Educational Administration: Theory and Practice
2024, 30(5), 9969-9975
ISSN: 2148-2403
https://kuey.net/ **Research Article**

Unmasking Phishing Attacks: A Content-Centric Machine Learning Defense

Vidhi Sinha[1], Stuti Garg[2], Latika Sharma[3*], Preeti Gupta[4] ,Dr. Sonia Deshmukh[5]

[1,2,3*,5]Department of CSIT, KIET Group of Institutions, Delhi-NCR, India
[4]Noida Institute of Engineering & Technology, Greater Noida, Plot -19, knowledge Park II, Greater Noida, Uttar Pradesh

***Corresponding Author:** Latika Sharma
*latikasharma@kiet.edu

Citation: Latika Sharma et al. (2024), Unmasking Phishing Attacks: A Content-Centric Machine Learning Defense,*Educational Administration: Theory and Practice*, *30*(5), 9969-9975
Doi: 10.53555/kuey.v30i5.4690

ARTICLE INFO **ABSTRACT**

Nearly all real-world processes have moved to digital platforms in recent years as a result of the Internet's unavoidable growth. Because mobile devices facilitate our connection with connected services at your convenience, there is a surge in the usage of the internet in all facets of our lives. But this inevitable growth also carries with it several security lapses, particularly for regular end users. To make things easier for them, phishing is one of the most popular attack tactics used by hackers. An innocent email or social media message is what initially starts this kind of assault, directing the victims to a malicious website. These attack types are very difficult for security administrators to identify. Consequently, a content-centric Phishing Detection strategy is suggested in this study. The idea aims to identify the optimal training models by implementing around six distinct machine learning models. According to experimental findings, the suggested methods are incredibly reliable and provide security administrators with respectable accuracy.

1 Introduction

Phishing involves deceivers who manipulate individuals into revealing sensitive information, such as usernames, passwords, bank account details, or credit card numbers, by creating deceptive web pages that closely resemble authentic sites. Phishing is not a recent form of online attack; rather, it is a traditional form of attack that attackers continue to employ because it is thought to be one of the most successful ways to access the target's system. Attackers choose phishing because it is simple to execute while yet having the potential to directly contact several online users as a target. This is supported by scientific data, which shows that phishing attack cases are significantly rising year. During the initial six months of 2023, there has been a staggering 464% rise in email-based phishing attacks compared to the previous year, 2022. Concurrently, there is a notable 24% uptick in attacks per organization over this period.

There are three categories of insider threats: intentional, careless, and unintentional. [1] Ever since the initial cyber-attack, millions of attacks of various kinds have been launched daily in every industry. By now, many nations have established defense departments and corresponding cyber-attack plans. The history of cyber-attacks is quite extensive. A vast range of attack techniques have been employed ever since the onset of the initial cyber-attack many years ago as a result of technological advancements. Today's most widely utilized attack types include malware, phishing, XSS, DOS, and DDOS assaults, listening, MitM attacks, birthday, password attacks, XSS attacks, and phishing. Every one of these attacks employs a distinct technological setup and methodology.

Considering the general stages of a phishing attack, Initially, the target audience receives the pre-made attack mail—whose content is tailored to make it attractive to the target user. After that, the unauthenticated sender of this email adds themselves to the list of susceptible users. There is no denying that the message has been expertly edited to appear deceptive. By participating in comparable situations, such as updating, clicking, logging in, or completing a payment, the user interacts with the content as intended by the sender of the message, potentially exposing, and sharing their information. The shared data is promptly assimilated by the attacker's system. To use the information obtained, funds and information are transferred from actual systems. In this manner, the process comes to an end, and in a matter of seconds, every user who interacts with the system falls victim to a straightforward phishing attack. Figure 1 illustrates a comprehensive phishing life cycle.

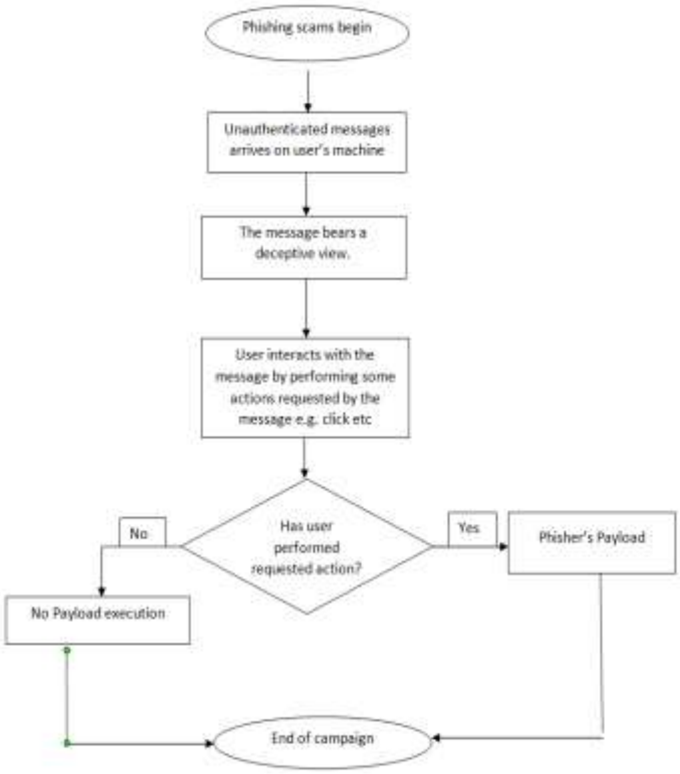

Phishing scams begin
Unauthenticated messages arrives on user's machine
The message bears a deceptive view.
User interacts with the message by performing some actions requested by the message e.g. click etc
Has user performed requested action?
No
Yes
Phisher's Payload
No Payload execution
End of campaign

Fig. 1.

This research will concentrate on safeguarding against phishing attacks that leverage machine learning and artificial intelligence. The goal is to offer a fresh viewpoint and substitute for current systems, all without requiring prior knowledge gained by user or experience. Using the strategy we've created, we will classify a suspect source as dangerous and make sure that precautions must be taken against any kind of possibility of phishing related attacks if it surpasses the limiting value we've set by verifying it with models based on artificial intelligence.

2 Related Work

Numerous anti-phishing defense techniques have been used up to this point. List-based, rule-based, and Intrusion detection systems relying on pattern recognition to identify any similarities have all been implemented in compliance with different scenarios. Each method of use has resulted in particular security weaknesses and problems that are specific to the environment in which it is used.
Let us have a brief look over these methods one-by-one.

2.1 List-based intrusion detection

In these systems, a blacklist or whitelist limits the content of the mail or address that can be accessed via a URL.
The blacklist programme safeguards and restricts access to websites that have been previously identified as phishing attacks.
The fundamental problem is that the system relies on URLs to function. Even a slight modification to the address of the URL can trick the vulnerable control mechanism, resulting in a security vulnerability. Furthermore, because these systems lack experience, it is unable to provide defense against recently created assaults. As long as the system is in operation, efficiency and production clearly diminish as the blacklist gets longer and longer every day.
The reverse application method is utilized in systems with white lists for applications. A restricted set of predetermined URL addresses that can be visited and only these URLs are accessible for universal-purpose, secured, and applications reliant on internal network. The biggest problem with this system is that access problems commonly occur as a result of its rigorous control mechanisms. As a result, there are two problems: access requests and work blocking.

2.2 Rule-based attack detection

If address of a web or it's URL is to be accessed, rule-based systems use simple criteria to assess whether it is a phishing attack. The strategy that has been used here often involves acting in accordance with the outcomes of Boolean type return rules, which have structures approximating IF THEN, OR AND, or Rule-based systems fall under a number of different areas. The first tactic makes use of rule structures based on search engines.

The if condition is employed in this method to determine if the domain address or URL is present in the indices of prominent search engines. The relevant address is marked as an attack if any of the information of URL or domain is missing from the search engine indexes.

2.3 Machine Learning-Based Attack Detection

Machine learning is the most recent and current approach for identifying invasions. The system functions generally in accordance with the findings of estimation based on probability by developing a model after gathering a sizable quantity of attacks and authentic website material, identifying their characteristics, and then querying these models whenever it encounters a new website. The reason for this knowledge's rise to prominence is the fact that it generates results with a significantly greater rate of performance than other systems using the characteristics and information extracted from multiple websites. Nevertheless, the system has made a noteworthy contribution to prioritizing the avoidance of performance issues over time. It can identify even recently generated phishing attacks, adapt and learn through usage, continually enhancing its capabilities, and furnishing more precise results. Constructing models through extraction of features using supervised learning methodologies, which are recognized in the area, it is determined in this study whether model is more successful for such a system. In the near future, it is predicted that this paradigm, which is based on artificial intelligence, would surpass all other systems in effectiveness, outperforming existing approaches across all systems. [2]

In [3, 4], Buber et al. used specific NLP techniques to analyze website URLs and assess whether or not they were counterfeit. Based on the results of their experiments, they attained very high rates of accuracy in their model. But they only concentrate on the address. However, using various URL shorteners makes changing the web page's address straightforward with today's technologies. Therefore, it is vital to use content-based detection.

3Objective

Phishing has developed into a serious cyber security issue, posing significant risks to people, businesses, and even the security of the entire country. These assaults usually make use of deceptive emails, fake websites, or social engineering methods. Phishing techniques have advanced, using tailored messaging, convincing website clones, and sophisticated social engineering techniques to trick naive website users. Effective and efficient methods are urgently needed to identify and prevent phishing attacks.

The primary goal of this research paper is to develop a robust and accurate fraudulent website detection system combining machine learning techniques and webpage content-based features. The major objective is to identify malicious websites that use the detailed information contained in the structure, layout, and content of webpages to deceive users and steal personal information. By looking at a wide range of content-based criteria, including HTML tags, input forms, images, links, and textual elements, the objective is to construct machine learning models that can precisely distinguish between legitimate and phishing websites. The objective of the research is to attain a notable level of precision, sensitivity, and specificity in detecting phishing URLs in order to enhance the security and trustworthiness of online surfing experiences. The results of this study will contribute to the advancement of more sophisticated and dependable systems.

4Methodology

Here is the visual representation of the entire methodology that has been followed in Figure 2.

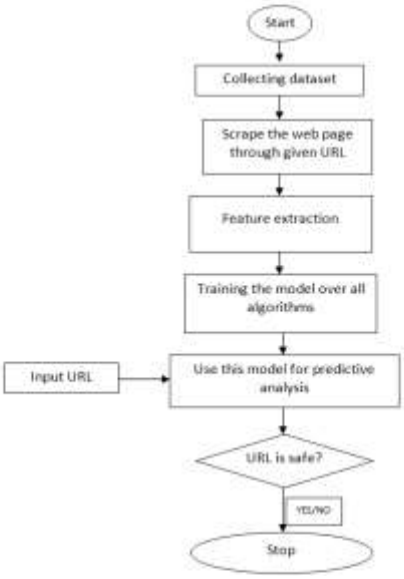

Start
Collecting dataset
Scrape the web page through given URL
Feature extraction
Training the model over all algorithms
Input URL
Use this model for predictive analysis
URL is safe?
YES/NO
Stop

Fig. 2.

4.1 Dataset

For phishing websites, we used the Phistank.com dataset. Phistank.com [5] is a website with many data sources and a variety of tools for spotting phishing scams. Many commercial enterprises and cyber security companies use the information in this to develop defense solutions. Additionally, it provided details on classification rules for either positive or negative data. We acquired a dataset of legitimate URLs from tranco-list.eu. The word "Tranco-list.eu" refers to a public dataset and website that provides information on the most popular websites online. As a part of this project, a comprehensive ranking of websites based on their popularity and traffic will be developed and maintained.

Table 1. Dataset Information

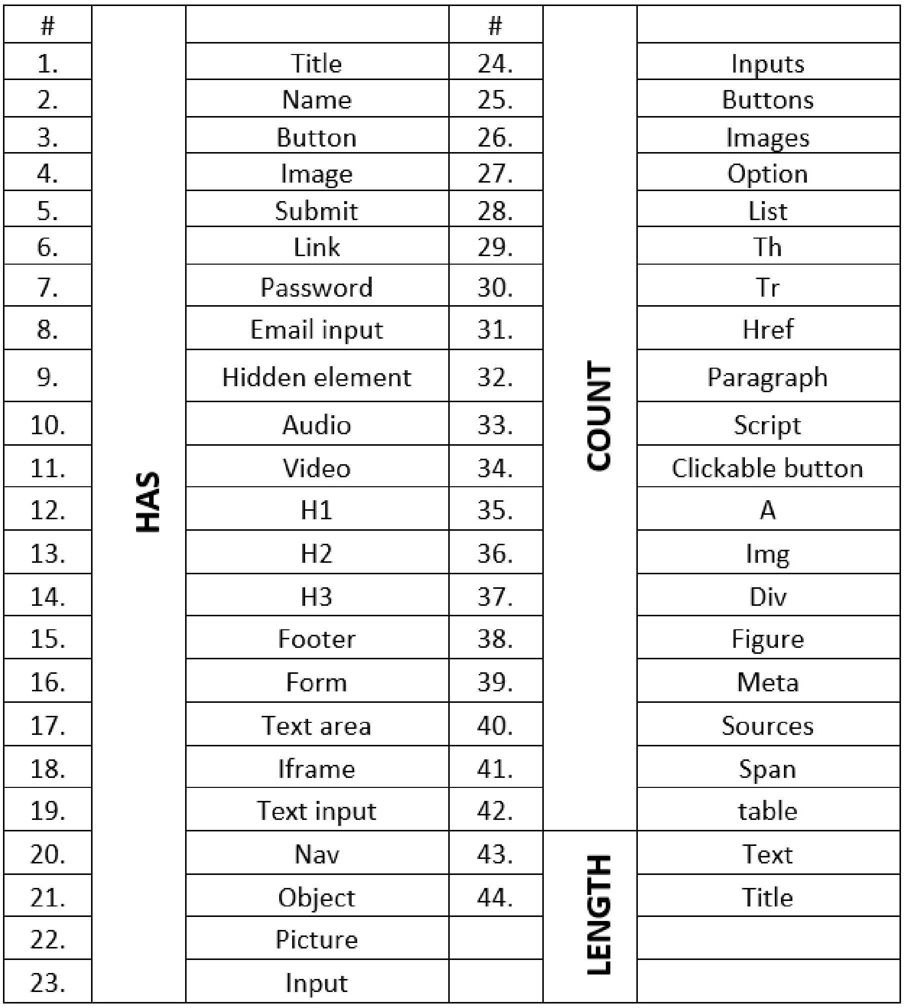
	Dataset
Phishing	8353
Non-Phishing	5438
Total	13791

4.2 Feature Extraction

Then, using a feature extraction technique, pertinent features are derived from the website content. In this process, HTML tags are analyzed while specific elements (such as buttons, photos, and links), text content, and other pertinent elements are sought after. This stage takes into account the elements you mentioned, including the presence of particular HTML tags, input forms, buttons, and images.

We carried out a careful feature identification analysis to identify important traits. By analyzing the contents and source codes of websites and emails, we focused on studying the more thorough investigations conducted in the background. The characteristics of email and websites with JavaScript and HTML coding have been the subject of extensive research. 44 different aspects in this text were found during our analysis.

The several features that are taken into account are-



#			#		
1.	HAS	Title	24.	COUNT	Inputs
2.		Name	25.		Buttons
3.		Button	26.		Images
4.		Image	27.		Option
5.		Submit	28.		List
6.		Link	29.		Th
7.		Password	30.		Tr
8.		Email input	31.		Href
9.		Hidden element	32.		Paragraph
10.		Audio	33.		Script
11.		Video	34.		Clickable button
12.		H1	35.		A
13.		H2	36.		Img
14.		H3	37.		Div
15.		Footer	38.		Figure
16.		Form	39.		Meta
17.		Text area	40.		Sources
18.		Iframe	41.		Span
19.		Text input	42.		table
20.		Nav	43.	LENGTH	Text
21.		Object	44.		Title
22.		Picture			
23.		Input			

4.3 Model training

In the context of phishing website detection utilizing a content-based approach, the learning algorithms part includes the use of several machine learning models to categorize and identify phishing websites according to their content characteristics. Many machine learning methods, including Support Vector Machine (SVM), AdaBoost, Gaussian Naive Bayes, Random Forest, Decision Tree, and K-Nearest Neighbours (KNN), have been used in this study.

Simple yet effective, Gaussian Naive Bayes is a probabilistic classifier. It assumes the traits are independent and have a Gaussian distribution. With regard to phishing website detection, it establishes whether a website is real or phishing based on the likelihood of each feature value given the class designation. Despite its ease of use, Gaussian Naive Bayes can deliver good accuracy and is particularly helpful when working with high-dimensional data.

A common supervised learning algorithm known as SVM divides data points by building hyperplanes in a high-dimensional space. SVM can be trained on content-based features taken from webpages to identify websites as either phishing or authentic in the event of detecting fraudulent websites. In order to separate and categorize webpages, SVM aims to maximize the difference across classes.

The Decision Tree represents decisions and their consequences as a tree-like structure. Every internal node is representative of a feature test, and each branch represents a potential feature value result. Decision Tree algorithms divide the data in recursive fashion according to feature thresholds in order to build a prediction model. Decision Tree can learn general rules to categorize websites based on content-based criteria in the context of phishing website detection.

In order to improve accuracy and decrease over fitting, Random Forest is a method of ensemble learning that uses numerous Decision Trees. A number of Decision Trees are generated, and the results are combined to arrive at a decision. A random subset of the characteristics and training data are used to build each tree. Using the combined knowledge of several Decision Trees, Random Forest can give reliable and accurate categorization in the context of detecting fake websites.

AdaBoost, also known as adaptive boosting, is a technique for ensemble learning that combines weak classifiers to produce a powerful classifier. When doing successive iterations, it gives misclassified samples more weight and assigns weights to each training sample based on how well it performed in the previous round. A more accurate and reliable phishing detection model can be created using AdaBoost by combining many weak classifiers that have been trained on content-based characteristics.

The K-Nearest Neighbours (K-NN) algorithm assigns a sample's class label based on the class labels of its K closest neighbours in the feature space. The choice of K controls how the classification determination is affected by nearby neighbours. By comparing a website's content-based attributes to those of known phishing and genuine websites, K-NN can effectively discriminate between the two classes of websites when it comes to the identification of fraud websites.

Each of these machine learning models has been applied to the issue of identifying phony websites and each one has distinctive capabilities. Each model's performance and effectiveness will be evaluated and compared in light of a number of metrics in order to ascertain how well suited each one is to the detection of content-based phishing attacks.

The model is evaluated on the testing set after it has been trained to determine how well it performs in terms of recall, accuracy, and other pertinent metrics. The usefulness of the content-based technique in precisely identifying phishing websites is evaluated using the results. Comparing the suggested model to current cutting-edge methods for phishing website identification will help to validate it.

5 Experimental Results

In order to determine whether machine learning can considerably increase accuracy, the author has tested six machine learning techniques on the same dataset: SVM, Decision Tree, Random Forest, AdaBoost, KNN, and Gaussian Naive Bayes. The tests were run using the Scikit-Learn library, and the results are listed below. Data for training and testing make up 80% and 20%, respectively. In terms of accuracy, it can be shown that feature selection process has a significant impact and is crucial for spotting this phishing website and this model is able to achieve greater accuracy than previous works in this field. The efficiency comparison table is depicted in Table 2.

Table 2.

ML algorithms	Accuracy	Precision	Recall
Naïve Bayes	0.7151	0.8653	0.8187
SVM	0.8762	0.9165	0.9234
Random Forest	0.9908	0.9798	0.9901
Decision Tree	0.9851	0.9812	0.9898
AdaBoost	0.9784	0.9614	0.9687
K-nearest neighbours	0.9779	0.9592	0.9716

6 Conclusion & Future Scope

The use of computers in all facets of our lives has increased dramatically in the last several years. To gain a larger share of the market, nearly all transactions that occur in the real world are therefore moved to the virtual world. While this advancement simplifies our lives in many ways, it also comes with a host of new issues, chief among them being security-related ones. Because of the Internet's anonymous nature, launching a cyber-attack is not too difficult. Certain basic attacks can even be made by an inexperienced user. Since phishing attacks aim to take advantage of computer users' vulnerabilities, they are among the most popular types of security breaches. This kind of attack cannot be prevented directly by conventional security measures. As a result, certain extra security measures must be created.

We developed a content-driven phishing detection strategy for this project, which examines the text and other features of the website to determine whether or not it is fraudulent. In order to comprehend and contrast the suggested models, we employed six distinct machine learning models in this method. The suggested model produces an efficient and acceptable security level for typical end users, according to experimental results.

We intend to use hybrid models in our ongoing research to try and improve the system's efficiency while working on more complex cases. The intention behind future model is to incorporate novel models, such as deep learning models. Improvements in results are anticipated, particularly as dataset sizes grow.

References

1. Types of Cyber Attackers, https://www.javatpoint.com/typesof-cyber-attackers.(2018)
2. Miyamoto, D., Hazeyama, H., & Kadobayashi, Y,An evaluation of machine learning-based methods for detection of phishing sites. In Advances in Neuro-Information Processing: 15th International Conference, ICONIP 2008, Auckland, New Zealand, November 25-28, 2008, Revised Selected Papers, Part I 15 (pp. 539-546). Springer Berlin Heidelberg(2009).
3. Buber, E., Diri, B., & Sahingoz, O. K. NLP based phishing attack detection from URLs. In Intelligent Systems Design and Applications: 17th International Conference on Intelligent Systems Design and Applications (ISDA 2017) held in Delhi, India, December 14-16, 2017 (pp. 608-618). springer international Publishing. (2018)
4. Buber, E., Dırı, B., & Sahingoz, O. K. Detecting phishing attacks from URL by using NLP techniques. In 2017 International conference on computer science and Engineering (UBMK) (pp. 337-342). IEEE. (2017, October)
5. Phistank, www.phishtank.com, 31 08 2020,
6. Ozker, U., & Sahingoz, O. K. Content based phishing detection with machine learning. In 2020 International Conference on Electrical Engineering (ICEE) (pp. 1-6). IEEE(2020, September).
7. Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. Machine learning based phishing detection from URLs. Expert Systems with Applications, 117, 345-357. (2019).
8. Wardman, B., Stallings, T., Warner, G., & Skjellum, A. High-performance content-based phishing attack detection. In 2011 eCrime Researchers Summit (pp. 1-9). IEEE. . (2011, November)
9. Korkmaz, M., Sahingoz, O. K., & Diri, B. Feature selections for the classification of webpages to detect phishing attacks: a survey. In 2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA) (pp. 1-9). IEEE. (2020, June).
10. Jain, A. K., Parashar, S., Katare, P., & Sharma, I. Phishskape: A content based approach to escape phishing attacks. Procedia Computer Science, 171, 1102-1109(2020)
11. Maktabar, M., Zainal, A., Maarof, M. A., & Kassim, M. N. Content based fraudulent website detection using supervised machine learning techniques. In Hybrid Intelligent Systems: 17th International Conference on Hybrid Intelligent Systems (HIS 2017) held in Delhi, India, December 14-16, 2017 (pp. 294-304). Springer International Publishing(2018).
12. Hannousse, A., & Yahiouche, S. Towards benchmark datasets for machine learning based website phishing detection: An experimental study. Engineering Applications of Artificial Intelligence, 104, 104347(2021).
13. Wu, C. Y., Kuo, C. C., & Yang, C. S. A phishing detection system based on machine learning. In 2019 International Conference on Intelligent Computing and its Emerging Applications (ICEA) (pp. 28-32). IEEE(2019, August).
14. Safi, A., & Singh, S. A systematic literature review on phishing website detection techniques. Journal of King Saud University-Computer and Information Sciences(2023).
15. Bai, W. Phishing website detection based on machine learning algorithm. In 2020 International Conference on Computing and Data Science (CDS) (pp. 293-298). IEEE(2020, August).
16. Buber, E., Demir, Ö., & Sahingoz, O. K. Feature selections for the machine learning based detection of phishing websites. In 2017 international artificial intelligence and data processing symposium (IDAP) (pp. 1-5). Ieee(2017, September).
17. Miyamoto, D., Hazeyama, H., & Kadobayashi, Y. An evaluation of machine learning-based methods for detection of phishing sites. In Advances in Neuro-Information Processing: 15th International Conference, ICONIP 2008, Auckland, New Zealand, November 25-28, 2008, Revised Selected Papers, Part I 15 (pp. 539-546). Springer Berlin Heidelberg(2009).
18. Bai, W. Phishing website detection based on machine learning algorithm. In 2020 International Conference on Computing and Data Science (CDS) (pp. 293-298). IEEE(2020, August).

19. Jain, A. K., Parashar, S., Katare, P., & Sharma, I. Phishskape: A content based approach to escape phishing attacks. Procedia Computer Science, 171, 1102-1109(2020).
20. Hannousse, A., & Yahiouche, S. Towards benchmark datasets for machine learning based website phishing detection: An experimental study. Engineering Applications of Artificial Intelligence, 104, 104347(2021).
21. Buber, E., Demir, Ö., & Sahingoz, O. K. Feature selections for the machine learning based detection of phishing websites. In 2017 international artificial intelligence and data processing symposium (IDAP) (pp. 1-5). Ieee(2017, September).
22. Korkmaz, M., Sahingoz, O. K., & Diri, BDetection of phishing websites by using machine learning-based URL analysis. In 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-7). IEEE. (2020, July).
23. Buber, E., Dırı, B., & Sahingoz, O. K. Detecting phishing attacks from URL by using NLP techniques. In 2017 International conference on computer science and Engineering (UBMK) (pp. 337-342). IEEE(2017, October).
24. Karatas, G., & Sahingoz, O. K. Neural network based intrusion detection systems with different training functions. In 2018 6th International Symposium on Digital Forensic and Security (ISDFS) (pp. 1-6). IEEE(2018, March).
25. Akinyelu, A. A., & Adewumi, A. O. Classification of phishing email using random forest machine learning technique. Journal of Applied Mathematics, 2014(2014).
26. Abawajy, J.User preference of cyber security awareness delivery methods. Behaviour & Information Technology, 33(3), 237-248. (2014).
27. El Aassal, A., Baki, S., Das, A., & Verma, R. M. An in-depth benchmarking and evaluation of phishing detection research for security needs. IEEE Access, 8, 22170-22192(2020).
28. Ramkumar, N., Kothari, V., Mills, C., Koppel, R., Blythe, J., Smith, S., & Kun, A. L. Eyes on URLs: Relating visual behavior to safety decisions. In ACM symposium on eye tracking research and applications (pp. 1-10) (2020, June).
29. Shirazi, H., Bezawada, B., Ray, I., & Anderson, C.Directed adversarial sampling attacks on phishing detection. Journal of Computer Security, 29(1), 1-23. (2021).
30. Ejaz, A., Mian, A. N., & Manzoor, SLife-long phishing attack detection using continual learning. Scientific Reports, 13(1), 11488. (2023).