THE INTERNATIONAL JOURNAL OF SCIENCE & TECHNOLEDGE

Phishing URL Detection: A Basic Machine Learning Approach

Dr. Onyiagha Chyke Godfrey
Professor, Faculty of Engineering (Ground & Communications),
Air Force Institute of Technology, Kaduna, Nigeria
Yanwalo George Fwa
Former Student, Department of Cybersecurity,
Air Force Institute of Technology, Kaduna, Nigeria
Ajimah Nnabueze Edmund
Head, Department of Electrical & Electronic Engineering,
Air Force Institute of Technology, Kaduna, Nigeria

Abstract:
Phishing is currently one of the most trending scam attacks in information transmission worldwide. Such occurrences can cause severe network disruption to corporate organizations, financial and/or educational institutions and even individual network subscribers. Phishing attacks would normally involve developing a malicious webpage mimicking a legitimate website. This lures unsuspecting users into giving out personal information, such as banking details, social security numbers, passwords, usernames, etc. A phishing fraudster sends an e-mail or text message to a target webpage that contains its Universal Resource Locator (URL). The result could be monetary loss, data breach, intellectual property theft, damaged reputation, and loss of customers. In this paper, we propose a simple phishing detection approach. Moreover, because the impostor is always evolving attacking techniques in a bid to evade detection, our method addresses this using a supervised machine learning system. From the website content, we extract the stochastic dynamical patterns and then use this to predict the authenticity of the website. The Extreme Gradient Boosting (XGBoost) algorithm is used to model the website features to obtain a better prediction result. The proposed technique can detect phishing websites with an accuracy of 86.6%.

Keywords: *Phishing URL, machine learning, XGBoost, phishing attacks, artificial intelligence, fraud detection bypass*

1. Introduction

Since its initial incidence in 1995, when some American Online (AOL) users discovered a technique to alter their screen names and make it appear as though it is from an AOL administrator, phishing fraudulent attacks have been a global issue in the internet community. They would "phish" for login credentials using such dubious screen names in order to gain free access to the Internet [1]. Phishing attacks are a common kind of social engineering scam that takes advantage of people's weaknesses to steal sensitive personal data.

Artificial intelligence is used by today's very high-speed networks, like 5G, to enhance the quality of service provided to a subscriber because of the unquenchable demand for more and more data per unit of time [2]. Despite this, phishing scammers are always improving their methods of attack. To get around or evade detection, for instance, they now employ Artificial Intelligence (AI) techniques and Large Language Models (LLM). According to recent studies, targeted phishing assaults are on the increase [2]. Of these, 88% are expected to deal with spear-phishing assaults, 83% with voice phishing (also known as vishing), 86% with social media attacks, 84% with SMS/text phishing (also known as SMishing), whilst 81% deal with malicious USB drops. According to a 2018 Proofpoint annual report, the frequency of all forms of phishing assaults increased from 76% in 2017 to 83% in 2018. Similarly, a report from the Anti-Phishing Working Group (APWG2) [3] stated that the number of phishing attacks detected in the second quarter of 2019 was significantly higher than the number recorded in the previous three quarters. This indicates that phishing attacks are becoming more common. These results have shown a clear picture of how phishing attacks have become more sophisticated and how they maliciously affect businesses and individuals with increasing severity.

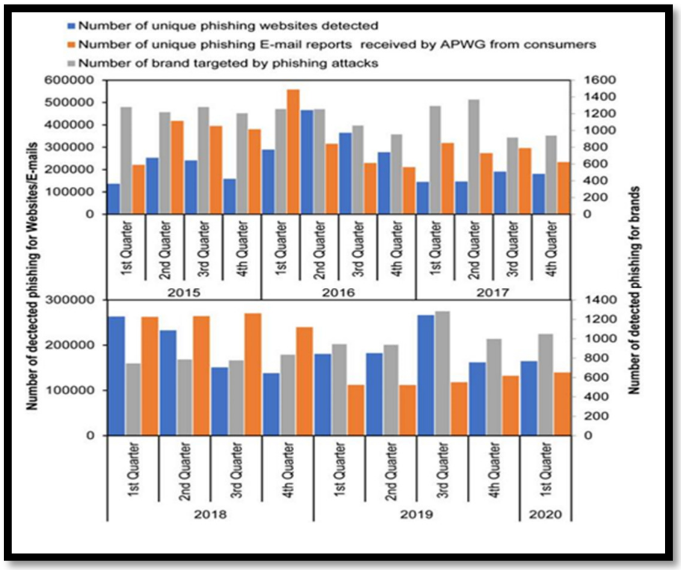


Figure 1: The Growth in Phishing Attacks 2015–2020 by Quarters Based on Data Collected from APWG Annual Reports. (Apwg.Org)

Since phishing attacks take advantage of a user's incapacity to confirm the legitimacy of a URL, it is possible to reduce the vulnerability of the user by developing and utilizing intelligent machine models. The data pattern distribution can be used to characterize content as malicious or acceptable if features of a website URL can be extracted. The majority of the time, users lack the technical knowledge necessary to recognize phishing URLs or distinguish dangerous URLs from legitimate ones. As a result, several ensemble machine learning algorithms are trained on diverse URL transaction datasets to increase their agility in detecting fraud.

2. Detection Challenges

The traditional approach to detecting phishing scams relies on detection blacklisting. The user report and the updated blacklisted websites are the main ways that this detection technology maintains its complexity. This method compares a website to a list of websites that have been blacklisted, such that if the website includes a pattern from the blacklist, it is deemed bogus; otherwise, it is deemed legitimate. However, the number of rogue websites is growing exponentially, making it intractable to identify new malicious URLs using that method. This allows 'zero-day' phishing assaults to spread. A user's report and the manual updating of the blacklist are other factors that determine how effective this strategy is. In this paper, our approach uses the supervised Machine Learning algorithm to achieve a more accurate solution, better than the blacklist approach.

3. Literature Review

3.1. Phishing Detection Using Extra Trees Classifier

According to a study by Arathi *et al.* [3], phishing is a type of attack in which the attackers try to deceive the victim into clicking on phishing links in order to steal important information like usernames or passwords. These connections might already have anti-phishing software and computational techniques in place for actively identifying phishing activity. However, they might not be able to identify phishing attacks that are changing with time. Using a machine learning technique, the researchers created a web-based tool to identify phishing URLs. In such an approach, they employed a feature extraction algorithm with emphasis on address bar-based features, abnormal-based features, HTML JavaScript-based features and Domain-based features. In assessing the accuracy and performance of their proposed model, two ensemble classifiers, random forest (RF) and Extra Trees (ET) were compared to find the one with a better performance procedure. The models were trained using a dataset from the University of California Irvine (UCI) Machine Learning Repository, with 30 features. Hyper-parameter tuning was performed on the models to check whether their predictive performance improved. The Extra Trees classifier without the tuning achieved the highest accuracy of 97.47% on the test dataset with the least false positive rate.

3.2. Phishing Detection Using Supervised Machine Learning.

In order to prevent the propagation of phishing attacks, Lakshmi *et al.* [4] proposed a model for detecting phishing websites using a supervised deep-learning algorithm. The model uses Deep Neural Networks (DNN) which can process data of the phishing URL on its own without any supervision. In this algorithmic approach, feature sets from the websites are analyzed using deep neural networks that can detect whether a website has been phished or not. This works by sending the results of the hidden layer to the next layer for processing. At the beginning, the input layer receives its input from the training and test dataset. The data is then processed without external help. Next, the hyper-planes information calculated from the hidden layers are passed over to the output layer.

The model uses thirty (30) well-defined features for training the deep neural networks, with 2 or 3 hidden layers to effectively determine if the URL is legitimate. This proposed model attained an accuracy of up to 90%.

In Vaneeta *et al*. [5], an intelligent phishing detection system based on machine learning classifiers with a wrapper features selection method was presented. The wrapper features selection method is based on a greedy search algorithm. This evaluates all possible feature combinations and then selects the combination that gives the best results for a specific Machine Learning (ML) algorithm. Datasets were collected from KAGGLE proprietary websites and the UCI to train the model. The work compared the accuracy scores of the following three algorithms: Neural Network with an accuracy of 95.296%, random forest with an accuracy of 97.744% and Support Vector Machine (SVM) with an accuracy of 94.00%. The wrapper feature selection proves to be more accurate in its prediction than the model without wrapper feature selection [6].

3.3. Overview of Phishing Attacks

Banik and Sarma [7] proposed an article with a detailed anatomy of phishing, which involves the attack phase, the attacker's types, vulnerabilities, threats, targets, the attack medium, and the attack techniques [7]. The proposed anatomy is targeted at aiding researchers in gaining a better understanding of the lifecycle of a phishing attack, which in turn will increase awareness of these phishing attacks and enhance the techniques being used to develop anti-phishing systems. The researchers highlighted the following reasons for human susceptibility to phishing attacks. Everyone is susceptible to phishing because the phisher plays on an individual's specific psychological and emotional fears and technical vulnerabilities [8]. Furthermore, curiosity and urgency were noted as the most common triggers that encourage individuals to respond to phishing attacks.

3.4. Phishing in E-commerce

E-Commerce has been plagued with problems since its inception and this research examines one of these problems: The lack of user trust in E-Commerce generally emanates from associated risks of phishing. Phishing has grown exponentially following the expansion of the Internet. This growth and the advancement of technology have not only benefitted honest internet users but have also enabled cyber-criminals to increase their attacking prowess. This has caused reasonable damage to this budding area of E-commerce. Furthermore, phishing has negatively impacted both the user and online businesses by breaking down the trust relationship between them. In an attempt to explore this problem, the susceptibility of phishing attacks to e-commerce has been examined. First, the Common Criteria Security Model was used to identify the key security areas as well as the weak points and vulnerable regions in e-commerce. Second, the strategies and tactics employed in phishing, including phishing e-mails, websites, and addresses, disseminated attacks, redirected attacks, and the information that phishers aim to acquire, have been scrutinized. Additionally, a method has been developed to lower the risk of phishing, which will enhance consumer trust in websites. The significance of trust, the Uncertainty Reduction Theory (URT), and the delicate equilibrium between control and trust have all been discussed in [9]. Finally, the study presented Critical Success Factors that aid in phishing prevention and control. These include User Authentication, Website Authentication, E-mail Authentication, Data Cryptography, Communication, and Active Risk Mitigation.

4. Architecture of the Proposed System Design

The architectural design of the proposed system follows the process of the user's engagement with the system by inputting a URL into the designed system. The URL inputted by the user is stored and processed by the classification algorithm. Features from the URL are extracted and analyzed. The model performs these processes to determine the legitimacy of the URL. Figure 2 is a block diagram of the architecture of the proposed system.

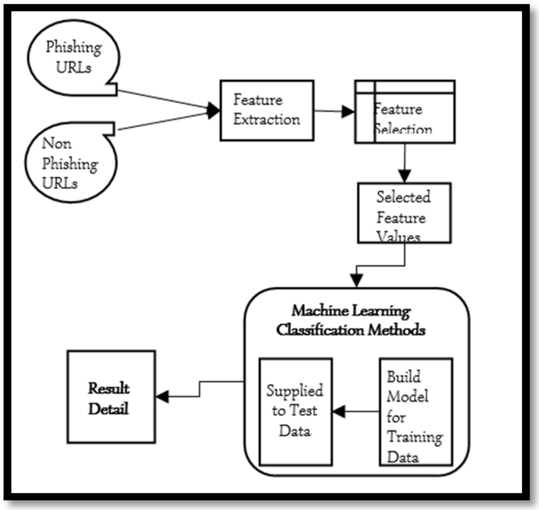


Figure 2: Methodology Architecture of the Proposed System

5. Methodology

This section outlines the procedures and techniques adopted to achieve the set goal of the research; some of the steps to achieve this goal are enlisted as follows:

- To execute this research work, datasets that are a collection of random phishing and legitimate URLs are used. The datasets are collected from two open-source databases called Phish Tank (for the phishing URLs) and open datasets of the University of New Brunswick (for the legitimate URLs).
- An Exploratory Data Analysis (EDA) is conducted on the collected data to fine-tune and eliminate unnecessary features from the datasets. This process brings out the salient features from the datasets that will yield the best result.
- Feature selection and feature ranking are conducted based on the pattern distribution properties of the Uniform Resource Locator (URL).
- Training the model using the Extreme Gradient Boost Tree (XGBOOST) classifier algorithm and thus understanding the bias of the model.
- Performance evaluation is carried out on the model to test out the algorithm's accuracy using performance metrics.

5.1. Feature Selection

The technique for feature selection is performed by selecting only important/salient features that will be useful in the training phase of the learning model. Since this approach uses the URL properties for its detection, webpage features will not be considered. The ability to classify a URL feature correctly is shown in their rank. The length of the URL amongst all features shows the highest rank [10].

An attacker would use a URL of a larger length to hide a suspicious path in the URL [11]. The feature with the second highest rank is that of a URL with a greater number of symbols. Characters which are rarely observed are mostly present in phishing URLs.

5.2. Feature Analysis of Dataset

Figure 3 gives Pattern Distribution Plot of Dataset based on selected features (generated by pandas) which gives an insight about the features in the datasets to enhance the dataset analysis.

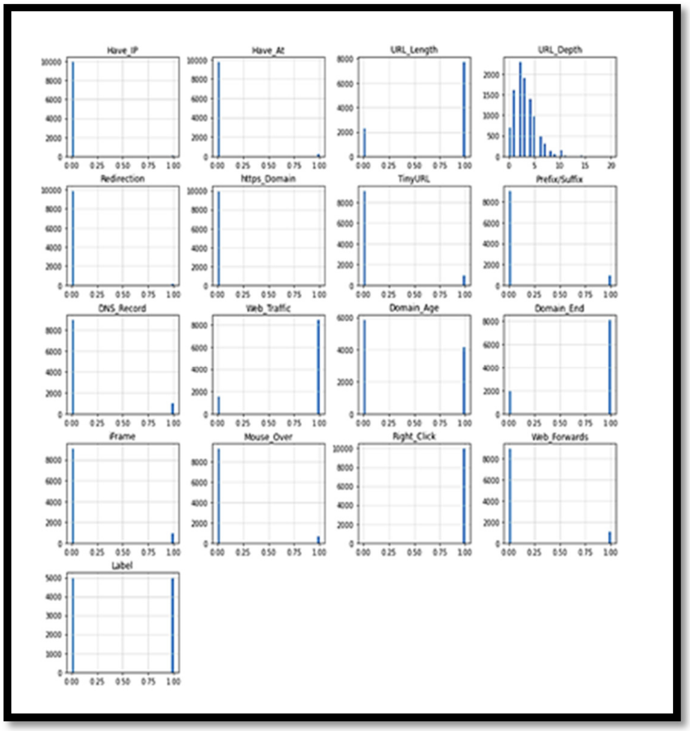


Figure 3: Pattern Distribution Plot of Dataset Based on Selected Features (Generated by Pandas)

The above graphs show a plot of the distribution pattern of authentic phishing datasets based on chosen features and their relationships to one another.

The data analysis made it clear that most phishing websites are lengthy and comprise symbols. Most phishing websites contain a small link depth, which classifies them as static pages, and they do not follow the same design principles as the authentic page.

The plot of a correlation heat map of the dataset shown in figure 4 illustrates the relationships between the various variables of the dataset.

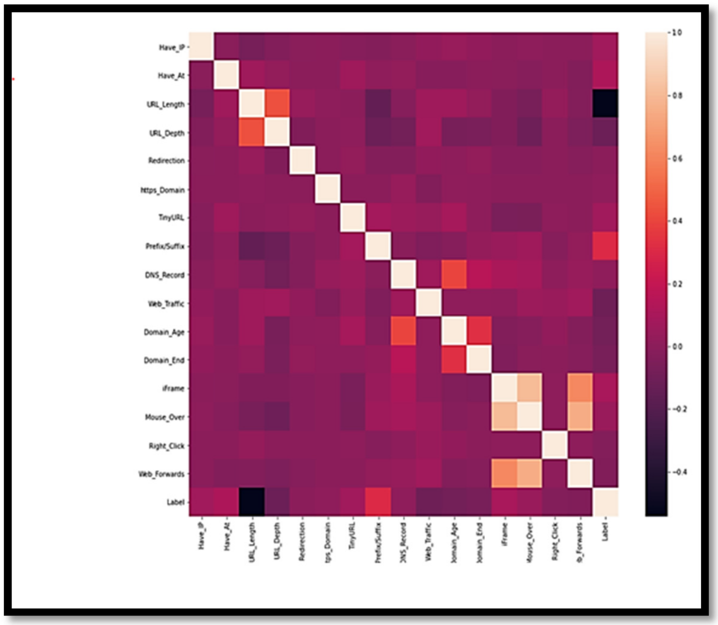


Figure 4: Correlation Heat Map of the Dataset (Generated by Pandas)

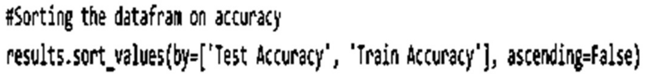
5.3. Dataset and Environment

The data used to generate the datasets on which the models were trained and tested were obtained from different open-source platforms. The dataset collection consists of phishing and legitimate URL datasets. As mentioned earlier, the set of phishing URLs was collected from an open-source service called Phish Tank. This service provides a set of phishing URLs in multiple formats, such as CSV, JSON, etc.

The collected URL information is updated hourly. This dataset is accessible from the "phishtank.com" website. Over 5000 random phishing URLs were collected to train and test the ML models. The set of legitimate URLs was obtained from the open datasets of the University of New Brunswick. This dataset is accessible from the university website. The dataset variables come from a collection of benign spam, phishing, malware and defacement URLs. From this spread, over 5000 random legitimate URLs were used to train the ML models.

5.4. Model Training and Testing

Training the Machine Learning models involves feeding the algorithms with data to help identify and learn good attributes of the dataset. This research aims to find a solution to a classification problem which falls under supervised machine learning. The algorithms used for phishing detection consist of supervised machine learning models to pre-process the impinging dataset, followed by a deep learning neural network, which was used to train the dataset. Model testing involves the method whereby the performance of a fully trained model is evaluated on a testing dataset. Therefore, after 80% of the data has been trained, 20% of the dataset is used to evaluate the trained dataset to appraise the performance of the models [21]. The Python code excerpt is shown as follows:



```
#Sorting the datafram on accuracy
results.sort_values(by=['Test Accuracy', 'Train Accuracy'], ascending=False)
```

	ML Model	Train Accuracy	Test Accuracy
3	XGBoost	0.866	0.864
2	Multilayer Perceptrons	0.865	0.864
0	Decision Trees	0.814	0.812
1	Random Forest	0.818	0.811
5	Support Vector Machine (SVM)	0.804	0.794
4	AutoEncoder	0.161	0.177

Table 1: Accuracy of the Models as Test and Training Data

6. Results

According to the approach used to develop the system, deep learning neural networks and machine learning models have been used. Examples of such models include Random Forests, Multilayer Perceptions, Auto Encoder Neural Networks, Support Vector Machines, Decision Trees, and Support Vector Machines with XGBoost [21]. The models determine whether a website URL is legitimate or a phishing type. A binary-class forecast (legal web-address = 0 and phishing web-address = 1) is provided by the models. More than six machine learning models and deep neural network algorithms were combined in the model development process. These helped to identify phishing URLs. The platform was Jupiter Notebook IDE with packages such as pandas, urllib, etc. The accuracy of the models was tested using sklearn matrices with accuracy scores shown in table 1. The XGBoost model had the highest performance score of 86.6%, the Multilayer Perceptions model had an accuracy of 86.5%, the Decision Tree model had an accuracy of 81.4%, the Random Forest model had an accuracy of 81.8%, the Support Vector Machine model had an accuracy of 80.4%, and the Auto Encoder Neural Network model had an accuracy of 16.1%.

7. Discussion

Phishing attacks cost Internet users billions of dollars every year and are a growing hazard and constant menace when it comes to security in cyberspace. Phishing incorporates a variety of sophisticated social engineering techniques to get sensitive information from users. As a result, phishing tactics can be propagated using a range of communication channels, such as e-mail, instant messaging, pop-up windows [21], and web pages. This project was able to classify and identify the various methods by which researchers have contributed towards the solution of phishing detection. In order to identify patterns in which URL links can be easily detected, the proposed system of this project employed different feature selection, machine learning, and deep neural network techniques, namely: Decision Tree, Support Vector Machine, XGBoost, Multilayer Perceptions, Auto Encoder Neural Network, and Random Forest. Users can enter website URL links to determine whether they are legitimate or phishing by using a web application that integrates these models. The highest accuracy is obtained based on the feature extraction algorithm used to distinguish phishing URLs from legitimate URL links.

8. References

i. Proofpoint. (2020). *the growth in phishing attacks, 2015–2020*. Retrieved from: https://www.proofpoint.com/sites/default/files/2020-06/pfpt-es-state-of-the-phish-2020-reports-a4.pdf

ii. Onyiagha, G., Anabi, H., Mbwahnche, R., & Achimugu, P. (2022). On the concept of artificial intelligence for quality of service management in 5G. *Academia Letters, Article: 4850*. Retrieved from: https://www.academia.edu/73269355/On_Concept_of_Artificial_Intelligence_for_Quality_of_Service_Management_in_5G_Network

iii. Anti-Phishing Working Group (APWG). (2018). *APWG trends report Q2 2018*. Retrieved from: https://docs.apwg.org/reports/apwg_trends_report_q2_2018.pdf

iv. Arathi Krishna V., Anusree A., Blessy Jose, Karthika Anilkumar, & Ojus Thomas Lee. (2021). Phishing detection using machine learning based URL analysis: A survey. *International Journal of Engineering Research & Technology (IJERT) NCREIS – 2021, 9*(13).

v. Lakshmi, L., Pushpa Rani, K., Vijay, K., & Priyanka, GVSS (2019). Phishing within e-commerce: A trust and confidence game. *International Journal of Recent Technology and Engineering (IJRTE)*, DOI:10.35940/ijrte.

vi. Mrs. Vaneeta M., Pratik N.N., Prajwal D., & Pradeep K.S. (2020). Phishing trend report. *International Journal of Emerging Technologies and Innovative Research, 7*(6), 117-123. Retrieved from: http://www.jetir.org/papers/JETIR2006018.pdf

vii. Banik, B., & Sarma, A. (2018). Phishing URL detection system based on URL features using SVM. *International Journal of Electronics and Applied Research, 5*, 40–55.

viii. Megaw, G., & Flowerday, S. (2010). Phishing within E-commerce: A trust and confidence game. In *Proceedings of the 2010 Information Security for South Africa Conference (ISSA 2010)* (pp. 1–8). DOI:10.1109/ISSA.2010.5588333.

ix. AK, & Gupta, B.B. (2017). Phishing detection: Analysis of visual similarity-based approaches. *Security and Communication Networks, 2017*. DOI:10.1155/2017/5421046

x. Jeeva, S.C., & Rajsingh, E.B. (2016). Intelligent phishing URL detection using association rule mining. *Human-Centric Computing & Information Sciences, 6*(10), 7–19.

xi. Dutta, A.K. (2021). Detecting phishing websites using machine learning techniques. *PLoS ONE, 16*(10), e0258361. DOI: 10.1371/journal.pone.0258361

xii. Abinadi, A., Akanbi, O., & Zainal, A. (2013). Feature extraction process: A phishing detection approach. In *13th International Conference on Intelligent Systems Design and Applications in Communications and Network Security (CNS)* (pp. 331-335). IEEE.

xiii. Jain, A.K., & Gupta, B.B. (2018). A machine learning-based approach for phishing detection using hyperlinks information. *Journal of Ambient Intelligence and Humanized Computing*. DOI:10.1007/s12652-018-0798-z

xiv. Le, H., Pham, Q., Sahoo, D., & Hoi, SCH (2018). Urlnet: Learning a URL representation with deep learning for malicious URL detection. *arXiv*. arXiv: 1802.03162.

xv. Rao, R.S., Vaishnavi, T., & Pais, A.R. (2019). CatchPhish: Detection of phishing websites by inspecting URLs. *Journal of Ambient Intelligence and Humanized Computing, 11*, 813–825.

xvi. James, J., L., Sandhya, & Thomas, C. (2013). Detection of phishing URLs using machine learning techniques. *ICCC*, 304–309. DOI:10.1109/ICCC.2013.6731669.

xvii. Elakya, M., Mohan, M., Kishore, V., Keerthivasan, M., & Solanki, M. (2019). Phishing scam detection using machine learning. *International Journal of Engineering and Advanced Technology, 9*, 114–118. DOI:10.35940/ijeat.A1023.1091S19.

xviii. Mehndiratta, M., Jain, N., Malhotra, A., Gupta, I., & Narula, R. (2023). Malicious URL: Analysis and detection using machine learning.

xix. Sapate, S. (2023). ML algorithm for detection of phishing sites. *International Journal of Innovative Research in Computer and Communication Engineering, 11*, 2086. DOI:10.15680/IJIRCCE.2023.1104064.

xx. Shreya Gopal Sundari. (2022). Phishing-website-detection-by-machine-learning-techniques. *Journal of Engineering Sciences, 13*(8). Retrieved from: https://jespublication.com/upload/2022-V13I8043.pdf.