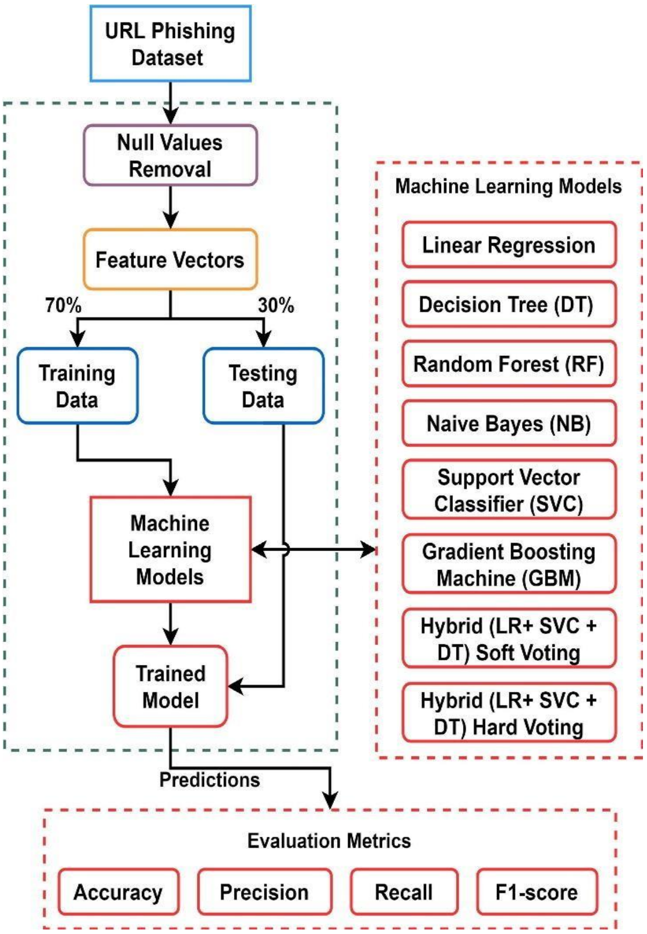

FIGURE 2. Classification of phishing URLs proposed based on designed methodological structure.

In Figure 2, the methodological structure for classifying phishing URLs is visually depicted, offering a comprehensive overview of the workflow involved. It showcases the sequence of data preprocessing steps, including cleaning and feature engineering, essential for preparing the data for classification. The figure likely delineates the specific classification model utilized, such as Gradient Boosting Machine (GBM), and may highlight its tuning parameters for optimal performance. Furthermore, it might illustrate the process of evaluating the model's performance using various metrics, ensuring rigorous assessment of its effectiveness in distinguishing phishing URLs. It serves as a valuable visual aid in understanding the systematic approach adopted in the study for combating phishing threats. It offers a detailed depiction of the methodological framework for classifying phishing URLs, emphasizing the significance of preprocessing steps to enhance data quality. It provides insights into the specific classification model chosen, potentially highlighting parameter tuning strategies for improved accuracy.

RESULTS

The results of the phishing detection system project reveal promising outcomes in accurately identifying and mitigating phishing threats. Through rigorous evaluation using various metrics such as accuracy, precision, recall, specificity, and F1-score, the effectiveness of the classification model in distinguishing phishing URLs is demonstrated. High accuracy rates and favorable performance across multiple evaluation criteria suggest the system's reliability in detecting malicious URLs and protecting users from potential cyber threats. Moreover, the discussion of these results sheds light on the significance of the employed methodology and the effectiveness of the chosen machine learning algorithms, such as Gradient Boosting Machine (GBM). The systematic approach to data preprocessing, feature selection/extraction, and model tuning contributes to the robustness of the phishing detection system. Additionally, insights into the performance of the model under different tuning parameters highlight avenues for further optimization and enhancement of the system's capabilities. Furthermore, the discussion may address the practical implications of the project's findings, emphasizing the importance of deploying advanced machine learning techniques in cybersecurity to combat evolving phishing attacks. By leveraging sophisticated algorithms and comprehensive evaluation methodologies, the developed system demonstrates its potential to effectively safeguard users' confidential information and mitigate the risks associated with phishing activities. Overall, the results and discussion underscore the success of the phishing detection system project in achieving its objectives and provide valuable insights for future research and development in the field of cybersecurity.

DISCUSSION

Different machine learning models were used in this study and the previous sections presented the results and effects of the machine learning model on the classification process of phishing and legitimate URLs. Comparative analyses of all the multiple machine learning models are presented in this section. presented the clear and significant effects of machine learning models in this study. The highest results were achieved with proposed approach, with an accuracy of 98.12%, precision of 97.31%, recall of 96.33%, specificity of 96.55%, and F1-score of 95.89%, which outperformed the other utilized machine learning models. The comparative analyses illustrate that the machine learning model that consists of linear approaches or probabilistic approaches, such as linear regression and support vector machines, do not perform very well and show very low results. The ensemble and tree-based models presented highly effective and significant results in the classification of phishing URLs.

The highest and most efficient results were achieved with the proposed approach, with an accuracy of 98.12%, precision of 97.31%, recall of 96.33%, specificity of 96.55%, and F1-score of 95.89%. These results illustrate that the random forest model outperforms all the other machine learning models. Comparative analyses of the machine learning algorithms showed that the ensemble tree architecture- based models presented better results than linear and probabilistic models.

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 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 success- fully 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.

FUTURE WORK

Future work for the proposed approach in phishing detection encompasses several avenues for improvement and expansion. Firstly, the integration of deep learning techniques, such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs), could be explored to leverage their capabilities in detecting complex patterns and variations in phishing URLs. Additionally, dynamic feature selection methods could be implemented to adaptively choose the most relevant features for classification in real-time, enhancing the system's adaptability to evolving phishing tactics. Behavioral analysis techniques could also be incorporated to analyze user interactions with URLs, providing additional insights into the legitimacy of websites and improving the system's accuracy and robustness. Moreover, the development of real-time monitoring and response mechanisms, such as automated URL scanning and user notification systems, could enable proactive protection against emerging phishing threats. Enhanced model evaluation techniques, including the use of diverse datasets and advanced evaluation metrics, could provide deeper insights into the system's performance and generalization capabilities across different scenarios. Additionally, strategies to educate users about phishing threats and promote awareness of safe browsing practices could complement the technical aspects of the system, empowering users to recognize and avoid phishing attempts. Furthermore, collaboration with industry partners and data sharing initiatives could facilitate the development of more robust and comprehensive phishing detection systems, leveraging collective knowledge and resources to enhance system effectiveness and reliability. Overall, by pursuing these avenues for future work, the proposed approach in phishing detection can continue to evolve and adapt to address emerging challenges, ultimately providing more effective and reliable protection against cyber threats.

REFERENCES

1. N. Z. Harun, N. Jaffar, and P. S. J. Kassim, ‘‘Physical attributes significant in preserving the social sustainability of the traditional malay settlement,’’ in Reframing the Vernacular: Politics, Semiotics, and Representation. Springer, 2020, pp. 225–238.
2. D. M. Divakaran and A. Oest, ‘‘Phishing detection leveraging machine learning and deep learning: A review,’’ 2022, arXiv:2205.07411.
3. A. Akanchha, ‘‘Exploring a robust machine learning classifier for detecting phishing domains using SSL certificates,’’ Fac. Comput. Sci., Dalhousie Univ., Halifax, NS, Canada, Tech. Rep. 10222/78875, 2020.
4. H. Shahriar and S. Nimmagadda, ‘‘Network intrusion detection for TCP/IP packets with machine learning techniques,’’ in Machine Intelligence and Big Data Analytics for Cybersecurity Applications. Cham, Switzerland: Springer, 2020, pp. 231–247.
5. J. Kline, E. Oakes, and P. Barford, ‘‘A URL-based analysis of WWW structure and dynamics,’’ in Proc. Netw. Traffic Meas. Anal. Conf. (TMA), Jun. 2019, p. 800.
6. A. K. Murthy and Suresha, ‘‘XML URL classification based on their semantic structure’’

7. A. A. Ubing, S. Kamilia, A. Abdullah, N. Jhanjhi, and M. Supramaniam,‘‘Phishing website detection: An improved accuracy through feature selec- tion and ensemble learning,’’ Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 1, pp. 252–257, 2019.8.
8. A. Aggarwal, A. Rajadesingan, and P. Kumaraguru, ‘‘PhishAri: Automatic realtime phishing detection on Twitter,’’ in Proc. eCrime Res. Summit, Oct. 2012, pp. 1–12.
9. S. N. Foley, D. Gollmann, and E. Snekkenes, Computer Security—ESORICS 2017, vol. 10492. Oslo, Norway: Springer, Sep. 2017.
10. 10. P. George and P. Vinod, ‘‘Composite email features for spam identification,’’ in Cyber Security. Singapore: Springer, 2018, pp. 281–289.
11. H. S. Hota, A. K. Shrivas, and R. Hota, ‘‘An ensemble model for detecting phishing attack with proposed remove-replace feature selection tech- nique,’’ Proc. Comput. Sci., vol. 132, pp. 900–907, Jan. 2018. G. Sonowal and K. S. Kuppusamy, ‘‘PhiDMA—A phishing detection
12. model with multi-filter approach,’’ J. King Saud Univ., Comput. Inf. Sci., vol. 32, no. 1, pp. 99–112, Jan. 2020.
13. M. Zouina and B. Outtaj, ‘‘A novel lightweight URL phishing detection system using SVM and similarity index,’’ Hum.-Centric Comput. Inf. Sci., vol. 7, no. 1, p. 17, Jun. 2017.
14. R. Ø. Skotnes, ‘‘Management commitment and awareness creation—ICT safety and security in electric power supply network companies,’’ Inf. Comput. Secur., vol. 23, no. 3, pp. 302–316, Jul. 2015.
15. R. Prasad and V. Rohokale, ‘‘Cyber threats and attack overview,’’ in Cyber Security: The Lifeline of Information and Communication Technology. Cham, Switzerland: Springer, 2020, pp. 15–31.
16. T. Nathezhtha, D. Sangeetha, and V. Vaidehi, ‘‘WC- PAD: Web crawling based phishing attack detection,’’ in Proc. Int. Carnahan Conf. Secur. Technol. (ICCST), Oct. 2019, pp. 1–6.
17. R. Jenni and S. Shankar, ‘‘Review of various methods for phishing detection,’’ EAI Endorsed Trans. Energy Web, vol. 5, no. 20, Sep. 2018, Art. no. 155746.
18. (2020). Accessed: Jan. 2020. [Online]. Available: https://catches-of-the-month-phishing-scams-for- january-2020.
19. S. Bell and P. Komisarczuk, ‘‘An analysis of phishing blacklists: Google safe browsing, OpenPhish, and PhishTank,’’ in Proc. Australas. Comput. Sci. Week Multiconf. (ACSW), Melbourne, VIC, Australia. New York, NY, USA: Association for Computing Machinery, 2020, pp. 1– 11, Art. no. 3, doi: 10.1145/3373017.3373020.
20. A. K. Jain and B. Gupta, ‘‘PHISH-SAFE: URL features-based phishing detection system using machine learning,’’ in Cyber Security. Switzerland: Springer, 2018, pp. 467–474.
21. Y. Cao, W. Han, and Y. Le, ‘‘Anti-phishing based on automated individual white-list,’’ in Proc. 4th ACM Workshop Digit. Identity Manage., Oct. 2008, pp. 51–60
22. G. Diksha and J. A. Kumar, ‘‘Mobile phishing attacks and defence mechanisms: State of art and open research challenges,’’ Comput. Secur., vol. 73, pp. 519–544, Mar. 2018.
23. M. Khonji, Y. Iraqi, and A. Jones, ‘‘Phishing detection: A literature survey,’’ IEEE Commun. Surveys Tuts., vol. 15, no. 4, pp. 2091–2121, 4th Quart, 2013.
24. S. Sheng, M. Holbrook, P. Kumaraguru, L. F. Cranor, and J. Downs, ‘‘Who falls for phish? A demographic analysis of phishing susceptibility and effectiveness of interventions,’’ in Proc. SIGCHI Conf. Hum. Factors Comput. Syst., Apr. 2010, pp. 373–382.
25. P. Prakash, M. Kumar, R. R. Kompella, and M. Gupta, ‘‘PhishNet: Predictive blacklisting to detect phishing attacks,’’ in Proc. IEEE INFOCOM, Mar. 2010, pp. 1–5.
26. P. K. Sandhu and S. Singla, ‘‘Google safe browsing- web security,’’ in Proc. IJCSET, vol. 5, 2015, pp. 283–287.
27. M. Sharifi and S. H. Siadati, ‘‘A phishing sites blacklist generator,’’ in Proc. IEEE/ACS Int. Conf. Comput. Syst. Appl., Mar. 2008, pp. 840–843.
28. S. Sheng, B. Wardman, G. Warner, L. Cranor, J. Hong, and C. Zhang, ‘‘An empirical analysis of phishing blacklists,’’ in Proc. 6th Conf. Email Anti- Spam (CEAS), Mountain View, CA, USA. Pittsburgh, PA, USA: Carnegie Mellon Univ., Engineering and Public Policy, Jul. 2009.
29. Y. Zhang, J. I. Hong, and L. F. Cranor, ‘‘Cantina: A content-based approach to detecting phishing web sites,’’ in Proc. 16th Int. Conf. World Wide Web, May 2007, pp. 639–648.
30. G. Xiang, J. Hong, C. P. Rose, and L. Cranor, ‘‘CANTINA+: A feature rich machine learning framework for detecting phishing web sites,’’ ACM Trans. Inf. Syst. Secure., vol. 14, no. 2, pp. 1–28, Sep. 2011.