Detection of Phishing Website Using Machine Learning and Features Extraction

Amaefule I.A[1], Ubochi C.I[2], Anamelechi F.C[3]

[1,2&3]Department of Computer Science, Imo State University, Owerri, Imo State Nigeria.

ABSTRACT

Phishing assaults are fast-expanding menace in cyberspace which costs web users and organizations huge financial loss annually. Sensitive information obtained from customers via various social engineering methods is prohibited, including Web sites, pop-up messages, instant messaging, email, and other communication channels can all be utilized to spot phishing attempts. This paper provides a framework that can identify phishing or real URL links. A collection of harmless, junk mail, malware, phishing, and defacement URLs are included in the data set employed for categorization. Additionally, phishing URLs from an open-source platform called "Phish Tank," which offers phishing URLs in various forms like JSON, CSV, and others, are included. Six (6) models of machine learning and deep neural network techniques are used to identify phishing URLs. With a collection of over 10,000 randomly chosen URLs, split into 60% training and 40% testing samples, and comprising up to 23,328 phishing and 4894 valid URLs, the research purpose is the development of online applications that can quickly recognize phishing URLs. The Uniform Resource Locator datasets has been trained and evaluated utilizing feature selections such as HTTPS & JavaScript-based features, domain-based features, address bar-based features in order to differentiate among legitimate and phishing URLs. The research provided a method for classifying URLs into legitimate and fraudulent URLs. In order to help individuals and organizations spot phishing links and stay one step ahead of the criminal, it would be very beneficial to authenticate each link that is delivered to them in order to verify its credibility.

Keywords: *Phishing, Detection, Machine Learning, Neural Network, Authentication, Identification*

I. INTRODUCTION

Our daily activities now heavily rely on online platforms, especially social media, for data gathering and dissemination. [1] Assertions that the global web is a computer connection that contains significant information. Many security mechanisms have been implemented to protect that data, but individuals are still a vulnerability. When users effortlessly give up their personal information or computer access, security mechanisms work much harder to safeguard their gadgets and information safe.

consequently, [2] among the most prevalent forms of attacks involving social engineering is social engineering, which is a form of attack intended to obtain user information, including debit and credit card information and login credentials. An attack happens when a perpetrator tricks his target into opening a quick text, chat message, or emails that seems to be of an authorized party. Upon clicking the link, the recipient mistakenly thinks they have received a gift and unintentionally hits a malicious link that can cause malware to be installed, system to freeze during a ransomware attack, or personal information to be exposed.

Cyber security worries have risen significantly in the past few decades owing to the swift implementation of technology breakthroughs, making people more vulnerable to abuse by others. individuals need to be aware of the strategies employed by hackers and how to prevent being a target of phishing attacks.

The strategies used by fraudsters get increasingly complex as technology advances. Apart from phishing, there are additional methods to obtain the private data of clients. [3] states that the following techniques are applicable:

a. Vishing, sometimes referred to as voice phishing, is when a phisher calls a target in an attempt to get private data about their bank account. The most common technique for phone phishing is the deployment of a phony caller ID.

b. Smishing (SMS Phishing): act of exploiting the Short Message Service (SMS) to deliver fraudulent messages. This approach involves using the SMS messaging services to lure a victim by sending them a link to a phishing site.

c. Ransomware: This type of assault prevents users from accessing their computers or information until they make a monetary ransom.

d. Malvertising: This type of malicious advertising uses active scripting to infect your computer with malware or pushes unsolicited content. Malicious advertising most frequently used are vulnerabilities for Flash and Adobe PDF.

As a result, this presents an increasing risk to individuals and enterprises of all sizes. There are greater numbers of phishing URLs and email distributed worldwide now, attackers have gained entry to industrial-strength solutions on the dark web, and, extremely alarmingly, they are becoming increasingly sophisticated and challenging to identify. The second-highest number of phishing assaults ever recorded happened in the first quarter of 2022, with payment systems being the most frequently hit sector. In the third quarter of 2021, 123,972 different phishing attacks were observed [4].

A. Phishing mechanism

Phishing websites are copies of legitimate websites that are currently aimed at, and they usually contain form inputs (textboxes, for example). The intruder gets the victim's details when they upload it. An intruder may obtain credentials for an individual from a non-target by doing the following:

i. Creating a website for phishing. In the initial stage, the criminal identifies the victim as a reputable business. Then, in order to obtain comprehensive details regarding the business, the intruder goes to its website. Then, employing these details, the intruder builds the fake website.

ii. Sending URLs: the hacker drafts a phony email and sends it to a lot of people. The criminal provided the URL of the fraudulent website in the spoof email. A spear-phishing assault targets a particular user category with the email. Furthermore, hackers may circulate the URL of fake websites via forum posts, blogs, and other social media platforms [5].

iii. Credential theft: A fake webpage opens in the individual's browser once they tap on the linked URL. An unwary guests' details are stolen via a fake signup form on the fraudulent website. The details entered by the individual is also accessible to the hacker.

iv. Identity Theft: The details are misused by the criminal. For example, a fraudster might employ an individual's credit card details to complete a purchase. [6]

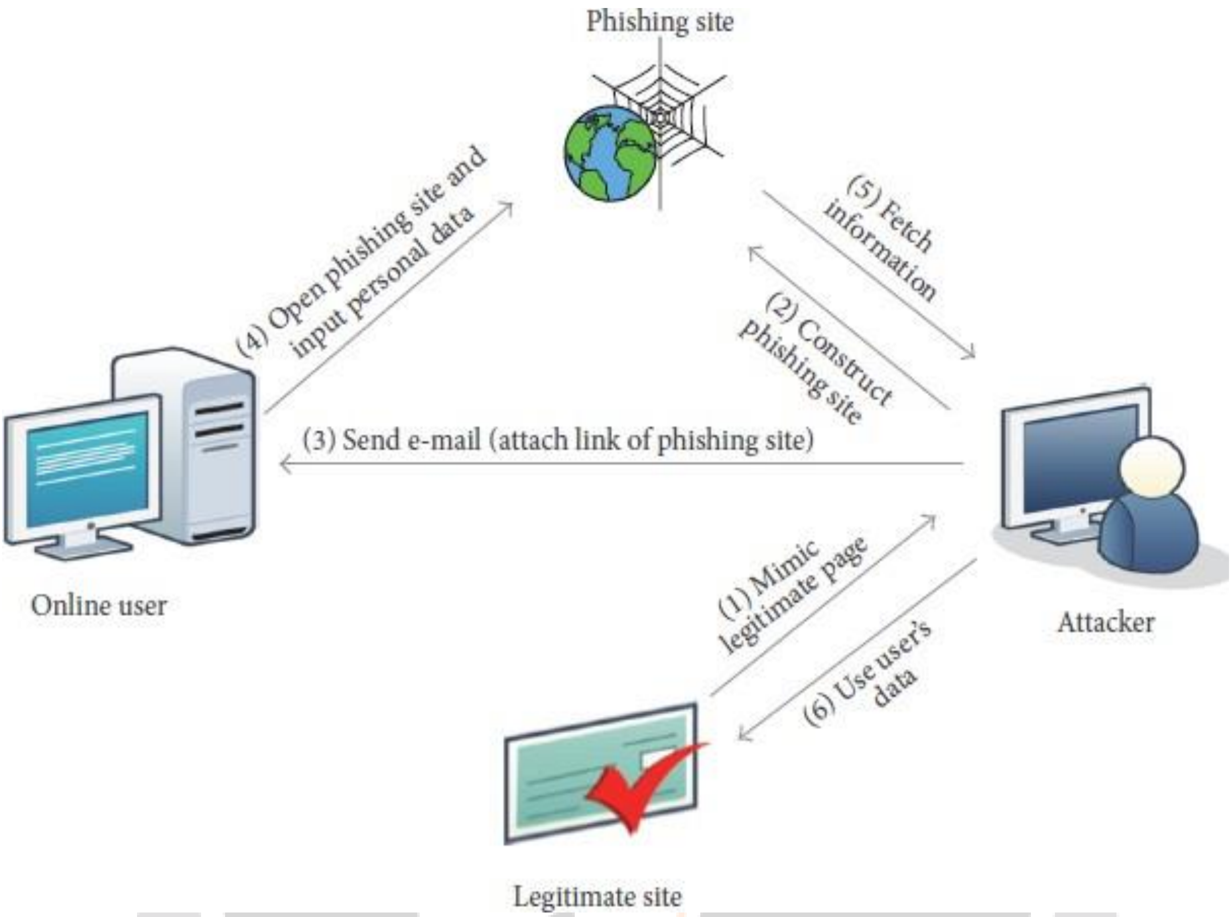

Phishing site
(4) Open phishing site and input personal data
(5) Fetch information
(2) Construct phishing site
(3) Send e-mail (attach link of phishing site)
Online user
Attacker
(1) Mimic legitimate page
(6) Use user's data
Legitimate site

Figure 1: phishing mechanism [6]

B. Taxonomy of phishing attack

[7] An intruder uses social engineering techniques and digital deceit to carry out a phishing assault. By forwarding a fake email, criminals carry out this assault via social engineering strategies. Hackers often ask their target to respond using the companies that issue credit cards, financial institutions, online merchants, as well as other businesses [8].

Malicious malware gets installed on the victim's device once they click on a phony email link, gathering and transmitting private data to the hacker. In the fraudulent emails, the hackers included fake links or uploaded malware (key logger software, for example, communicates the individual's key hit information). Criminals may also be able to gain entry to the victim' system and collect data anytime they desire.

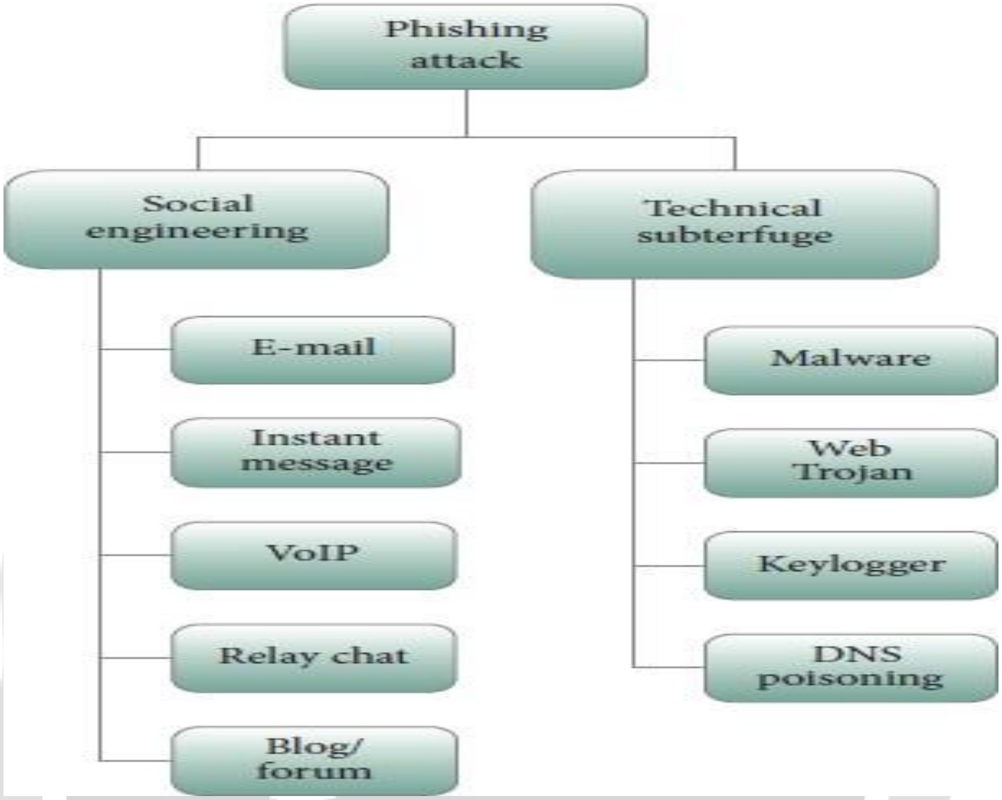


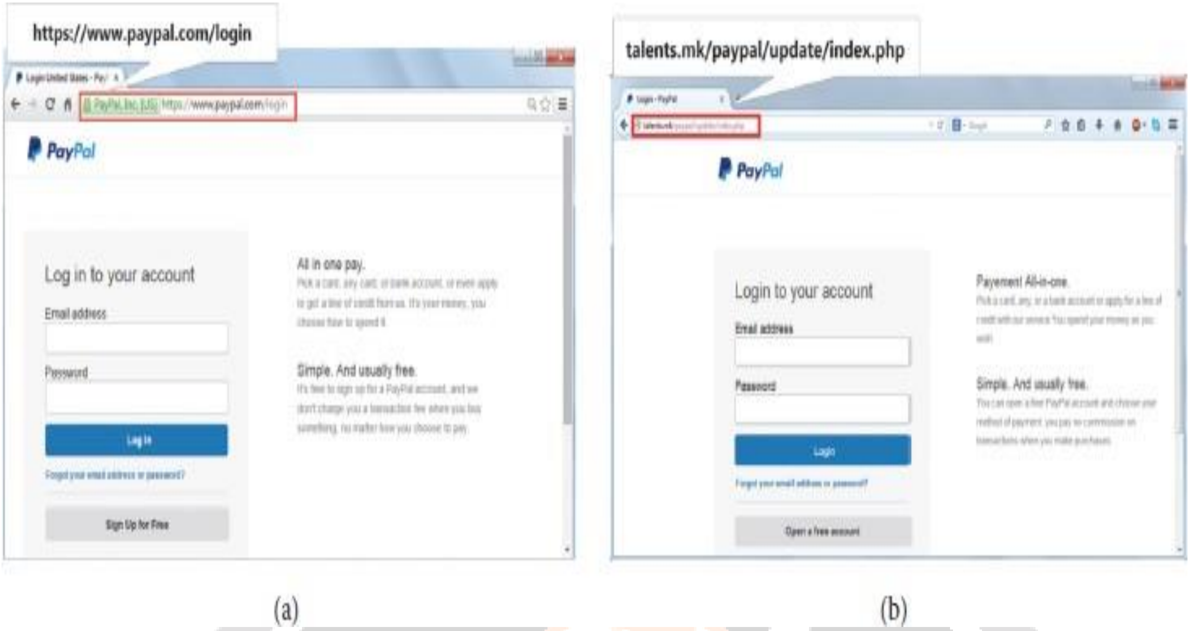
Figure 2: Phishing attacks Classification [6]

C. Visual similarity-based phishing detection and filtering approaches

An individual might be a target of the assault's by likening the impressive visual resemblances amid the phishing site and the authentic site they are targeting, including page layouts, images, written content, font sizes, and colour. Take, for example, the PayPal websites in Figure 3, which are both authentic and fraudulent. Both pages have an identical visual style even though their URLs are distinct. SSL (Secure Socket Layer) certificates and a website's URL are not always important details for users to focus on. If an attacker fails to closely replicate the visual layout of a targeted site, there is little chance that customers will submit their login details. A hacker can trick an individual in one of these methods:

1. **Visual Appearance:** The phony websites look a lot like the real one. The assail could obtain the HTML source code of a real site and use it to develop a fake one.

2. **Address Bar:** Hackers may conceal the address of a site or URL bar by employing scripts or a photo. The customer will assume they're on the right site when they provide their information. [6]


https://www.paypal.com/login
PayPal, Inc. [US] https://www.paypal.com/login
PayPal
Log in to your account
Email address
Password
Log In
Forgot your email address or password?
Sign Up for Free
(a)
talents.mk/paypal/update/index.php
PayPal
Login to your account
Email address
Password
Login
Forgot your email address or password?
Open a free account
(b)

All in one pay.
Pick a card, any card, or bank account, or even apply to get a line of credit from us. It's your money, you choose how to spend it.

Simple. And usually free.
It's free to sign up for a PayPal account, and we don't charge you a transaction fee when you buy something, no matter how you choose to pay.

Payement All-in-one.
Pick a card, any, or a bank account or apply for a line of credit with our service. You spend your money as you wish.

Simple. And usually free.
You can open a free PayPal account and choose your method of payment, you pay no commission on transactions when you make purchases.

Figure 3 (a) Legitimate PayPal webpage and (b) phishing webpage of PayPal. [6]

II. LITERATURE REVIEW

More than 38 percent of the global's population, or 2.97 billion people, used the Internet in 2014, based on Internet World Stats [9]. Hackers can fool unwary individuals into falling for phishing tricks since the Internet is not safe. Phishing emails are utilized online to trick financial organizations and individuals alike. The Anti-Phishing Working Group (APWG) is a multinational group that works to tackle digital cybercrime and fraud by promoting, education, research and law enforcement, as stated by [10]. Phishing attacks set a record in 2012, with a 160% increase in total attacks compared to 2011. The roughly 450,000 phishing assaults that were found in 2013 resulted in losses of over 5.9 billion USD [11].

The total attack incidence in 2013 was 1% higher than in 2012. The first quarter of 2014 saw the discovery of 125,215 phishing attacks, an increase of 10.7% over the fourth quarter of 2013. Over fifty-five percent of phishing sites exploit a targeted website's name in a manner designed to deceive users, and 99.4 percent of them use port 80 [4].

[7] Phishing assaults are carried out by a perpetrator who uses methods such as social engineering and cyber trickery. By forwarding a fake email, hackers carry out the assault via social engineering methods. Criminals usually ask victims for the names of financial institutions, credit card companies, online retailers, and other organizations. [8]. Dangerous malware is installed on the user's computer when they click on a phony email link, gathering and transmitting private data to the attacker. In the fraudulent emails, the perpetrators included malicious hyperlinks or attached malware (key logger programs, for example, communicates the user's key press details). Criminals may also be able to gain entry to the victim's computer and collect data anytime they desire.

Numerous critiques of earlier literary studies are given [12]. The study recommends the use of (CRI) - Crime, Prevention Review and Investigation of Knowledge Gap, to counteract phishing attacks.

[13]; proposed numerous models (techniques), along with different components of phishing attacks and methods to identify websites that are phishing. One of the paper's outstanding features is its examination of different phishing detection strategies and approaches. It also offers a proposed method for accurately identifying fake sites.

[14], suggest a grouping method for phishing assault classification. This method includes both classifying websites and extracting features from them. A total of thirty variables were identified and obtained via the repository data set

of (UCI) - Irvine machine learning after the concepts for phishing extracting features were outlined. Support Vector Machine (SVM), Naive Bayes (NB), and Extreme Learning Machine (ELM) were used to classify the data based on these attributes [14]. Having a precision of 95.34%, the Extreme Learning Machine (ELM) outperformed Support Vector Machine. It made use of six function activation. MATLAB was used to help with the outcomes.

[15], provide a technique that uses natural language processing and machine learning to detect phishing email attempts. The content undergoes a semantic examination to detect fraudulent intention. Natural language processing (NLP) methods are used to parse each phrase and identify the semantic functions of the words in respect to its premise. This method is developed using Python scripts and the Nazario email phishing dataset. Outcomes from Net-craft and SEA-Hound are compared; which show 98% and 95% accuracy, accordingly.

III. METHODOLOGY

The current phishing identification method uses machine learning algorithms and deep neural networks. The entire structure consists of two primary parts: an online application and machine learning models, which includes Auto Encoder Neural Network, Multilayer XGBooster, Random Forest, Support Vector Machine, and Decision Tree.

These models were selected after a number of comparisons between the outcomes of different machine learning techniques. A characteristic of websites derived from both genuine and phish datasets is used to test and train each of these algorithms. The best effective model is thus selected and incorporated into an online application that enables individuals to identify if a URL link is genuine or fraudulent.

A. Model Development

Supervised and unsupervised learning were used in the phases of developing machine learning models for phishing identification methods. The data needed to build the datasets for the training of the models is sourced from many open-sources platforms. Both legitimate and fraudulent URL datasets are included of the dataset gathering. This website provides a database of hourly-updated phishing URLs in a number of formats, such as CSV, JSON, etc. In order to train the machine learning models, more than 24,442 randomized phishing URLs were extracted from this set of data. Furthermore, a dataset of URLs which aren't harmful, spamming, phishing, or alteration can be found from the same freely available sites. Regardless of the various forms, the valid URL record is considered for this investigation. To train the machine learning models, more than 5000 genuine URLs were chosen at random from this set of data. The actual database was subsequently prepared for the machine learning model by pre-processing it; removing unnecessary and inaccurate data and encode it for phishing identification employing the One-Hot Encoder technique.

In order to identify trends and knowledge in the data set, it underwent exploratory analysis of the data, which involved examining, exploring, and summarizing it employing data visualization methods. Heat maps, scatter plots, and pair plots are some examples of such representations. Phishing and valid data sets were employed to obtain website content-based attributes, like the address bar-based feature (that possesses eight features), domain-based feature (that possesses three features), and the HTML & JavaScript-based feature (that possesses four features), through obtaining fresh attributes from the existing ones; in a data set. As a result, a total of fifteen (15) attributes were obtained for detecting phishing attacks.

These includes providing data to the machine-learning algorithms so they can identify and comprehend the dataset's positive features. The categorization challenge, the focus of this study's problem, was generated by supervised learning. The dataset is trained for identification of phishing using the following algorithms: Decision Tree, Random Forest, Support Vector Machines, XGBooster, Multilayer Perceptron, and Auto-encoder Neural Network. The dataset was used to create each of these models. A training set and an evaluation set are created from the dataset. Half of the data is used in the training phase to enable machine learning algorithms to comprehend the data and the ability to differentiate between legitimate and fake URLs. fifty (50) percent of the training sample is utilized for evaluation so as to measure the results of the trained dataset after fifty percent of the data sets was successfully trained. The process of evaluating a model involves figuring out how well it performs and measuring its generalization precision. In order to evaluate the efficacy of the algorithms used for identifying phishing attacks, numerous ranking and utility features were constructed utilizing the Scikit-learn (sklearn matrices) tool.

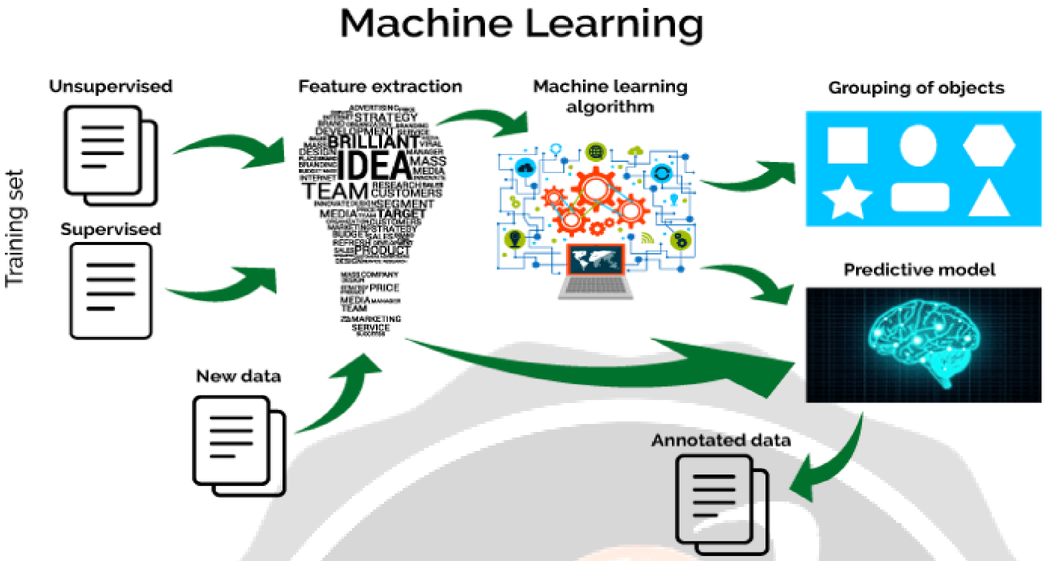


Figure 4: Machine learning development process [16].

A. System Architecture

Figure 5 shows the design of the suggested phishing identification method. Before the most effective algorithm with the greatest degree of precision is selected, a URL address that an individual submits goes past a number of trained machine learning and deep neural network algorithms. The selected model is integrated into a web application after being created as an API (Application Programming Interface). The web application, that is accessible via a variety of display gadgets such as desktops, laptops, tablets, and smart phones, is therefore used by the individual using it.

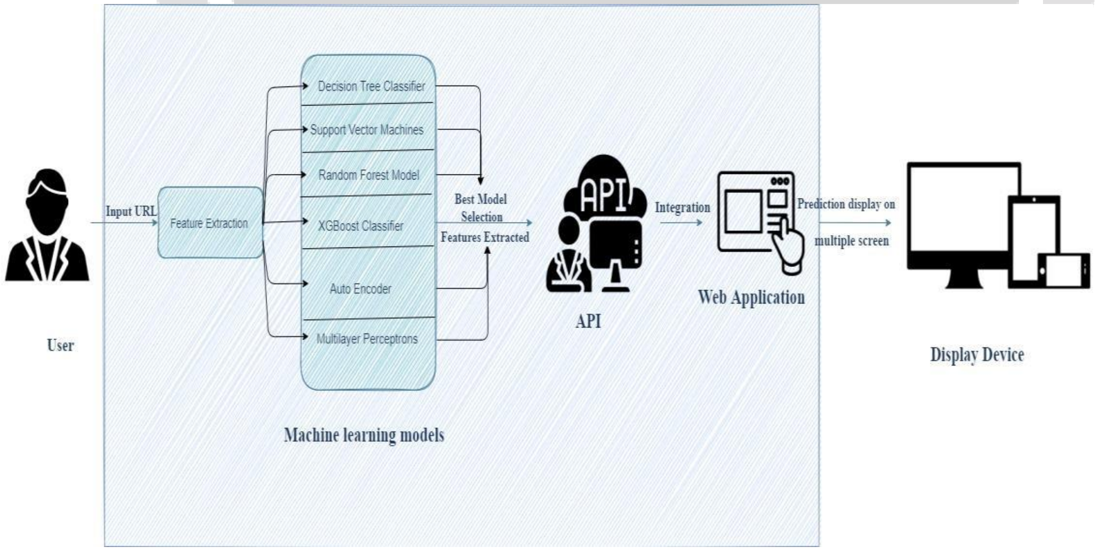


Figure 5: Architectural Design of the Phishing Detection System

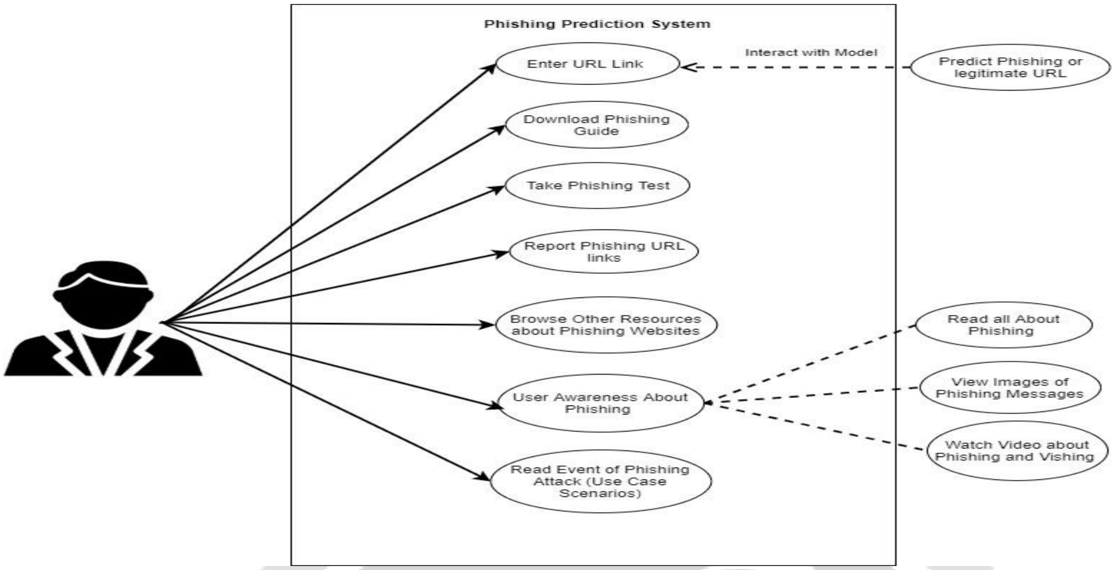

Phishing Prediction System
Enter URL Link
Interact with Model
Predict Phishing or legitimate URL
Download Phishing Guide
Take Phishing Test
Report Phishing URL links
Browse Other Resources about Phishing Websites
User Awareness About Phishing
Read all About Phishing
View Images of Phishing Messages
Watch Video about Phishing and Vishing
Read Event of Phishing Attack (Use Case Scenarios)

Figure 6: Use Case diagram of the new System

Figure 8 shows a schematic diagram of the machine learning method that phishing detection techniques employ. Figure 7 shows the phishing identification online method. The website evaluates the URL syntax once an individual submits a link to ascertain if it is authentic or fake.

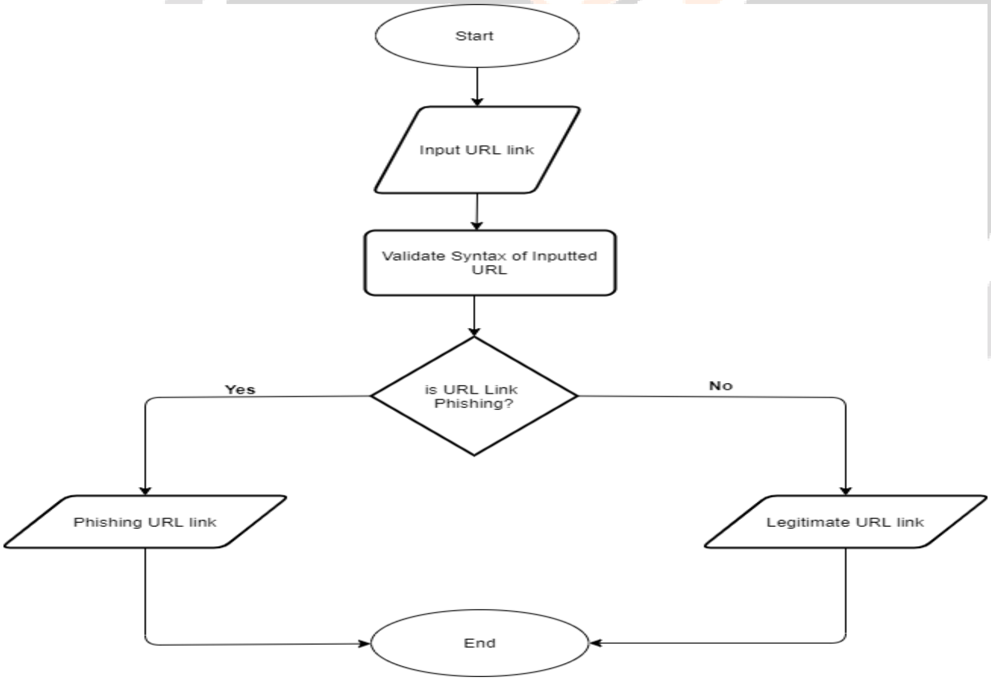


Figure 7: The Flowchart of the phishing detection web interface system

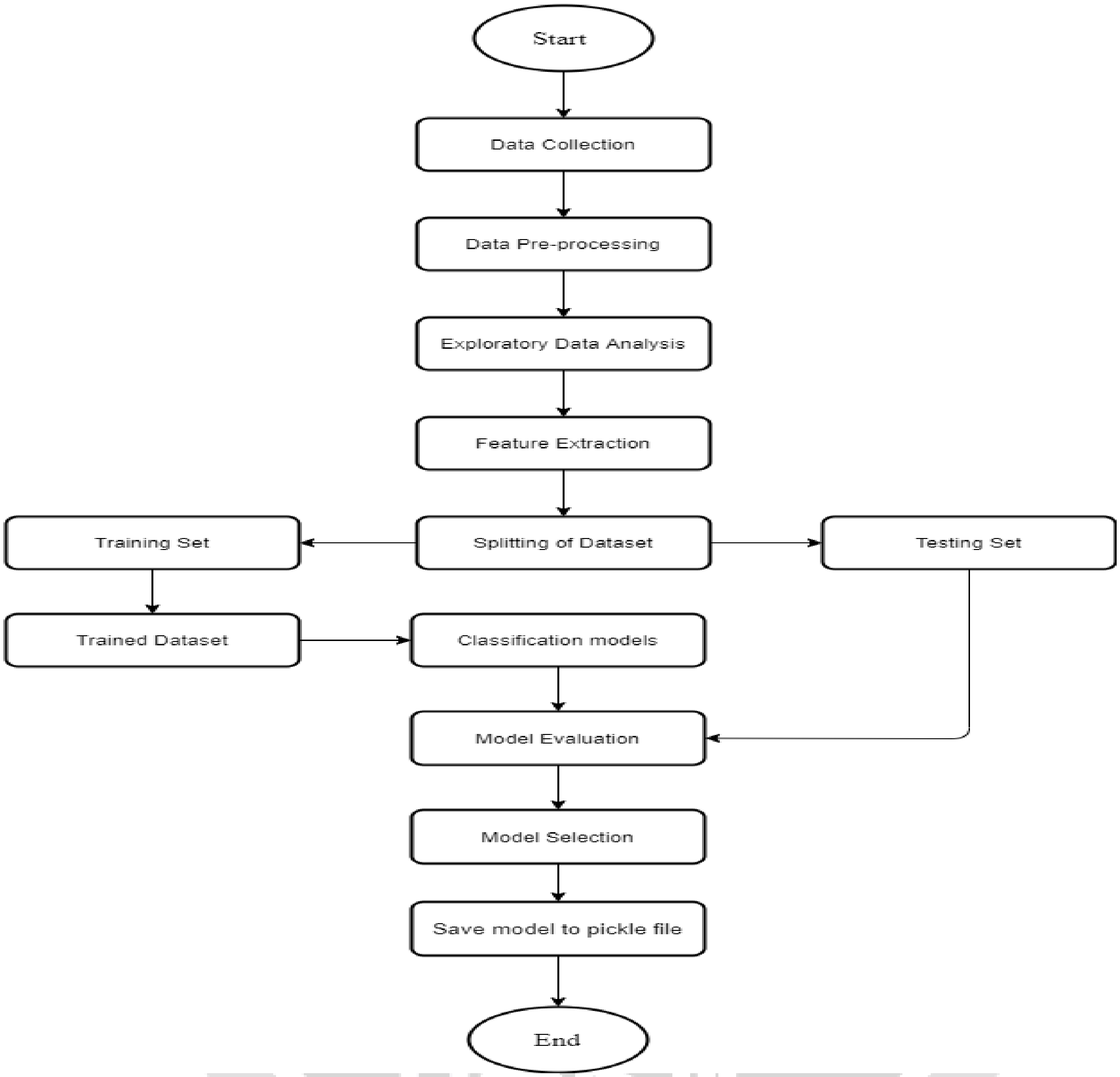


Figure 8: The flowchart of the machine learning technique for phishing detection systems.

The technology employed machine learning models and deep neural networks. Random Forest, Auto Encoder Neural Network, Multilayer Perceptions, Decision Tree, Support Vector Machine and XGBooster are some of these algorithms. The algorithms determine if the URL provided on the website is authentic or fraudulent. The algorithms produce a two-class prediction (genuine (0) and Phishing (1)). The method of constructing the model involved combining more than six (6) machine learning models with deep neural network algorithms to detect fake URLs using the Jupyter notebook IDE and packages such as pandas, Beautiful Soup, who-is, urllib, and others. The Sklearn matrix having a precision score, were used to determine the algorithms' efficiency.

IV. MITIGATION STRATEGIES

Though, phishing attempts have become increasingly sophisticated and difficult to identify, there continue to be warning indicators which may assist you identify phishing before it's become too late. Check out these crucial clues which security professionals employ to spot phony links: [17].

1. **Verify Suspicious URLs:** Phishing URLs frequently contain unusual characters, are lengthy, or are unclear. These are used by hackers to conceal the actual endpoint of the link and deceive clients. Examining the URL thoroughly is the initial measure in self-defense. Verify that it always starts with "HTTPS," since the "s" denotes a secure connection made with an SSL license. But remember that SSL certificates are not sufficient on their own. illicit material is now frequently being distributed by cybercriminals over HTTPS URLs that appear authentic. For this reason, URLs that appear to be an array of symbols or are extremely complicated ought to generate suspicions in your mind.

2. **Be Aware of Redirect Chains:** As the aforementioned instance illustrates, one of the primary strategies employed by cybercriminals is redirection. In addition to taking the URL's intricacy into account, see where the link takes you. By confusing users and lengthening the distribution chain, this strategy makes it more difficult to identify the fraudulent intent. Another frequent occurrence is when hackers send an email stating that a file must be downloaded. However, they transmit a URL that leads through redirects and eventually requests login information in order to access the file, rather than an attachment or direct link.

3. **Look for Odd Page names and Missing Favicons:** Examining website names and icons is an additional method to identify suspicious links. A genuine website ought to be free of odd characters and jargon, and its title should correspond to the service you are using. Incomplete titles or suspicious, arbitrary strange are frequently indicators that something is amiss. Authentic pages have a logo that matches the product or service in addition to the website's name. A phishing assault is indicated by a missing or nondescript favicon.

4. **Watch Out for Abused CAPTCHA and Cloudflare Checks:** Abuse of CAPTCHA systems, especially the "I'm not a robot" authentication, is a typical strategy employed in phishing attempts. Phishing criminals can take advantage of CAPTCHAs by including pointless, recurring CAPTCHA tasks on fraudulent web pages, even though they are intended to authenticate individuals and guard against machines. A comparable method is the abuse of services such as Cloudflare, where hackers can impede delay to victims and conceal the phishing attack by using Cloudflare's authentication processes.

5. **Check Microsoft Domains Before Entering Passwords:** Hackers frequently construct webpages that imitate reputable companies, such as Microsoft, in an attempt to fool customers into divulging their login details. Even though Microsoft usually requests passwords for a select few authorized domains, it's still advisable to exercise caution. Remember that your company can also utilize its corporate domain to seek verification. Verifying the link prior to releasing the login details is consequently; usually a smart practice.

6. **Examine Links with Familiar Interface Features:** By attentively scrutinizing program interface elements, you can also identify suspicious links. Remember that program interface components on a browser page that has an input field for entering a password are a serious red flag. By imitating well-known software interfaces, like those from Adobe or Microsoft, and inserting password entry forms inside, phishers frequently try to win over customers' confidence. People at risk become more at ease and let their guard down as a result, which eventually leads them to fall into the hacking bait. Verify URLs with these components twice prior to providing confidential details.

V. CONCLUSION

Phishing attacks are an ever-increasing hazard in cyberspace and cost internet users billions of dollars annually. It uses a range of social engineering strategies to obtain confidential data from consumers. Therefore, a variety of communication pathways, including websites, pop-up notifications, messaging apps, and electronic mail, may be utilized to identify phishing actions. The different approaches investigators have used to help solve the problem of phishing detection were categorized and determined in this investigation. The developed system used various feature selection, deep neural network and machine learning methods, including Decision Tree, Support Vector Machine, XGBooster, Multilayer Perceptions, Auto Encoder Neural Network, and Random Forest, to find trends that made it easy recognize URL links. Utilizing the feature extraction method, the model has been linked with the web application that lets individuals enter links to websites and decide if they are fake or authentic. Employing extracted features and algorithms deployed to the data set, malicious URLs were precisely recognized, improving the models' computational precision. Also, incredibly good at figuring out if the web address is authentic.

VI. ACKNOWLEDGMENT

We would like to sincerely thank the TETFund for funding and facilitating the effective completion of the study, as well as the Vice Chancellor of Imo State University in Owerri for encouraging collaboration in research at the institution.

VII. REFERENCE

[1] Pamela (2021). Phishing attacks. Retrieved from https://www.khanacademy.org/computing/computersandinternet/xcae6f4a7ff01e7d:online-data-security/xcae6f4a7ff015e7d:cyber-attacks/a/phishing-attacks.

[2] Imperva. (2021). Phishing attacks. Retrieved from https://www.imperva.com/learn/application-security/phishing-attack-scam/.

[3] KnowBe4 (2021). Phishing Techniques. Retrieved from https://www.phishing.org/phishing-techniques.

[4] Anti-Phishing Working Group (APWG) Phishing activity trends report the first quarter. (2022) Retrieved from http://docs.apwg.org/reports/apwg trends report q1.

[5] Krugel C., Vigna G, Robertson W. (2005). A Multi-Modal Approach to the Detection of Web-Based Attacks. Computer Networks 48(5): 717-738; doi: 10/1016/j.connet.2005.01.009.

[6] Ankit K.J and Gupta B.B (2017). Phishing Detection: Analysis of Visual Similarity Based Approaches. Security and Communication Networks 2017 (4): 1-20; doi: 10:1155/2017/5421046.

[7] Almomani, A., Gupta, B. B., Atawneh, S., Meulenberg, A., & Almomani, E. (2015). A survey of phishing email filtering techniques, Proceedings of IEEE Communications Surveys and Tutorials, vol. 15, no. 4, pp. 2070–2090.

[8] Tewari, A., Jain, A. K, & Gupta, B. B. (2016). A recent survey of various defense mechanisms against phishing attacks. Journal of Information Privacy and Security, vol. 12, no. 1, pp. 3–13

[9] Internet world stats usage and population statistics. (2014). Retrieved from http://www.internetworldstats.com/stats.htm.

[10] APWG report. (2012). Retrieved from http://apwg.org/download/document/245/APWG Global Phishing Report 2H.

[11] RSA Anti-Fraud Command Center (2014) Retrieved from https://www.emc.com/collateral/fraud-report/rsa-online-fraud-report-012014.pdf.

[12] Anjum N. S., Antesar M. S., & Hossain M.A. (2016). A Literature Review on Phishing Crime, Prevention Review and Investigation of Gaps. Proceedings of the 10th International Conference on Software, Knowledge, Information Management & Applications (SKIMA).

[13] Ashritha, J. R., Chaithra, K., Mangala, K., & Deekshitha, S. (2019). A Review Paper on Detection of Phishing Websites using Machine Learning.Proceedings of International Journal of Engineering Research & Technology (IJERT), 7, 2. Retrieved from www.ijert.

[14] Sönmez, Y., Tuncer, T., Gökal, H., & Avci, E. (2018). Phishing web sites features classification based on extreme learning machine. 6th Int. Symp. Digit. Forensics Secure. ISDFS 2018 - Proceeding, vol. 2018–Janua, pp. 1–5.

[15] Peng, T., Harris, I., & Sawa, I. (2018). Detecting Phishing Attacks Using Natural Language Processing and Machine Learning. Proc. - 12th IEEE Int. Conf. Semant. Comput. ICSC 2018, vol. 2018–Janua, pp. 300–301.

[16] Ayush, P. (2019). Workflow of a Machine Learning project. Retrieved from https://towardsdatascience.com/workflow-of-a-machine-learning-project-.

[17] Expert Tips on How to Spot a Phishing (2024). Linkhttps://thehackernews.com/2024/09/expert-tips-on-how-to-spot-phishing-link.html?m=1