ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAY353

Development of a Phishing Detection System Using Support Vector Machine

Akinwole Agnes Kikelomo
Computer Technology Department,
School of Technology,
Yaba College of Technology, Lagos,
Lagos, Nigeria.

Ogundele Israel Oludayo
Computer Technology Department,
School of Technology,
Yaba College of Technology,
Lagos, Nigeria.

Abstract:- Phishing represents a significant and escalating threat within the cyber domain, inflicting substantial financial losses on internet users annually. This illicit practice leverages both social engineering tactics and technological means to unlawfully obtain sensitive information from individuals online. Despite numerous studies and publications exploring various methodologies to combat phishing, the number of victims continues to surge due to the inefficiencies of current security measures. The inherently anonymous and unregulated nature of the internet further compounds its susceptibility to phishing attacks. While it's commonly believed that successful phishing endeavours involve the creation of replica messages or websites to deceive users, this notion has not undergone systematic examination to identify potential vulnerabilities. This paper endeavours to fill this gap by conducting a comprehensive evaluation of phishing, synthesizing diverse research perspectives and methodologies. It introduces an innovative classification method utilizing Support Vector Machine (SVM), achieving an impressive accuracy rate of 96.4% in detecting phishing attempts. By implementing this model to distinguish between phishing and legitimate URLs, the proposed solution offers a valuable tool for individuals and organizations to promptly identify and mitigate phishing threats. The findings of this study hold significant implications for bolstering internet security measures and enhancing user awareness in navigating potentially malicious online content.

***Keywords:-** Phishing, Software Detection, Cybersecurity, Support Vector Machine, URL.*

I. INTRODUCTION

Systems that are connected to the internet, including their data, software, and hardware, are protected from cyber threats by cybersecurity. To prevent illegal access to data centers and other digital systems, both private persons and commercial organizations utilize this technology. A strong cybersecurity plan helps provide a good defense against harmful attacks that aim to access, change, remove, destroy, or extort sensitive information and systems that belong to a person or company. For all users of computers and the internet, information security is essential, and when navigating various websites, one should reduce the possibility of fraud [1].

One of the most challenging issues to prevent and eliminate is phishing, a problem that has been since the beginning of the internet era. A sort of cybercrime known as "phishing" occurs when an attacker contacts one or more targets by phone, text, or email while pretending to be a reputable business in an attempt to coerce them into disclosing sensitive information like passwords, bank account information, and credit card details. One of the biggest online security risks of our time is phishing attacks. Typically, attackers use phony websites to obtain the victims' private information. The Anti-Phishing Working Group [2] reports that 138,328 phishing pages were reported in the fourth quarter of 2018 and that the trend of phishing attempts is rising annually. Numerous financial losses result from it. Based on the incidents reported to the Federal Bureau of Investigation [3], about $48 million was lost in the United States in 2018. Furthermore, phishing attempts are quickly rising to the top of the malware delivery list [4-5]. According to a recent Microsoft security intelligence report [6], the most common online threat in 2018 was phishing.

Phishing websites have the potential to cause malicious software to be installed, giving hackers remote access to the entire system. The majority of phishing websites have an exact replica of legitimate websites, and when their domain names are compared to those of the original websites, the only differences are minor typos or the use of characters that look similar to trick the target into entering sensitive information or granting access by installing malicious software[7][8].

There are two primary groups into which phishing detection techniques fall. The first, known as user awareness, tries to teach consumers how to distinguish between phishing and non-phishing emails. Using a combination of blacklists, heuristics, visual resemblance, and machine learning (ML), the second method is called software detection and is used to identify phishing attempts [9].

Teaching computers to learn from experiences in the same manner that humans and other animals do is the aim of machine learning research. Machine learning algorithms "learn" directly from data using computer technology, as opposed to using a predefined equation as a model. There are two approaches to it: unsupervised learning, which employs internal structures or hidden patterns in the input data, and supervised learning, which trains a model to predict future outcomes using known input and output data [9].

According to Bambrick [10], one supervised machine learning technique that can be applied to regression and classification is the support vector machine (SVM). Because SVMs can solve both linear and nonlinear issues and are based on the idea of dividing a dataset into two classes using the best hyperplane, they are more frequently used in classification challenges [11].

These techniques' effectiveness and performance are influenced by the feature set, quantity of training instances, and classification algorithms. The performance of classification models can be improved by keeping the number of training instances and the classification models separate and by taking into account a wide range of attributes. Unfortunately, the longer model creation takes, the more difficult it is to identify phishing websites quickly. It is therefore important to choose the right characteristics in order to construct the classification models faster without sacrificing accuracy. In light of this, this chapter presents a study on feature selection techniques' utility in identifying phishing websites.

The effectiveness of machine learning methods with and without feature selection is therefore compared. Thus, the goal of the study is to reduce the amount of people who fall victim to fake websites and reveal their personal information by creating a phishing website detection system utilizing the Support Vector Machine (SVM) search engine.

The research is organized as follows. Section 2 present on the existing literature review in phishing attack, machine learning and its related work on detection of phishing attack using machine learning. Section 3 explains the methodology used in this study and the design of the web application. Section 4 analyze and discuss on the experimental results of phishing attack to determine legitimate and non-legitimate. Section 5 was on the conclusion of the research work.

II. LITERATURE REVIEW

There are researches in the literature that concentrate on identifying phishing assaults. Some useful and efficient defense strategies are emphasized in the recent surveys, where authors classify the technical approaches employed in these kinds of assaults and discuss the general characteristics of the current phishing schemes [12][13]. According to a relevant study on the experiences of phishing assaults, computer users fall victim to phishing for the following five primary reasons:

- Users don't have a thorough understanding of URLs,
- They also don't know which websites are trustworthy,
- They can't see the entire address of a website because of redirections or hidden URLs
- They don't have enough time to check the URL or accidentally enter some pages; and
- They can't tell phishing websites from genuine ones.

A. *Types of Phishing Attacks*

Attackers study the weaknesses in internet security through a variety of tactics. They constantly search for ways to get beyond the security system's restrictions. The following list includes a variety of unique phishing attacks.

- *Phishing using Algorithms:* America Online (AOL), which was built with an algorithm, detected the initial phishing attempt. The credit card numbers of America Online accounts were matched by an algorithm created by the fraudster [15].

- *Deceptive Phishing:* Users on the internet are tricked by fraudsters using various techniques. To validate the account, fishermen send emails to the users. They require that you click buttons and links. The website that lies behind the links is where hackers steal and store user personal information.

- *URL Phishing*: This type of phishing assault uses a hidden link to target victims via the Universal Resource Locator (URL). The hackers' website can be accessed by clicking the provided link. Upon clicking the link, the user's information is stored on the hackers' website and is redirected [16].

- *Hosts File Poisoning:* This technique is used on Windows operating systems to contaminate host files. The intended website is either redirected to a hacker's website or returns the error message "The Page Not Found" when the user finds it. The user's data is captured and taken if it can be redirected to the false website.

- *Injection of Content Phishing:* When a user is targeted by hackers, they pose as genuine websites. The intention is to mislead the user or portray the company incorrectly. Another name for it is "content spoofing." This tactic is employed by the attackers to trick the user and gather data for their server.

- *Clone Phishing:* "Clone" refers to the biotechnological process of creating an individual that is identical to the original. That occurs frequently in genetic engineering. Another type of phishing attack is called "clone phishing," in which the sender or recipient of the email is compromised by an adversary. A similar original email is created by the malevolent attacker and sent to the first or second person along with an attachment or link. They ask for the original to be sent in an updated version [17].

- *Whalering:* The organization's upper leaders are the focus of this kind of phishing attack. The email, which is addressed to the executives, discusses significant matters. The email's message may mirror the grievances of the clients.

- *Spear Phishing:* This type of email scam is designed to target particular individuals and businesses. To get a response from the intended recipient, the attacker emails them. The email is written in such a way that it appears as though they are aware of numerous details about the victims, including their name, email address, and place of employment [18].

B. *Phishing Detection Methods*

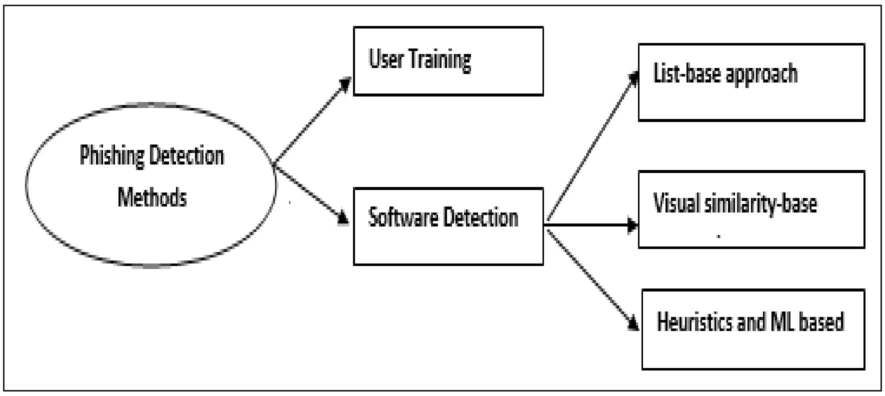


Fig 1: Phishing Detection Methods

Figure 1 shows various phishing methods for mitigating attacks. Every step of the attack cycle has a different set of suggested defences against phishing attempts. While certain cybersecurity techniques operate automatically and provide user alerts, others require training to ensure users are prepared for potential attacks. Below is a list of these techniques:

- ➢ *User Training*
- ➢ *Software Detection*

➢ *User Training*

Educating users and staff members about phishing attacks and providing them with warnings can help in avoiding such threats. Various techniques have been proposed for training users, with several studies indicating that interactive training is the most effective approach for helping users distinguish between phishing and legitimate websites. User training works well, but human error still happens, and individuals tend to forget what they've been taught. Non-technical users do not find training to be very valued and it takes a lot of time [19].

➢ *Software Detection*

Although some phishing attacks can be avoided with user training, there are hundreds of websites that we encounter every day, making it difficult and occasionally impractical to apply our training to every one of them. Using the program is another option for identifying phishing websites. Beyond human judgment, the machine is more accurate in its analysis of numerous variables, including the URL, email message text, and website content, before reaching a conclusion. Phishing detection software is available in several forms and is divided into the following categories:

- List-based approach: Utilizing blacklist-based anti-phishing methods integrated into web browsers is among the widely adopted approaches for detecting phishing attempts. These methods rely on two distinct lists: the blacklist, which logs known malicious websites, and the whitelist, which catalogues verified legitimate sites. Blacklists are usually curated from user submissions or third-party reports sourced from alternative phishing detection systems. Research suggests that blacklist-based anti-phishing techniques can successfully flag approximately 90% of fraudulent websites during initial screening [20].
- Visual similarity-based approach: The visual similarity between phishing websites and their legitimate counterparts is a significant factor in deceiving individuals into believing they are accessing a trustworthy site. This similarity often leads users to unknowingly input their information into malicious platforms. To detect phishing websites, certain tools examine various elements such as text content, formatting, HTML, CSS, and images on web pages to identify visual resemblances [21][22]. Additionally, discriminative key point features that treat phishing detection as an image matching problem were proposed by Chen et al. [23]. There are drawbacks to visual similarity-based techniques. For instance, techniques based on a website's content won't identify websites that utilize images rather than words. Image matching techniques take a long time and are difficult to collect sufficient data from them [23].
- Heuristic and machine learning-based: Machine learning techniques have shown to be an effective means of categorizing hostile behaviors or artefacts, such as phishing websites or spam emails. Fortunately, there are plenty of phishing website samples available to build a machine learning model, as most of these methods require training data. Some machine learning systems detect phishing by using the content and features of the website, while others analyze a webpage snapshot using vision techniques [24]. A variety of machine learning techniques, including those covered in the section below, have been employed to identify phishing websites. These techniques include Ada boost, SVM, KNN, neural networks, gradient boosting, XGBoost, decision trees, random forests, and logistic regression [25][26].

C. *Machine Learning Techniques*

Artificial intelligence (AI) techniques, such as machine learning, enable computers to learn from experience. Machine learning algorithms utilize mathematical methods to extract insights from data without relying on predefined equations as models. As more data becomes available, these algorithms adapt and improve their performance. Deep learning represents a sophisticated subset of machine learning [27]. In order to predict future outputs, machine learning uses two main methods: unsupervised learning, which looks for innate structures or hidden patterns in input data, and supervised learning, which includes training a model using known input and output data.

- ➢ *Supervised Learning*

When applied to uncertain situations, supervised machine learning creates a model that predicts using available data. Supervised learning involves training a model using a predefined set of input data and corresponding outputs. This enables the model to make accurate predictions for new data inputs. When you have known data and want to predict outcomes, supervised learning is the method to use. Regression algorithms and classification are employed in supervised learning to develop machine learning models. Regression analysis is a tool for predicting continuous responses, like changes in temperature or electricity usage. Examples of standard regression techniques encompass neural networks, boosted and bagged decision trees, regularization, stepwise regression, adaptive neuro-fuzzy learning, as well as linear and nonlinear models. Regarding classification, common approaches include support vector machines (SVMs), logistic regression, boosted and bagged decision trees, k-nearest neighbors, Naive Bayes, discriminant analysis, and neural networks [28].

- ➢ *Unsupervised Learning*

Unsupervised learning involves the detection of underlying structures or hidden patterns. It is employed to remove entries from datasets that have input data but no labelled answers. A popular technique for unsupervised learning is clustering. Cluster analysis is employed in exploratory data analysis to uncover concealed patterns and groupings within datasets [29]. Its applications span various domains such as product identification, market research, and gene sequence analysis. For instance, a cellular phone company can utilize machine learning to predict tower usage in order to enhance the locations where it installs towers. The team uses clustering algorithms to produce the best possible cell tower location because a phone can only communicate with one tower at a time. This allows the phone to receive the best possible signal for their groups or groups of clients. Among the common clustering algorithms are subtractive clustering, Gaussian mixture models, hidden Markov models, fuzzy C-means clustering, hierarchical clustering, k-means, k-medoids, and self-organizing maps [30][31].

D. Support Vector Machine

The support vector machine (SVM) is among the most commonly utilized classifiers. This algorithm aims to identify the point closest between two classes by maximizing the distance between them. This method of linear and nonlinear classification uses a supervised learning model. A kernel function is used in nonlinear classification to translate the input to a higher-dimensional feature space. SVM have some weaknesses despite being quite powerful and frequently used in classification. Higher computations are required for data training. Furthermore, they are prone to overfitting since they are sensitive to noisy input. The sigmoid, polynomial, linear, and RBF (radial basis function) are the four common kernel functions at the SVM. The plane's equation is show in figure 2, which divides the data, is $(x)=wTx+b$, where w and b are weights that the model has learned [32][33].

$$\mathbf{Y}=\begin{cases}\mathbf{1}, & \mathbf{if\ g(x)\geq 1}\\ \mathbf{0}, & \mathbf{if\ g(x)\leq -1}\end{cases}$$

Fig 2: Plan Equation for SVM

Since the plane serves as a division between them, locations below $(x)=-1$ are classified as class 0, and ones above $g(x)=1$ plane are classified as class 1.

E. Related works

Various techniques exist for identifying phishing websites, such as heuristic-based, visual-similarity, blacklisting, whitelisting, and more. Nonetheless, machine learning and deep learning methods possess inherent capabilities to detect zero-hour or recently emerged phishing attacks. Researchers have dedicated significant efforts to tackle this ongoing challenge, driven by attackers' continuous efforts to exploit weaknesses in existing anti-phishing systems. Machine learning and deep learning approaches offer promising solutions in this regard.

[34] used the Random Forest classifier (RF) obtained 98.11% accuracy in phishing detection using machine learning. The study was carried by using a dataset that included 10,000 evenly distributed phishing and legal websites. Twenty features were chosen during the feature selection procedure out of the 48 features in the dataset. The study came to the conclusion that feature selection helped to increase the detection method's accuracy.

[35] developed a machine learning model for distinguishing between phishing and authentic websites and included over thirty URL-based variables. The machine learning repository offered by UCI is where the URL dataset was acquired. Next, dimensional reduction was applied to the dataset. Utilizing the smaller dataset, they conducted machine learning algorithms such as regression trees, recursive partitioning, random forest, support vector machines, generalized linear models, and generalized additive models. The most effective algorithm for the situation was the RF algorithm with 300 trees, which yielded an accuracy of 96.65 percent and a precision of up to 97.4 percent. By using higher-order dimensionality reduction methods, such as the variance inflation factor (VIF), the results can be enhanced.

[36] uses Internet of Things datasets to identify phishing attempts. The data between authentic and phishing was distinguished using the standard machine learning techniques of SVM, RF, DT, NN, and linear models. We then compared the output of these machine learning models on various datasets and with the prior dataset. With the exact conventional traffic that may have been targeted by botnet

attacks, UNSW-NB15, a dataset created by the Australian Center of Cyber Security at UNSW Canberra, was collected. 96.85% accuracy was the best resulted from the random forest algorithm. Non-numeric values will impact efficiency and results. Hybrid machine learning techniques are suggested to improve outcomes efficiency.

[37] offer an effective method for phishing detection by machine learning in their journal article, "An Efficient Approach for Phishing Detection using Machine Learning." In order to create high-performance classification models faster, the authors use a feature selection technique to shortlist a collection of characteristics. They used an 11,055-person phishing dataset with 30 features for their studies. In order to accelerate the build time of classification models for phishing detection while maintaining accuracy, a number of machine learning techniques are employed.

[38] proposed a technique using web page similarity and URL-based discovery for phishing detection and prevention. To examine the extracted URL that directs users to the website and the virtual URL that they view, they employ the LinkGuard4 algorithm. The system switches to visual similarity-based identification if the URL-based technique fails to identify phishing. Due to the small number of websites they chose for their test trial, their test result is regrettably not robust; therefore, in order to increase precision, a thorough study must be done.

III. METHODOLOGY

The primary goal of this research is to design a model that enables the detection of phishing URLs using machine learning. To capture these illegal websites, the Support Vector Machine Algorithm was used. Given the dynamic nature of the phishing attack, the machine learning algorithm employed in this system allows for the identification of much improved performance. Before classifying a website as phishing or legitimate, the model needs to extract specific attributes from the provided URL. Once this website information is collected and categorized, the user receives a phishing report. One aspect addressed in this architecture is the creation of new models as fresh data sets arrive. This involves repeated training of the model after the initial training to help the system adapt to new attack vectors through the relearning process. This approach proves useful when dealing with large or non-stationary data. When it comes to non-stationary data, batch algorithms typically don't work well if there is confusing information present, such as distinct distributions that change over time, and the batch method integrates it incorrectly. The dataset that is kept in the URL list will be tracked in order for the learning model to continuously train itself in order to accomplish this functionality.

➢ *Data Collection and Gathering*

The information gathered from www.kaggle.com is utilized to create the datasets that are used for training the models. The dataset collection includes datasets with authentic and phishing URLs. A feature extractor technique will be utilized to create the feature vector from this data, making use of the dataset for both training and testing.

➢ *Software Tools Used for Development*

Multiple tools and frameworks were utilized in crafting the different components and prototypes. Python was chosen as the programming language and configured for both developing and testing the model. Various machine learning models and libraries such as pandas, scikit-learn, python-whois, BeautifulSoup, NumPy, TensorFlow, and numerous other essential packages were incorporated into the development process.

Table 1: Attributes Used for Phishing Detection

S/n	Attributes	S/n	Attributes	S/n	Attributes
i	IP address	xi	Using Non-Standard Port	xxi	Disabling right-click
ii	Length of the URL	xii	HTTPS Token	xxii	Using pop-up windows
iii	URL shortening Service	xiii	Abnormal request URL	xxiii	iframe
iv	URLs having the “@” symbol	xiv	Abnormal URL of anchor	xxiv	Age of the domain
v	Redirecting using “//”	xv	Links in <meta>, <script>, and <link> tags	xxv	Abnormal DNS record
vi	URLs with prefix and suffix	xvi	Server Form Handler (SFH)	xxvi	Web Traffic
vii	Sub-domain(s) in URL	xvii	Submitting to Email	xxvii	Page Rank
viii	Using Secure Sockets Layer (SSL) Certificate	xviii	Abnormal URL Address	xxviii	Google index
ix	Favicon	xix	Redirect Pages	xxix	Links pointing to page
x	Domain Registration	xx	On Mouse over to hide the link	xxx	Statistical Report

➢ *Features Extraction*

A genuine website can be distinguished from a phishing one by the characteristics and attributes of the users accessing the website domain. In this research work, we gathered Thirty (30) phishing features and indicators, and they were clustered into six. The feature extraction model will try to extract these Thirty (30) features from a specific webpage.

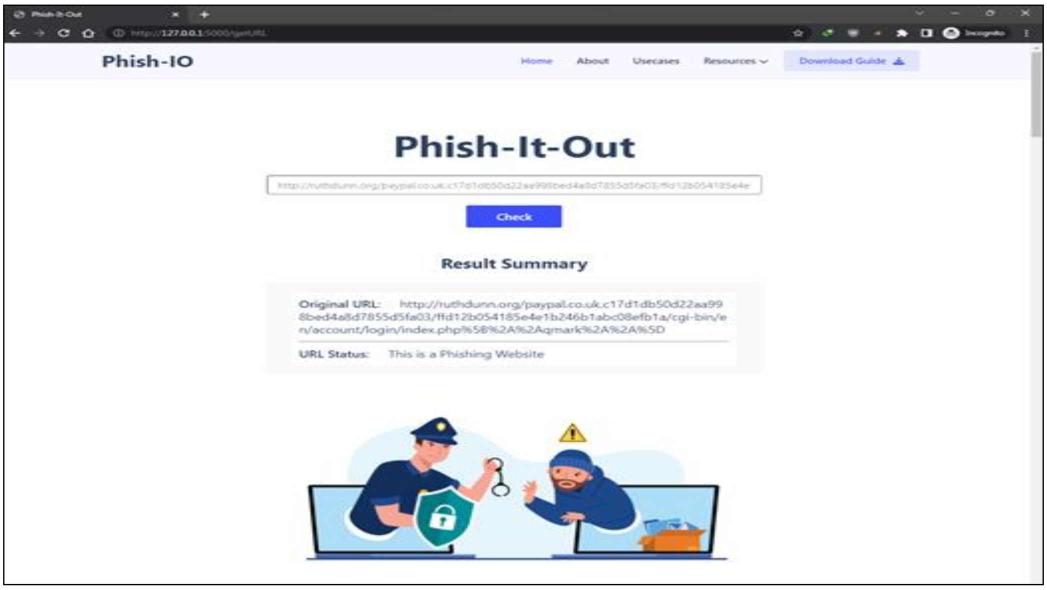
- *System Design of the Development*

The phishing detection web application called "Phish-It-Out" has been developed to run on any browser. The application was developed using programming languages such as Python, HTML, CSS & JavaScript.

The phishing detection web application has the following pages:

- *The Home Page*

The home page contains an input form for the user to enter a URL and check if it is a phishing or legitimate website as shown in figure 3. It checks the state of the URL based on the feature selection as discussed earlier. The purpose of this page is to help its users validate a URL link and also provide various resources on phishing attacks. The User can also take a google phishing test to help understand how to detect phishing messages and URLs. Also, users can download a book that contains information and other resources on phishing.



Phish-It-Out

Result Summary

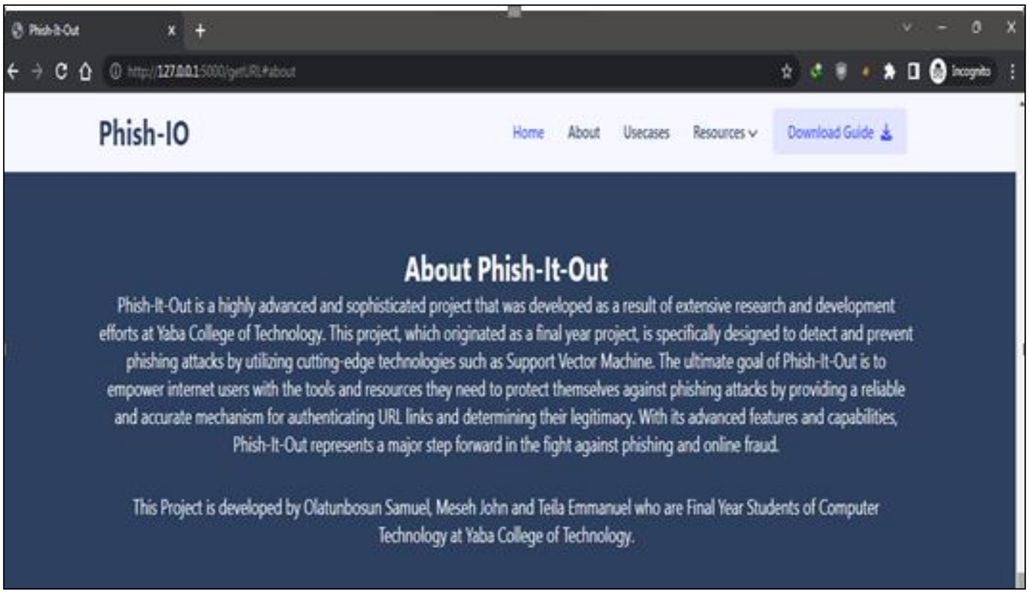
Original URL: http://ruthdunn.org/paypal.co.uk.c17d1db50d22aa998bed4a8d7855d5fa03/ffd12b054185e4e1b246b1abc08efb1a/cgi-bin/en/account/login/index.php%5B%2A%2Aqmark%2A%2A%5D

URL Status: This is a Phishing Website

Fig 3: The User Home Page for phish-it-out

- *The About Page*

The about page contains details about the web application and information on the phish-it-out application shown in figure.



About Phish-It-Out

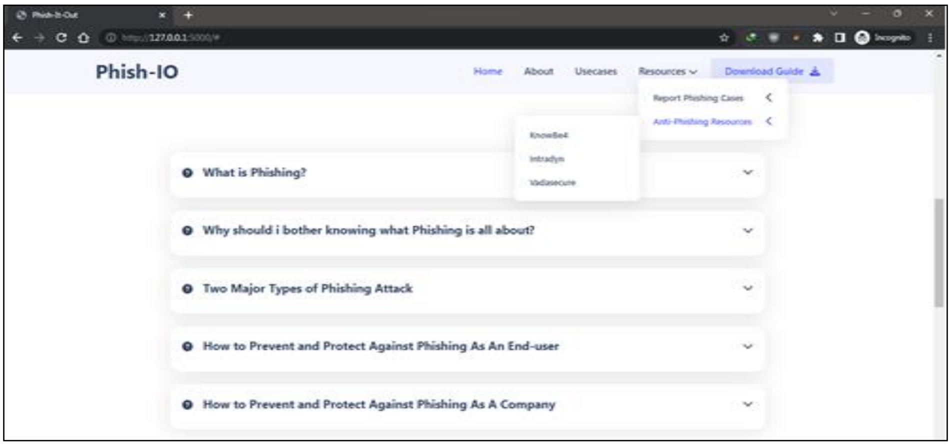
Phish-It-Out is a highly advanced and sophisticated project that was developed as a result of extensive research and development efforts at Yaba College of Technology. This project, which originated as a final year project, is specifically designed to detect and prevent phishing attacks by utilizing cutting-edge technologies such as Support Vector Machine. The ultimate goal of Phish-It-Out is to empower internet users with the tools and resources they need to protect themselves against phishing attacks by providing a reliable and accurate mechanism for authenticating URL links and determining their legitimacy. With its advanced features and capabilities, Phish-It-Out represents a major step forward in the fight against phishing and online fraud.

This Project is developed by Olatunbosun Samuel, Meseh John and Teila Emmanuel who are Final Year Students of Computer Technology at Yaba College of Technology.

Fig 4: The About Page

- *Resource Page*

The resource page contains different resources regarding phishing, such as the definition, types, and techniques of a phishing attack, as well as reference links to the source from which the content where retrieved as shown in figure 5. Also, it contains two (2) sub-section links: the first section can be used to report phishing cases, and the second section consists of some tools and solutions that can also help prevent phishing attacks.



What is Phishing?

Why should i bother knowing what Phishing is all about?

Two Major Types of Phishing Attack

How to Prevent and Protect Against Phishing As An End-user

How to Prevent and Protect Against Phishing As A Company

Fig 5: The Resource Page

IV. RESULT AND DISCUSSION

The prototype for phishing detection, reporting, and evaluation is showcased. Following sections delve into the development tools used, the operational prototype model comprising distinct components, experimentation, performance assessments, evaluation metrics, and a discussion of the outcomes.

The data used for the classifications were collected from www.kaggle.com. The dataset contains over 11,000 web urls that were used for both training and testing of the model and it includes both phishing and legitimate URL.

- *Data Visualization*

Few plots and graphs are displayed in figure 6 and figure 7 to visualize how the data is distributed and how the extracted features are related to each other. Also, figure 8 displays the feature importance graph which shows the different extracted features and their level of importance

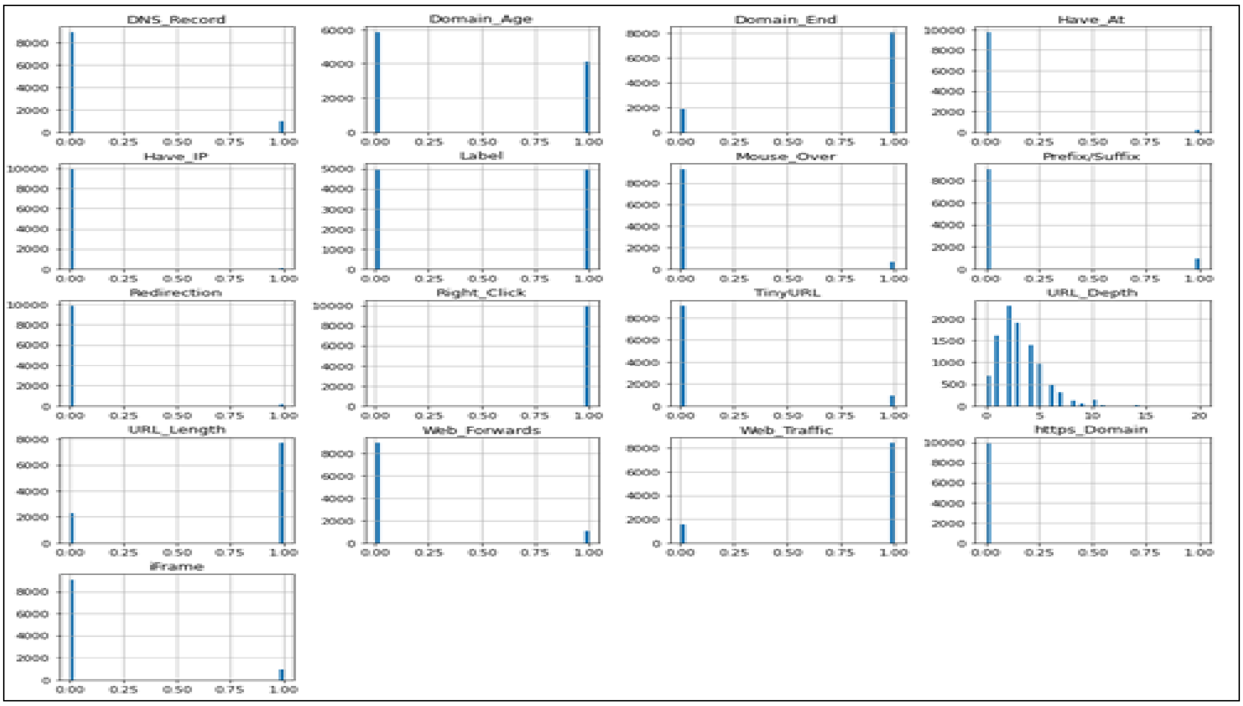


Fig 6: Distribution Plot of the Dataset

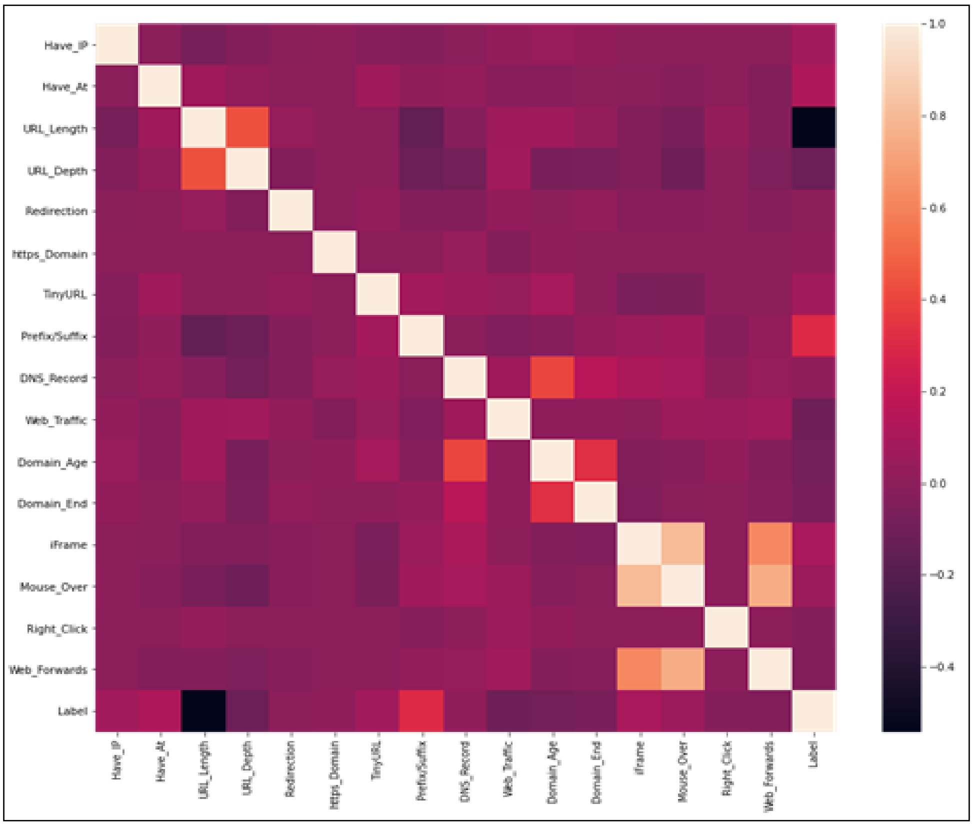


Fig 7: Correlation Heat Map of the Dataset

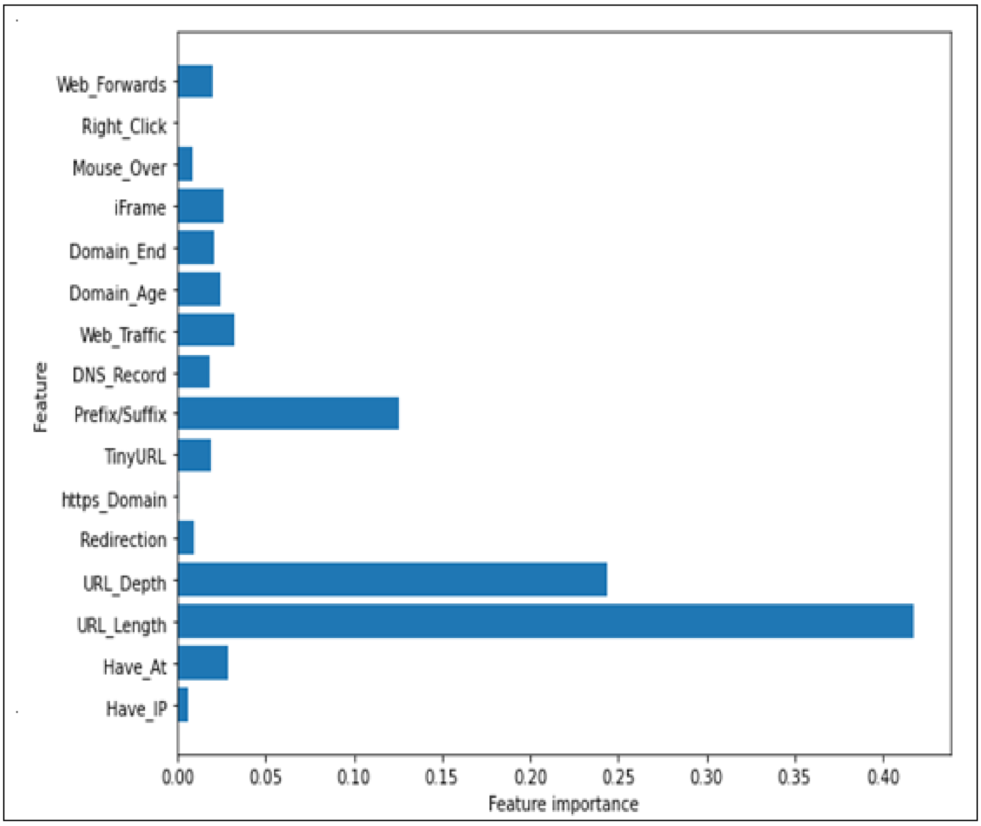
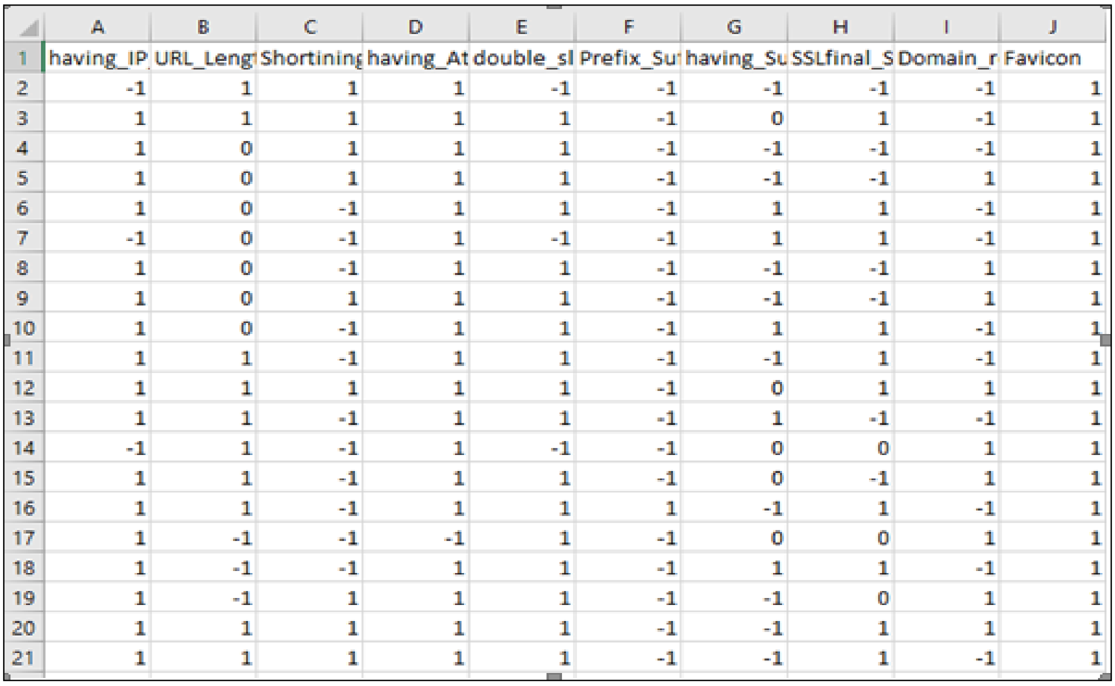


Fig 8: Feature Importance Graph

➢ *Results*

The data used for the classifications were obtained from www.kaggle.com. The dataset contains over 11,000 web urls that were used for both training and testing of the model and it includes both phishing and legitimate URL. Samples of the dataset are shown below in Figure 9



	A	B	C	D	E	F	G	H	I	J
1	having_IP	URL_Leng	Shortining	having_At	double_sl	Prefix_Su	having_Su	SSLfinal_S	Domain_r	Favicon
2	-1	1	1	1	-1	-1	-1	-1	-1	1
3	1	1	1	1	1	-1	0	1	-1	1
4	1	0	1	1	1	-1	-1	-1	-1	1
5	1	0	1	1	1	-1	-1	-1	1	1
6	1	0	-1	1	1	-1	1	1	-1	1
7	-1	0	-1	1	-1	-1	1	1	-1	1
8	1	0	-1	1	1	-1	-1	-1	1	1
9	1	0	1	1	1	-1	-1	-1	1	1
10	1	0	-1	1	1	-1	1	1	-1	1
11	1	1	-1	1	1	-1	-1	1	-1	1
12	1	1	1	1	1	-1	0	1	1	1
13	1	1	-1	1	1	-1	1	-1	-1	1
14	-1	1	-1	1	-1	-1	0	0	1	1
15	1	1	-1	1	1	-1	0	-1	1	1
16	1	1	-1	1	1	1	-1	1	-1	1
17	1	-1	-1	-1	1	-1	0	0	1	1
18	1	-1	-1	1	1	-1	1	1	-1	1
19	1	-1	1	1	1	-1	-1	0	1	1
20	1	1	1	1	1	-1	-1	1	1	1
21	1	1	1	1	1	-1	-1	1	-1	1

Fig 9: Sample of dataset for Detection

➢ *Features Extracted from the Datasets*

The features extracted from the dataset are categorized into: URL & Domain Features, Security & Encryption Features, Source Code & Java script Features, Page Style & Contents Features, Web Address Bar Features and Search Engine-based Features.

The URL & Domain Features are:

- Using IP address: The purpose of the function is to determine whether the input URL contains an IP address or not.
- Abnormal request URL: The purpose of this function is to check if the given domain is present in the given URL. This can be useful in detecting abnormal URLs that may contain unexpected or suspicious domain names.
- Abnormal URL of anchor: This function is used to determine whether the hyperlinks (anchors) in a webpage are safe or not.
- Abnormal DNS record: This check if a given domain has a DNS record or not.
- Domain Registration: The purpose of this function is to determine if the length of time between the current date and the domain registration expiration date is less than one year.
- Age of the domain: This function determines the age of a domain by calculating the difference between the current date and the date the domain was registered.
- Favicons: This check if the favicon of a website is legitimate or not.

V. CONCLUSION

Phishing attacks pose a rapidly growing threat in the cyber world, resulting in significant financial losses for internet users annually. They stand as major threats to individuals, organizations, and public institutions seeking to safeguard their web assets and underlying data. Phishing relies on diverse social engineering tactics to illicitly obtain sensitive information from users. These techniques can be executed through various communication channels, including email, instant messaging, pop-up notifications, and fraudulent web pages. The research was to categorize and recognize how phishers carry out phishing attacks and the different ways in which researchers have helped to solve phishing detection with the use of Support Vector Machine techniques. The system developed is based on a feature extraction algorithm which was used to identify phishing URLs from legitimate URL links and is integrated into a web application where users can input website URL links to detect if it is legitimate or phishing. The feature extraction and the models used on the dataset helped to uniquely identify phishing URLs. The research endeavor holds significant promise for enhancing web security by offering a foundational framework. This framework could potentially serve as a viable solution in the evolution from traditional blacklist-based detection methods to fully integrated phishing detection web services, integrating machine learning approaches.

REFERENCES

[1]. S. Shea, A. S. Gillis, and C. Clark, "What is Cybersecurity?," *Search Secur.*, 2021.

[2]. K. M. Bakarich and D. Baranek, "Something phish-y is going on here: A teaching case on business email compromise," *Curr. Issues Audit.*, vol. 14, no. 1, pp. A1–A9, 2020.

[3]. Razorthorn phishing report https://www.razorthorn.co.uk/wp-content/uploads/2017/01/Phishi ng-S

[4]. K. M. Bakarich and D. Baranek, "Something phish-y is going on here: A teaching case on business email compromise," *Curr. Issues Audit.*, vol. 14, no. 1, pp. A1–A9, 2020.

[5]. D. Gupta and R. Rani, "Improving malware detection using big data and ensemble learning," *Comput. Electr. Eng.*, vol. 86, p. 106729, 2020.

[6]. Microsoft Security Intelligence Report (2019) vol 24 https://www.microsoft.com/security

[7]. G.-G. Geng, Z.-W. Yan, Y. Zeng, and X.-B. Jin, "RRPhish: Anti-phishing via mining brand resources request," in *2018 IEEE International Conference on Consumer Electronics (ICCE)*, IEEE, 2018, pp. 1–2.

[8]. Z. Alkhalil, C. Hewage, L. Nawaf, and I. Khan, "Phishing attacks: A recent comprehensive study and a new anatomy," *Front. Comput. Sci.*, vol. 3, p. 563060, 2021.

[9]. J. VanderPlas, *Python data science handbook: Essential tools for working with data*. " O'Reilly Media, Inc.," 2016.

[10]. N. Bambrick, "Support vector machines: A simple explanation," *línea]. Dispon. en https//www. kdnuggets. com/2016/07/support-vector-machines-simple-explanation. html*, 2018.

[11]. R. Pupale, "Support vector machines (svm)—an overview," *A post Towar. data Sci. available https//towardsdatascience. com/https-medium-compupalerushikesh-svm-f4b42800e989*, 2018.

[12]. K. L. Chiew, K. S. C. Yong, and C. L. Tan, "A survey of phishing attacks: Their types, vectors and technical approaches," *Expert Syst. Appl.*, vol. 106, pp. 1–20, 2018.

[13]. I. Qabajeh, F. Thabtah, and F. Chiclana, "A recent review of conventional vs. automated cybersecurity anti-phishing techniques," *Comput. Sci. Rev.*, vol. 29, pp. 44–55, 2018.

[14]. M. Volkamer, K. Renaud, B. Reinheimer, and A. Kunz, "User experiences of torpedo: Tooltip-powered phishing email detection," *Comput. Secur.*, vol. 71, pp. 100–113, 2017.

[15]. A. Basit, M. Zafar, X. Liu, A. R. Javed, Z. Jalil, and K. Kifayat, "A comprehensive survey of AI-enabled phishing attacks detection techniques," *Telecommun. Syst.*, vol. 76, pp. 139–154, 2021.

[16]. D. M. Y. Beh and R. Bahuang, "Detecting Phishing Uniform Resource Locator (URL) using Machine Learning," *J. Comput. Technol. Creat. Content*, vol. 7, no. 2, pp. 35–41, 2022.

[17]. M. N. Alam, D. Sarma, F. F. Lima, I. Saha, and S. Hossain, "Phishing attacks detection using machine learning approach," in *2020 third international conference on smart systems and inventive technology (ICSSIT)*, IEEE, 2020, pp. 1173–1179.

[18]. P. Dewan, A. Kashyap, and P. Kumaraguru, "Analyzing social and stylometric features to identify spear phishing emails," in *2014 apwg symposium on electronic crime research (ecrime)*, IEEE, 2014, pp. 1–13.

[19]. R. Dhamija, J. D. Tygar, and M. Hearst, "Why phishing works," in Proceedings of the SIGCHI conference on Human Factors in computing systems, pp. 581–590, 2006.

[20]. C. Ludl, S. McAllister, E. Kirda, and C. Kruegel, "On the effectiveness of techniques to detect phishing sites," in International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment, pp. 20–39, Springer, 2007.

[21]. A. P. Rosiello, E. Kirda, F. Ferrandi, et al., "A layout-similarity-based approach for detecting phishing pages," in 2007 Third International Conference on Security and Privacy in Communications Networks and the Workshops-SecureComm 2007, pp. 454–463, IEEE, 2007.

[22]. S. Afroz and R. Greenstadt, "Phishzoo: Detecting phishing websites by looking at them," in 2011 IEEE fifth international conference on semantic computing, pp. 368–375, IEEE, 2011.

[23]. K.-T. Chen, J.-Y. Chen, C.-R. Huang, and C.-S. Chen, "Fighting phishing with discriminative keypoint features," IEEE Internet Computing, vol. 13, no. 3, pp. 56–63, 2009.

[24]. S. Rao, A. K. Verma, and T. Bhatia, "A review on social spam detection: Challenges, open issues, and future directions," *Expert Syst. Appl.*, vol. 186, p. 115742, 2021.

[25]. D. D. Rufo, T. G. Debelee, A. Ibenthal, and W. G. Negera, "Diagnosis of diabetes mellitus using gradient boosting machine (LightGBM)," *Diagnostics*, vol. 11, no. 9, p. 1714, 2021.

[26]. A. K. Dutta, "Detecting phishing websites using machine learning technique," *PLoS One*, vol. 16, no. 10, p. e0258361, 2021.

[27]. H. Nozari and M. E. Sadeghi, "Artificial intelligence and Machine Learning for Real-world problems (A survey)," *Int. J. Innov. Eng.*, vol. 1, no. 3, pp. 38–47, 2021.

[28]. P. C. Sen, M. Hajra, and M. Ghosh, "Supervised classification algorithms in machine learning: A survey and review," in *Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018*, Springer, 2020, pp. 99–111.

[29]. S. Naeem, A. Ali, S. Anam, and M. M. Ahmed, "An unsupervised machine learning algorithms: Comprehensive review," *Int. J. Comput. Digit. Syst.*, 2023.

[30]. S. M. Miraftabzadeh, C. G. Colombo, M. Longo, and F. Foiadelli, "K-means and alternative clustering methods in modern power systems," *IEEE Access*, 2023.

[31]. O. E. Olawade, S. A. Onashoga, and O. Arogundade, “Comparative analysis of machine learning techniques in health system,” in *2020 international conference in mathematics, computer engineering and computer science (ICMCECS)*, IEEE, 2020, pp. 1–6.

[32]. J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, “A comprehensive survey on support vector machine classification: Applications, challenges and trends,” *Neurocomputing*, vol. 408, pp. 189–215, 2020.

[33]. V. Shahrivari, M. M. Darabi, and M. Izadi, “Phishing detection using machine learning techniques,” *arXiv Prepr. arXiv2009.11116*, 2020.

[34]. M. Almseidin, A. A. Zuraiq, M. Al-Kasassbeh, & N. Alnidami, Phishing detection based on machine learning and feature selection methods, *International Association of Online Engineering*, Retrieved July 9, 2023, (2019).

[35]. A. Suryan, C. Kumar, M. Mehta, R. Juneja, and A. Sinha, “Learning model for phishing website detection,” *EAI Endorsed Trans. Scalable Inf. Syst.*, vol. 7, no. 27, pp. e6–e6, 2020.

[36]. S. Naaz, “Detection of phishing in internet of things using machine learning approach,” *Int. J. Digit. Crime Forensics*, vol. 13, no. 2, pp. 1–15, 2021.

[37]. E. Gandotra and D. Gupta, “An efficient approach for phishing detection using machine learning,” *Multimed. Secur. Algorithm Dev. Anal. Appl.*, pp. 239–253, 2021.

[38]. N. M. Shekokar, C. Shah, M. Mahajan, and S. Rachh, “An ideal approach for detection and prevention of phishing attacks,” *Procedia Comput. Sci.*, vol. 49, pp. 82–91, 2015.