Hybrid Machine Learning: A Tool to Detect Phishing Attacks in Communication Networks

Ademola Philip Abidoye[1], Boniface Kabaso[2]
Department of Information Technology
Cape Peninsula University of Technology
Cape Town, South Africa

***Abstract*—Phishing is a cyber-attack that uses disguised email as a weapon and has been on the rise in recent times. Innocent Internet users if peradventure clicking on a fraudulent link may cause him to fall victim to divulging his personal information such as credit card PIN, login credentials, banking information, and other sensitive information. There are many ways in which attackers can trick victims to reveal their personal information. In this article, we select important phishing URLs features that can be used by an attacker to trick Internet users into taking the attacker's desired action. We use two machine learning techniques to accurately classify our data sets. We compare the performance of other related techniques with our scheme. The results of the experiments show that the approach is highly effective in detecting phishing URLs and attained an accuracy of 97.8% with 1.06% false-positive rate, 0.5% false-negative rate, and an error rate of 0.3%. The proposed scheme performs better compared to other selected related work. This shows that our approach can be used for real-time applications in detecting phishing URLs.**

***Keywords*—Phishing attack; data sets; URL classification; phishing URL; attackers; machine learning; classifiers; Internet**

I. INTRODUCTION

In the last decade, Internet usage has been increasing tremendously and makes our lives easy, simple, and transforms our daily lives. It plays a major role in the areas of communication, education, business activities, and commerce [11, 27]. A lot of useful data, information, and knowledge can be obtained from the Internet for personal, organizational, economic, and social development. Positive and productive use of the Internet will assist users to become successful in their careers and businesses. The Internet makes it easy to provide many services online and enables us to access various information at any time, from anywhere around the world. Online banking, including transferring money between accounts, online bills paying, and so on. These services have become very prevalent as more financial institutions start to provide almost free online services. Presently, about 40% of the world population are connected to the Internet [22]. The main purpose of the Internet is to provide worldwide access to various types of data for advancing research in engineering, science, design, and medicine as well as in maintaining global defense and surveillance [7]. However, as more people are using the Internet globally, different kinds of attacks have been identified including denial-of-service and distributed denial of service attacks, drive-by attack, man-in-the-middle attack, password attack, eavesdropping, and phishing attack [30]. Over the last decade, phishing has skyrocketed to staggering proportions and will continue to increase due to various phishing groups using different methods of attacks. Therefore, it is imperative to comprehensively study the mode of operation of attackers. The word phishing comes from the fact that cyber-attackers are fishing for sensitive data and information. The "ph" is coined from the advanced methods the phishers employ to distinguish their activities from the more simplistic fishing. The concept of phishing is a form of social engineering and can be traced back to the early 1990s via America Online (AOL) [8].

Phishing is the act of sending a fake email, messages, or malicious websites to trick the recipient/Internet users into divulging sensitive personal information such as personal identification number (PIN) and password of their bank account, credit card information, date of birth, or social security numbers. To perpetuate this type of attack, the attacker usually poses as a trustworthy organization. For instance, an attacker may send an email that looks like it is from a financial institution or a reliable credit card company requesting for their account information by tricking the target that there is a problem or a need to update his/her within a stipulated time. There were 112163 unique phishing attacks and 60889 unique phishing sites reported in the U.S. in June 2019 [3]. Phishing attacks affect hundreds of thousands of internet users across the globe. Individuals and organizations have lost a huge sum of money and private information through phishing attacks [12].

What differentiates phishing from other Internet attacks is the form the message takes: the attackers disguise as a real person, trusted entity of some kind, or an organization the target might transact business with. It is one of the fastest-growing types of cyber-attack and most widespread due to financial gain the attackers derive from any successful phishing. The attackers capitalize on some recipients' desire to respond to urgent requests from their "financial institutions" by clicking a link or download an attachment provided in a spoofed email that looks "official", but it is linked to a fraudulent website(s) which may result in financial losses, identity theft, or other fraudulent activity.

A. Statistics of Phishing Attacks

The sudden attack of phishing against financial institutions was first known in July 2003. Since then, commercial banks, E-gold, and E-loan are the main target of the phishers. Among financial institutions that have been attacked in the U.S.,

commercial banks account for 91 percent of the attacks while insurance companies account for 7 percent. Similarly, about 39 percent of the total retail banking activities and 25 percent of the credit-card companies have been attacked in 2018 [6].

The number of global phishing attacks rose to 129.9 million during the second quarter of 2019; it increased by 21% more than the same quarter of 2018. Greece has the highest number of phishing attacks at 26.2%, followed by Venezuela, Brazil, Australia, and Portugal. In terms of financial institutions and establishments, commercial banks have the highest percentage of phishing emails at 30.7%, followed by payment systems at 20.1%, worldwide Internet portals at 18%, and social networks at 9% [15]. The act of phishing is not limited to a particular country; it occurs everywhere and every day. The reason is that phishers are using the Internet to phish unsuspecting Internet users for financial gain [9]. Phishing information flow is shown in Fig. 1.

Phishers are looking for more effective and advanced ways to launch phishing attacks. They are developing new techniques for attacks and improving on the old ones. With the advancement in technology, they have refined their attacks both in the usage of websites and emails. They can develop more innovative and effective methods of targeting innocent victims. It is essential to note that different phishers have various methods they use for phishing, but all have similar techniques and tools. These methods can be majorly grouped into three namely impersonation, forwarding, and popups [28].

In recent years, researchers and stakeholders have paid much attention to the problem of phishing and how it could be solved. They have developed different approaches in the literature for detecting malicious uniform resource locators (URLs) and emails. Some of these approaches are presented below.

B. Aim of Research

This work aims to develop a technique that can detect all forms of phishing strategies created by attackers in communication networks. We generate our set of rules which rely on our observations and hybrid machine learning techniques. We gather different methods and tricks used by attackers to entice unsuspecting victims to fabricated web pages and use those attributes to design our rule data sets.

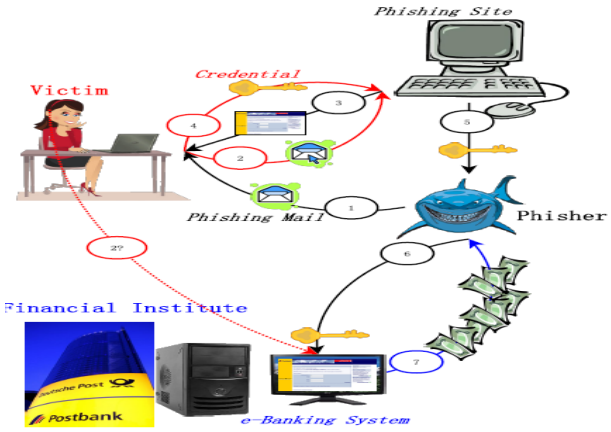

Fig. 1. Phishing Information Flow [20].

C. The Significance of the Study

In recent times, there is an increasing need to identify phishing URLs and emails because of the negative effect they have on their targets. Researchers have developed various methods and applications for exposing phishing websites and detecting malicious emails, but only a few scholars have used machine learning methods for detecting phishing websites. In this study, we are using a hybrid machine learning technique for detecting phishing URLs. A combination of Support Vector Machines (SVM) and Naïve Bayes techniques are used for accurate phishing URLs detection and to minimize false positive detection. This approach provides up to date protection against zero-day phishing attacks.

D. Problem Statement

Phishing detection methods do suffer from low detection accuracy and high positive false alarm, particularly when new phishing techniques are invented. Besides, a blacklist is a common method for detecting phishing URLs but it is ineffective in responding to new phishing attacks since it is now very easy to register a new domain, no comprehensive blacklists can ensure an adequate up-to-date database.

Researchers have developed various approaches to detect phishing websites using different learning algorithms, but this problem still needs more attention of the researchers because new phishing websites are being deployed every day and phishers are using different techniques to lunch their attacks. Consequently, most of the solutions provided for phishing attacks were based on small experimental data sets, the accuracy and effectiveness of these algorithms on real large data sets cannot be ascertained. Thus, the number of malicious websites increases very fast, how to detect phishing websites from a large number of legitimate websites in real-time with high accuracy must also be addressed. It is imperative to design intelligent anti-phishing algorithms that are capable of detecting ever-increasing phishing attacks. A hybrid machine learning technique is used for the detection of phishing URLs. We use both SVM and Naïve Bayes classifiers for the detection since no single classifier is perfect. SVM scales relatively well to high dimensional data, and error can be explicitly controlled. Also, it is very easy to implement. However, it does not scale very well for a large data set. Naïve Bayes classifier is used to overcome the weakness in SVM. This classifier is capable of handling large data sets and scales linearly with the number of predictors and data points.

E. Contributions

This research work uses hybrid machine learning techniques to accurately classify our data sets into either phishing or benign URLs in communication networks. These two classifiers are used together because strengths in one classifier complement the weaknesses in the other classifier. Besides, we use 13 important lexical features to model our classifiers to achieve high precision and to provide a better-accuracy trade-off. We observe that using important lexical features increases the overall classification across all the data sets and minimize the error rate. This shows that the proposed approach can be used for near real-time applications in detecting phishing URLs.

The rest of this article is organized as follows. In Section 2 related work is discussed. Section 3 discusses the proposed approach. Data used for the experiments, relevant features in predicting phishing URLs, and the classifiers used are discussed in this section. In Section 4, we present the various experiments conducted and also discuss the performance evaluation of the two machine learning techniques used. Finally, the conclusion is presented in Section 5.

II. Related Work

Blacklisting and whitelisting are the two widely methods that have been used to manage which entities get access to our system.

A blacklist is a list of suspicious or forbidden URLs that should be blocked or denied access on a network or system. This method is very simple to implement. It is just to deny any strange or suspected URLs access to the network. However, this method is too weak to detect the majority of phishing incidents since new threats are many and constantly emerge every day, such as a zero-day attack. This approach is incapable of detecting or stopping any new kind of attack. It requires keeping a comprehensive list of suspicious websites and their reports which consume a lot of system resources [18]. Phishers sometimes design URLs specifically to evade detection by tools that use a blacklist system. Finally, this approach fails to identify some types of attacks that target a profitable organization.

On the other hand, a whitelist allows several websites to be accessed and blocks other websites that are not on the list. It denies any new URL unless it is proven to be benign (legitimate). Whitelist applications can be used to identify websites by their file name, size, and directory path. Thus, whitelisting access control is higher than blacklisting, as the default is to block websites and allows only those websites that are proven to be legitimate to be accessed. However, its implementation is more complex and hard to assign because it requires more information on the application being used to create the whitelist. Also, it is infeasible to create a whitelist that contains all the list of legitimate sites due to their large number [19]. Another challenge of whitelisting is that a user must remember to check the interface each time he visits any website. Thus, there is a need to develop innovative methods that are capable to detect any recent methods the phishers are using for phishing.

A recent increase in suspicious URLs has attracted the attention of many researchers, and they have developed different techniques for website phishing detection. The definition of phishing constantly changes concerning the way phishing is performed. Email and website are the two major methods the phishers are using for phishing. These two methods have the same goal but there are some differences between the two.

Aburrous et al. [1] proposed an intelligent system for phishing webpage detection in e-banking. They developed a model that combines fuzzy logic with a data mining algorithm to detect phishing websites and categorize the phishing type using 10-fold cross-validation. This model achieved 86.38% grouping accuracy. However, this model has a high percentage of false positive.

Basnet et al. [5] proposed a heuristic-based approach to group phishing URLs by using the data available only on URLs. The authors used a binary classification method to detect phishing URLs and grouped URLs into phishing URLs and legitimate URLs. The results of the experiments show that the proposed approach is very effective in detecting phishing URLs compared to related work. However, this approach is only tested on a data set that is less than 300. It may not be effective on a large data set.

Jain and Richariya [13] developed a new method for detecting phishing emails using link-based features. A prototype web browser was used as a means to process each incoming email to detect a phishing attack. A combination of the prototype and their algorithm assist the system users to be notified of possible attacks and prevent them from clicking any malicious URLs.

Mahmood and Rajamani [21] proposed an anti-phishing detector (APD) technique based on association rule mining for detecting phishing websites. APD dynamically traces out any possible phishing attacks during message transmission between computer users. In addition, the authors developed an algorithm to extract frequently reoccurring words and forward the information to APD for further processing. The results of the approach shown to be effective.

Ajlouni et al. [2] proposed a method for detecting phishing websites based on associative classification algorithms. It is an improvement over [1]. The results of the experiment show that the method achieved 98.5% accuracy in detecting phishing webpages. However, there is no information about how many rules they used for the extraction.

Zhang et al. [32] proposed a new classification method based on a Sequential Minimal Optimization classifier algorithm that consists of features of websites. The results show that the algorithm performs better than the selected baseline. However, this approach can only detect phishing webpages with the Chinese language.

A new rule-based approach for detecting phishing attacks in internet banking is presented in [23]. The authors used two feature sets that have been developed to find webpage identity and support vector machine algorithm to classify webpages. The proposed features are independent web browser history or search engine results. The results of the experiments show that the method can detect phishing webpages with an accuracy of 99.14% true positive and only 0.86% false-negative alarm.

Ramesh et al. [25] developed a method for detecting phishing webpages. The webpage is scrutinized and classified as indirect and direct links associated with the page. Indirect link features are extracted from the search engine result while direct links are extracted from the page contents. Also, they used a third-party DNS lookup to match the domains of a malicious webpage and phishing target to the corresponding IP address. The results of this approach achieve 99.62% accuracy. However, the efficiency of this method depends on largely the speed of search engine and DNS lookup time

which can affect its performance. A comparison of the related studies that have been used to detect phishing URLs in the literature with our work is presented in Table I.

TABLE I. EVALUATION OF RELATED WORK WITH PROPOSED APPROACH

Work	Approach	A	B	C	D
[1]	Fuzzy logic	No	No	Yes	Yes
[5]	Binary clarification	Yes	No	Yes	Yes
[13]	Web browser	No	No	Yes	Yes
[21]	Rule-based (APD)	No	No	Yes	Yes
[2]	Data mining	Yes	Yes	Yes	no
[32]	Sequential Minimal Optimization	No	Yes	No	Yes
[23]	Rule-based approach	Yes	No	No	No
[25]	Domain identification	Yes	Yes	Yes	No
	Proposed approach	Yes	Yes	Yes	Yes

where A = Zero-day phishing detection
B = 3rd-party services' Sovereignt
C = Search engines sovereignty
D = Language sovereignty

III. PROPOSED APPROACH

In this section, we present in detail our method for detecting malicious URLs. The approach is divided into two parts, and each part's output is an input to the next part as shown in the proposed framework in Fig. 2.

The first part is based on data collection, processing of data sets, and URLs feature extraction. We consider different heuristic features in the structure of URLs, ranging from a generic social engineering feature, lexical feature in the URL, multiple alphabets, and phishing target brand name. The feature vector is constructed with 13 important features to model our classifiers. The second part is based on the classification of data set using a hybrid of machine learning classifiers to evaluate our approach. -We performed different experiments. The results of the experiment show that our scheme achieves 97.8% accuracy on average. The description of each part is briefly discussed in the following subsections.

A. Processing of Data Sets and URLs Features Extraction

A large number of data sets (36,874), discussed in sub-Section 3.1, were collected and processed to make them suitable for the requirement of this study. The processing involved many stages, these include webpages feature extraction, data standardization, and attribute weighing. These steps are very important so that the classifiers would be able to understand the data sets and appropriately categorize them into their classes. The classifier is regularly trained with new phishing web pages to learn new trends in phishing. The outcome of this phase is used as input to the next part of the appropriate classifiers.

We propose a hybrid machine learning approach to effectively classify phishing URLs based on the information available to an individual URL. Phishing URLs are treated as a binary classification problem with the benign URLs belong to the negative class and phishing URLs belonging to the positive class. We collected our phishing and benign URLs from PhishTank, Yahoo directory, and the Google engine to form our data sets. Thereafter, we extract many features that have proved to be effective in predicting phishing URLs by employing different publicly available resources to classify the data sets into their respective classes. We apply both SVM and Naïve Bayes algorithms to create models from training data sets which consist of feature extractions and class labels. Fig. 2 shows the proposed framework for phishing URLs detection.

We use two types of data sets for this research. The first set is phishing data sets and the other one is benign data sets. The data sets are collected from different credible sources [10, 24],

The data sets contain 36874 URLs with their related features. We wrote Python scripts code to automatically download certified phishing URLs from PhishTank.

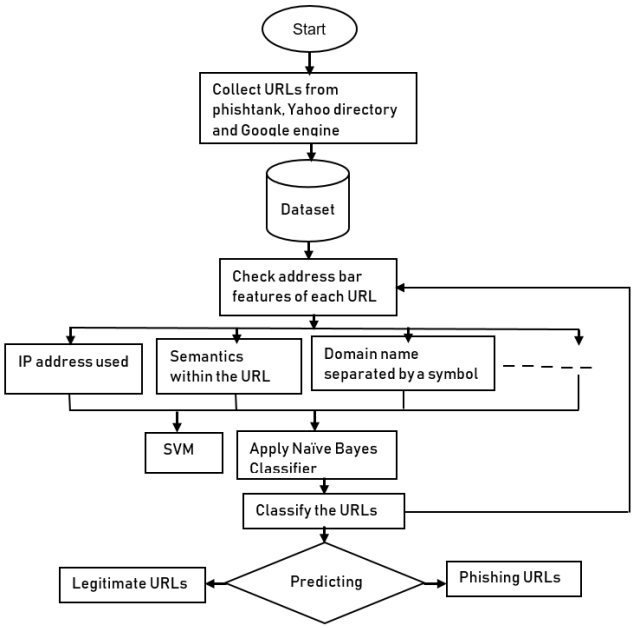

Fig. 2. The Proposed Framework for Detecting Phishing URLs.

B. Phishing Data Sets

PhishTank is a joint project to which people can submit suspicious phishing URLs for confirmation. It is a public clearinghouse for phishing URLs [4]. Suspicious URLs are further scrutinized by many people before being confirmed as phishing URLs and added to a blacklist. PhishTank provides a comprehensive list of current and active phishing URLs.

Researchers and developers can download phishing URLs from the Phishing Web site after signing up. They would be able to download the URLs from PhishTank in different file formats with an API key.

We downloaded two sets of phishing URLs. The first data-set is referred to as DTS1, contains 14,298 phishing URLs. They were collected from March 4, 2019, to April 19, 2019, based on the reports in [26] which shows that phishing attacks are usually higher during this period than the preceding

months. Also, we observe that phishers constantly develop new tactics to get personal information from unsuspecting users, to explore various and recent methods the attackers are using motivated us to collect the second sets of data. The second set of data, referred to as DTS2, contains 7,350 phishing URLs. They were collected from November 1 to December 4, 2019. We chose this period because it has a special day "Black Friday" (November 29, 2019) in which many people have been waiting for to buy cheap goods from stores, online using their credit or debit cards. Phishers also use this period as an opportunity to display their tactics and launch different attacks on unsuspecting users. A total of 21,648 phishing URLs was collected from the PhishTank Web site.

C. Legitimate Data Sets

Our benign URLs were collected from the Yahoo directory. Yahoo provides a generator that arbitrarily produces an URL in its directory each time the Web page is visited. This service is used to randomly choose an URL and download the contents of the Web page with the server header information. This service is used to collect 9,045 random URLs from May 6, 2019, to June 10, 2019. Our list consists of URLs from financial institutions, e-commerce, online services, cloud storage, religious organizations to get different URL structures and Web page contents [16]. To provide more learning instances for legitimate URLs, we chose 6,181 legitimate URLs from the Open Directory Project (DMOZ) Web directory [29]. DMOZ is a multilingual open-content directory of World Wide Web links containing more than three million URLs.

We use a Google tool to analyze the list of benign URLs collected and crawled. These URLs are used as legitimate webpages based on the assumption that all the URLs extracted were benign since they were downloaded from legitimate Internet sources.

Python and Java scripts are used to parse the legitimate and phishing URLs and extract the features discussed in subsection 3.2. Web pages that we could not extract features from their contents were discarded to get only valid URLs for our data sets. The total number of our data sets is presented in Table II.

D. Data Authentication

Data sets collected need to be authenticated to ascertain the real status of the URLs, particularly in the case of phishing websites as it is known that the phishing website only lasts a few weeks [31]. Thus, every URL needs to be authenticated before processing.

In this section, we present relevant features that are effective in predicting phishing web sites. Each feature is discussed with its associated rules.

TABLE II. DATASETS FOR PHISHING URLS DETECTION

Data set	Phishing	Non-phishing	Total data sets
DTS1	14,298	9,045	23,343
DTS2	7,350	6,181	13,531
DTS1 + DTS2	21,648	15,226	36,874

A generic social engineering feature

Phishers use generic greetings in their messages such as "Sir", "Dear Bank Customer", "Dear Customer", and "Dear Member" to address their target victims. The content of the message is always threatening such as "please update your bank account to prevent it from being blocked", "Your account has been compromised!", "Urgent action required!", "Your account will be closed!" These intimidation strategies are becoming more common than the promise of "instant riches"; taking advantage of victims' anxiety and concern to get them to provide their personal information.

$$Rule \begin{cases} \textbf{if } \textit{the greeting is directed to account owner and do not require to supply} \\ \text{a piece of } \textit{personal information via a link in the message} \rightarrow \textit{Legitimate} \\ \textbf{else if } \text{the } \textit{greeting is generic} \rightarrow \textit{Suspicious} \\ \textbf{else } \textit{update your information via a given link} \rightarrow \textit{Phishing} \end{cases}$$

Lexical features explain lexical patterns of phishing URLs such as long IP addresses, special characters, number of dots, and so on.

IP-based URL

Internet Protocol (IP) address is one of the ways to hide the webpage address. If an IP address is used instead of a Domain Name System (DNS) address in the URL, it will be difficult for innocent users to ascertain where they are being directed to when they click the link or press the Enter key on their system to load the page. Another reason for using the IP address is that phishers would not like to spend money to buy a domain for their phony web pages.

$$Rule: \begin{cases} \textbf{If } \textit{the domain name has an IP Address} \rightarrow \textit{Phishing} \\ \textbf{else} \rightarrow \textit{Legitimate} \end{cases}$$

Long URL to hide the fake part

Attackers can use lengthy URLs to mask the fake part in the address bar. For instance,

"http://prudentbank.com/2k/ab51e2e319e51502f416dbe46b773a5e/?cmd=_home&dispatch=11004d58f5b74f8dc1e7c2e8dd4105e811004d58f5b74f8dc1e7c2e8dd4105e8@phishing.net.html"

We computed the length of URLs in our data sets and determined their average length to ensure the accuracy of our research. The findings showed that if the URL length is less than 52 characters, it is classified as legitimate; it is suspicious if the length is between 52 and 73 characters, and it is a phishing URL if the URL is more than 73 characters. A method based on frequency has been used to update this feature rule, which improves its accuracy.

$$\text{Rule: } \begin{cases} \textbf{If } \textit{URL length} < 52 \textit{ characters} \rightarrow \textit{Legitimate} \\ \textbf{else if } \textit{URL length} \geq 52 \textit{ and } \leq 73 \textit{ characters} \rightarrow \textit{Suspicious} \\ \textbf{else} \rightarrow \textit{Phishing} \end{cases}$$

Shortened URL "TinyURL"

Short URL enables to reduce long links from social networks and top sites on the Internet. This is achieved by the service provider through an "HTTP Redirect" on a domain name that is short and redirects to the corresponding long URL [17]. For instance, an URL for Wiki's article "http://en.wikipedia.org/wiki/URL_shortening" contains 64 characters and its corresponding short URL http://bit.ly/c1htE; it contains 16 characters with Bitly's default domain name "bit.ly" and the hash "c1htE" as the back-half. A hash only consists of letters and numbers "a-z, A-Z,0- 9". Attackers use this shortened URL feature to hide links to infected websites or phishing.

$$\text{Rule: } \begin{cases} \textbf{if } TinyURL \rightarrow Phishing \\ \textbf{else} \rightarrow Legitimate \end{cases}$$

URL's having "@" Symbol

Using "@" symbol within the URL causes the Web browser to read the right side of the browser address and ignore everything preceding the "@" symbol. For instance, in this URL www.prudentbank.com@www.google.com, the browser will ignore "www.prudentbank.com" and only read www.google.com which it may be used to hide a phishing URL.

$$\text{Rule: } \begin{cases} \textbf{if } URL\ having\ @\ symbol \rightarrow Phishing \\ \textbf{else} \rightarrow Legitimate \end{cases}$$

Hovering of a Mouse over Hyperlink Feature

One of the tactics of phishers is that they use legitimate domain names for their links to send messages to their potential victims while the destination URLs are hidden from them using HTML code. For instance, a phisher may send this link www.prudentbank.com to unsuspecting Internet users which looks like a Prudent Bank Website whereas the destination URL "http://phishing.com" is hidden from the user. If the user clicks the link "www.prudentbank.com" it will take him to "http://phishing.com" thinking that they are surfing a legitimate website. To check if a link is malicious or not, a mouse is hovering over the link to view the destination URL.

$$\text{Rule: } \begin{cases} \textbf{if } destination\ URL\ is\ the\ same\ with\ the\ domain\ name\ and\ the\ link\ leads\ to\ the\ homepage \rightarrow Legitimate \\ \textbf{else if } the\ destination\ URL\ cannot\ be\ determined \rightarrow Suspicious \\ \textbf{else } the\ destination\ URL\ does\ not\ the\ same\ with\ the\ domain\ name \rightarrow Phishing \end{cases}$$

Redirecting using "//"

The presence of "//" in the URL path shows that an innocent user will be redirected to another infected website. For example, http://www.legitimate.com//http://www.phishing.com.

This study examines the position of "//" in a legitimate URL. If the URL begins with "http" then "//" should appear in the 6th position and the 7th position if it begins with "https".

$$\text{Rule: } \begin{cases} \textbf{if } the\ position\ of\ \text{"//"}\ in\ the\ URL > 7 \rightarrow Phishing \\ \textbf{else} \rightarrow Legitimate \end{cases}$$

Domain name separated by a dash symbol

It is very rare for a legitimate domain name to be separated by a dash symbol (-). Phishers use this method to trick Internet users by adding a dash symbol (-) within the domain name so that users will think that they are surfing a legitimate webpage. For instance, http://www.pay-pal.com/.

$$\text{Rule: } \begin{cases} \textbf{if } Dash\ symbol\ (-)\ is\ part\ of\ a\ domain\ name \rightarrow Phishing \\ \textbf{else} \rightarrow Legitimate \end{cases}$$

Subdomain of a subdomain

A URL might include an Internet country code top-level domain (ccTLD) to identify a particular country. For instance, http://www.prudentbank.com.za/login/. "za" is a ccTLD, and the ".com" portion of the extension shows that the domain name is a commercial entity. Taking the two extensions together ". com.za" is called a second-level domain (2LD) and "prudent bank" is the real domain name. To minimize rules for extracting this feature, first, we remove subdomain "www" from the URL and ccTLD if the extension is part of the URL. Thereafter, the number of dots in the URL is counted. If the number of dots is one, then the URL is legitimate. It is suspicious if the number of the dots is two since the URL has one subdomain. It is declared phishing if the number of dots is more than two since it will contain many subdomains.

$$\textit{Rule: } \begin{cases} \textbf{if } \text{the } number\ of\ dots\ in\ domain\ portion = 1 \rightarrow Legitimate \\ \textbf{else if } dots\ in\ domain\ portion = 2 \rightarrow Suspicious \\ \textbf{else} \rightarrow Phishing \end{cases}$$

A domain name containing multiple alphabets

It is possible to register domain names in other alphabets such as Chinese, Arabic, French, German, or anything that can be represented with the Unicode standard since 1998. Phishers have taken advantage of this unique feature by finding characters in other alphabets which look similar to the Latin ones to lure users into a phishing website. For instance, in this URL "https://apple.com", the domain name can be registered with "xn--pple-43d.com". The URL is equivalent to "https:// xn--pple-43d.com". Thus, most users will fall for this trick because their browsers will show the green padlock icon, showing that the user is on a secure connection but in fact, a bunch of Cyrillic characters is embedded within the multiple alphabets.

$$\text{Rule: } \begin{cases} \textbf{if } domain\ name\ containing\ multiple\ alphabets \rightarrow Phishing \\ \textbf{else} \rightarrow Legitimate \end{cases}$$

Phishing website longevity

We believe that legitimate websites will be hosted and regularly paid for one or more years in advance. It has been shown that a phishing website exists for a short period to avoid being detected [14]. In our data sets, the longest fake domains that have been used are only for six months.

Rule: $\begin{cases} \boldsymbol{if}\ domains\ expire \leq\ six\ months\ \rightarrow\ Phishing \\ \boldsymbol{else} \rightarrow\ Legitimate \end{cases}$

IV. Detection of Phishing URLs

We use a hybrid machine learning classification techniques in detecting phishing URLs. A feature vector matrix is built from our data sets presented in Table I. Each vector-matrix consists of 13 important lexical features described above. We use two variables to classify the data sets: -1 for a legitimate URL and 1 for a phishing URL as shown in equation (1). This gives a feature matrix-vector of 36,874 denoting the total number of the data sets.

There are many machine learning classification algorithms, we classified our data sets using the following classification algorithms. Metrics for classification are discussed thereafter.

A. Support Vector Machines (SVMs) Classifiers

In any classification process, both a parameter and a model technique should be chosen to achieve a high level of performance of the machine learning. Recent methods enable different kinds of models of varying complexity to be selected.

This study uses a linear classifier of the form: $f(X_i) = W.X_i + b$ where . represents the dot product, W denotes the weight vector, X_i is input data, and b denotes a learned bias vector.

Let $\{X_i\}$ denote the features of our data sets for all $i = 1,\ 2,3, \ldots, n,\quad X_i \in \mathbb{R}^d$, and $y_i \in \{-1, 1\}$ denote class labels (indicator variable). Our goal is to classify the data sets correctly. The following mathematical equations need to be satisfied to achieve this goal as shown in equation (1). SVM data sets classification is contained in Algorithm 1.

$$f(X_i) = \begin{cases} \geq 0 & y_i = +1 \\ < 0 & y_i = -1 \end{cases}$$

$$.X_i + b \geq 1 \qquad (1)$$

$$W.X_i + b < 1$$

$$y_i(W.X_i + b) \geq 1, \text{ for all } i$$

B. Naïve Bayes Classifiers

Naive Bayes classifiers are a group of classification algorithms based on Bayes' Theorem. The underlying assumption of these classifiers is that all the features used for the classification are autonomous of each other. In other words, it assumes that the existence of a specific feature in a data set is unrelated to the existence of any other feature. The Bayes can consider all the features of data sets and correctly classify them. It provides a way of determining posterior probability $P_r(y|X_i)$ from $P_r(X_i)$, $P_r(y)$, and $P_r(X_i|y)$ as shown in equation (2).

Fig. 3 shows the process of experimenting before arriving at our results.

$$P_r(y|X_i) = \frac{P_r(X_i|y) * P_r(y)}{P_r(X_i)} \qquad (2)$$

Above,

$P_r(y|X_i)$ is defined as the posterior probability of class (legitimate or phishing URL) given the predictor (feature).

$P_r(X_i)$ is the probability of a predictor.

$P_r(y)$ is the probability of the class.

$P_r(X_i|y)$ is the probability of the predictor given class.

The variable $y = y_k$ denote the class defined above and variable X_i denote the features of our data sets such that

$$X_i = (X_1, X_2, X_3, \ldots \ldots, X_n)$$

Substituting for X_i in equation (3) and expanding using the chain rule

$$P_r(y|X_1, X_2, .., X_n) = \frac{P_r(X_1|y)P_r(X_1|y) \ldots\ldots P_r(X_n|y)P_r(y)}{P_r(X_1)P_r(X_2) \ldots\ldots P_r(X_n)} \qquad (3)$$

The value of the denominator remains static for all values in our data set. Thus, the denominator is eliminated and proportionality is introduced as follows.

$$P_r(y|X_1, X_2, .., X_n) \propto P_r(y) \prod_{i=1}^{n} P_r(X_i|y) \qquad (4)$$

The above function is further used to classify our data sets, X_i, into two classes: legitimate or phishing URLs. Model in Fig. 3 is developed to classify the data sets.

Algorithm 1: SVM Data Classification

Begin

1 Given a hyperplane $W.X + b$

2 $f(X_i) = W.X_i + b$ for all $i = 1,2,3, \ldots \ldots n$

3 The classifier can be expressed as

4 $f(X_i) = \widetilde{W}.\widetilde{X}_i + w_o = W.X_i$

5 where $W = (\widetilde{W}, w_o),\ X_i = (\widetilde{X_1},\ 1)$

6 Let $W = 0$

7 Considering the data sets and class labels, $\{X_i, y_i\}$

8 $f(X_i) = sign\ (\sum w[i]\, x[i] + b)$

9 **if** X_i is wrongly classified **then** $W \leftarrow W + \beta * sign(W.X_i + b)$

10 **Else**

11 Continue until all the data sets are correctly classified

12 **end if**

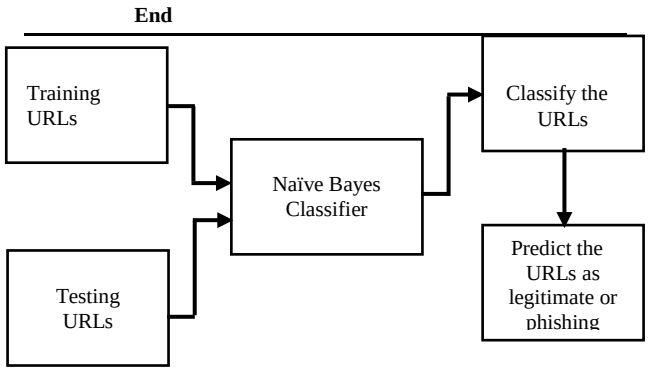

Fig. 3. The Proposed Model to Detect Phishing Attacks.

C. Metrics used for Evaluation

The following metrics are used for evaluation of the proposed scheme to eliminate or minimize misclassification in our data sets. We assume that a legitimate website is negative and a phishing website as positive. include i) True positive rate (TPR) also called sensitivity, ii) False positive rate (FPR) also called specificity, iii) true negative rate (TNR), and false-negative rate (FNR). These prediction outcomes are summarised in Table III.

TABLE III. PREDICTION OUTCOMES FOR PHISHING URLS DETECTION

Expected class			
Classes		True	False
	True	True positive (TP)	False-positive (FP)
	False	False-negative (FN)	True negative (TN)

*True positive rate (*Sensitivity*)*: It is defined as the proportion of legitimate websites that are correctly classified as legitimate. It is mathematically expressed as follows.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad (5)$$

*False-negative rate (*Specificity*):* FN is defined as the proportion of phishing websites that are correctly classified as phishing.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \qquad (6)$$

False-positive rate (FPR): It is defined as the proportion of phishing websites that are wrongly classified as legitimate websites. It is mathematically expressed as follows.

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \qquad (7)$$

True negative rate (TNR): It is defined as the proportion of legitimate websites that are wrongly classified as phishing websites. It is mathematically expressed as follows.

$$\text{TNR} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad (8)$$

Accuracy: Accuracy (ACC) is determined as the number of all correct predictions divided by the total number of the dataset. It is mathematically expressed as follows.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FN+ FP}} = \frac{\text{TP+TN}}{\text{P+N}} \qquad (9)$$

Error rate: Error rate (ERR) is determined as the number of all wrong predictions divided by the total number of the dataset. It is mathematically expressed as follows.

$$\text{Error rate} = \frac{\text{FP+FN}}{\text{TP+TN+FN+ FP}} = \frac{\text{FP+FN}}{\text{P+N}} \qquad (10)$$

V. EXPERIMENTS

To evaluate the proposed scheme, we used two machine learning techniques: Support Vector Machines (SVM) and Naïve Bayes to classify our train data sets into two classes. Many experiments were performed on the data sets to test whether the input URLs are malicious or benign. The URLs were entered into the python program and extracted the URLs features. The results of the classification are presented in Table IV. The table shows the 5th percentile, 95th percentile, median, and standard deviation (SD) values for the *Accuracy* of each classifier for four different number of runs using all the important features discussed above.

A. Analysis and Discussion of Results

To test the accuracy of the algorithms, we obtained the following experimental results and present them in a tabular form as shown in Table IV.

Also, we conducted more experiments on the classification of the URLs. Fig. 4 shows the graphical representation of phishing and benign values for the next experiment. A total number of 18108 URLs are phishing and 1892URLs are benign.

Moreover, Fig. 5 shows the graphical representation of phishing and benign values. A total number of 22897 URLs are for phishing and 2103 are benign.

Similarly, Fig. 6 shows the graphical representation of phishing and benign values. A total number of 23851 URLs are phishing and 6149 are benign.

TABLE IV. EXPERIMENTAL RESULTS OF THE PHISHING CLASSIFIERS

Experiment	**URLs**	**Phishing**	**Benign**
Exp1	1000	991	9
Exp2	2000	1987	13
Exp3	3000	2947	53
Exp4	4000	3850	150
Exp5	5000	4766	234
Exp6	6000	5683	317
Exp7	7000	6708	292
Exp8	8000	7671	329
Exp9	9000	8518	482
Exp10	10000	9376	624
Exp11	11000	10431	569
Exp12	12000	11498	502
Exp13	13000	12602	398
Exp14	14000	13255	745
Exp15	15000	13943	1057

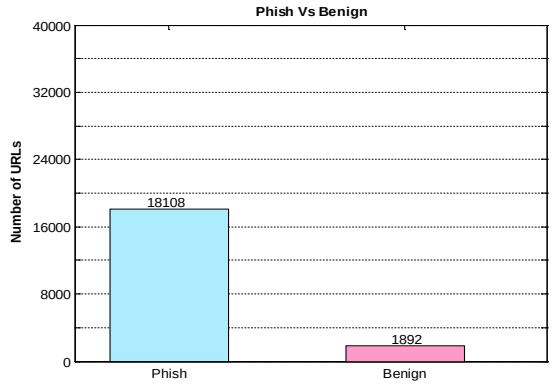

Fig. 4. Graphical Classification for 20,000 URLs.

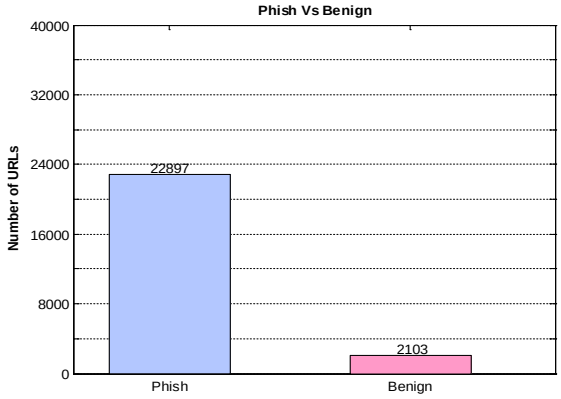

Fig. 5. Graphical Classification for 25,000 URLs.

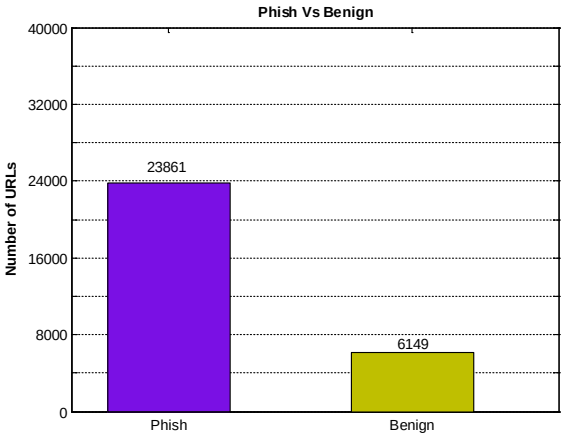

Fig. 6. Graphical Classification for 30,000 URLs.

Finally, Fig. 7 shows the graphical representation of phishing and benign values. A total number of 27629 URLs are phishing and 7371 are benign.

In order to provide further information about confidence intervals of URLs classification, each classifier runs for 100, 150, 200, and 250. Table V shows the 5th percentile, 95th percentile, median, and standard deviation (SD) values for the accuracy of each classifier.

More experiments were performed to ascertain which malicious schemes and attack methods are successful at tricking innocent Internet users to reveal personal information. We use 30 phishing features and randomly distributed them across 40 phishing URLs from our data sets. Thus, one phishing feature could be in many phishing URLs; similarly, one phishing URL could have one or more features. The results of the experiments are presented in Table VI.

We observe that the “Anomalous Request URL” featured in all the selected 40 phishing URLs having a 100% appearance. In addition, spelling errors are 85% having appeared 34. It shows that most of the messages sent by the attackers to innocent users have spelling errors. However, the "Disabling right-click button" feature has the highest percentage (7.5%) with 3 appearances. We ensured that every phishing feature had featured at least once in all the selected phishing URLs.

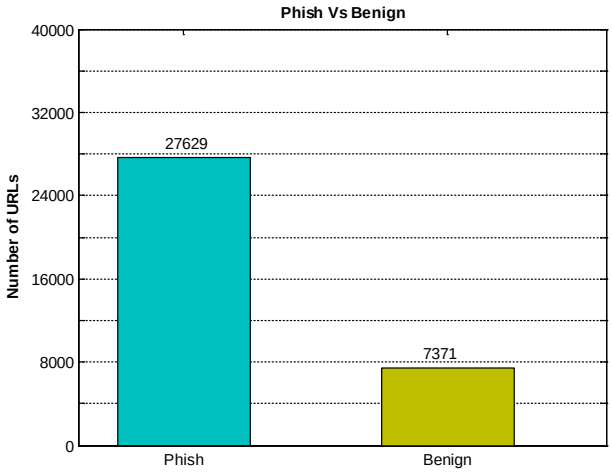

Fig. 7. Graphical Classification for 35,000 URLs.

TABLE V. CLASSIFICATION RESULTS FOR THE CLASSIFIERS

Number of Runs	Classifier	5th Percentile	95th Percentile	Median	SD
100	SVM	95.25	97.31	96.78	0.31
	Naïve Bayes	96.37	98.42	97.81	0.27
150	SVM	92.95	94.10	93.45	0.42
	Naïve Bayes	95.07	95.29	94.62	0.38
200	SVM	89.09	90.73	90.39	0.57
	Naïve Bayes	91.51	93.48	94.62	0.49
250	SVM	86.37	88.06	87.41	0.67
	Naïve Bayes	89.28	91.50	90.59	0.61

TABLE VI. PHISHING FEATURE INDICATORS

Lexical features	No. of appearance	Percentage of appearance (%)
IP-based URL	23	57.5
Long URL to hide the fake part	28	70.0
Shortened URL	7	17.5
URL's having "@" Symbol	9	22.5
Using forms with the 'Submit' button	5	12.5
Hovering of a Mouse over Hyperlink	21	52.5
Spelling errors	34	85.0
Redirect pages	29	72.5
Anomalous Request URL	40	100.0
Domain name separated by a dash symbol	13	32.5
Subdomain of a subdomain	28	70.0
Copying Website	15	37.5
Anomalous cookie	7	17.5
Website Traffic	5	12.5
Domain name having multiple alphabets	11	27.5
1.1.1.1 Phishing website longevity	25	62.5
Generic salutation	31	77.5
Pharming attack	6	15.0
Using Non-Standard Port	18	45.0
URL of Anchor	14	35.0
Disabling right-click button	3	7.5
Adding Prefix or Suffix	12	30.0
Status Bar Customization	16	40.0
Age of Domain	23	57.5
Google Index	19	47.5
Server Form Handler (SFH)	5	12.5
Number of Links Pointing to Page	17	42.5
Using Hexadecimal Character Codes	12	30.0
Replacing Similar Characters for URL	21	52.5
Using the pop-up window	7	17.5

VI. CONCLUSION AND FUTURE WORK

Phishing is a type of social engineering attack often used to steal user personal information. In this project, we explore several tactics in which phishers use to trick innocent Internet users into divulging their personal information. We added new features to our design and in addition to some important features, we identified in the literature. An efficient approach is developed for detecting malicious URLs. Hybrid machine learning algorithms are used to classify our data sets. Several experiments were performed to determine the efficiency of our scheme. These experiments showed better performance and achieved a classification accuracy of 97.8% with a low false-positive rate of 1.06%.

In the future, we would consider more machine learning algorithms to compare their accuracy and false-positive rates.

ACKNOWLEDGMENTS

This work is supported by the Centre for Postgraduate Studies, Cape Peninsula University of Technology, Cape Town, South Africa.

DECLARATION OF INTERESTS

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

AUTHORS' CONTRIBUTIONS

All authors contributed and approved the final manuscript.

DATA AVAILABILITY

The raw data of the IoT devices used to support the findings of this study are available from the corresponding author upon request.

CONFLICTS OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this paper.

REFERENCES

[1] M. Aburrous, M. A. Hossain, K. Dahal, and F. Thabtah, "Intelligent phishing detection system for e-banking using fuzzy data mining," Expert systems with applications, Vol.37, No.12, pp.7913-7921, 2010.

[2] M. Ajlouni, W. e. Hadi, and J. Alwedyan, "Detecting phishing websites using associative classification," image, Vol.5, No.23, pp.36-40, 2013.

[3] APWG. (2019, November 13,). Anti-Phishing Working Group Phishing Activity Trends Report. Available: https://docs.apwg.org/reports/apwg_trends_report_q2_2019.pdf

[4] R. B. Basnet, A. H. Sung, and Q. Liu, "Feature selection for improved phishing detection," in Proceedings of the International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, 2012, pp.252-261.

[5] "Learning to detect phishing URLs," International Journal of Research in Engineering and Technology, Vol.3, No.6, pp.11-24, 2014.

[6] A. Bouveret, Cyber risk for the financial sector: a framework for quantitative assessment: International Monetary Fund, 2018.

[7] M. Büchi, N. Just, and M. Latzer, "Caring is not enough: the importance of Internet skills for online privacy protection," Information, Communication & Society, Vol.20, No.8, pp.1261-1278, 2017.

[8] J. A. Chaudhry, S. A. Chaudhry, and R. G. Rittenhouse, "Phishing attacks and defenses," International Journal of Security and Its Applications, Vol.10, No.1, pp.247-256, 2016.

[9] K. L. Chiew, K. S. C. Yong, and C. L. Tan, "A survey of phishing attacks: their types, vectors, and technical approaches," Expert Systems with Applications, Vol.106, pp.1-20, 2018.

[10] L. M. Ellram and W. L. Tate, "The use of secondary data in purchasing and supply management (P/SM) research," Journal of purchasing and supply management, Vol.22, No.4, pp.250-254, 2016.

[11] M. Graham and W. H. Dutton, Society and the internet: How networks of information and communication are changing our lives: Oxford University Press, 2019.

[12] J. Hong, "The current state of phishing attacks," 2012.

[13] A. Jain and V. Richariya, "Implementing a web browser with phishing detection techniques," arXiv preprint arXiv:1110.0360, 2011.

[14] L. James, Phishing exposed. Canada.: Syngress, 2005.

[15] Kaspersky. (2019, November 25). How to protect yourself against spam email and phishing. Available: https://www.kaspersky.co.za/resource-center/threats/spam-phishing

[16] J. LaCour, "Phishing Trends and Intelligent Report," 2019.

[17] S. Le Page, G.-V. Jourdan, G. v. Bochmann, J. Flood, and I.-V. Onut, "Using url shorteners to compare phishing and malware attacks," in Proceedings of the 2018 APWG Symposium on Electronic Crime Research (eCrime), 2018, pp.1-13.

[18] L.-H. Lee, K.-C. Lee, H.-H. Chen, and Y.-H. Tseng, "Poster: Proactive blacklist update for anti-phishing," in Proceedings of the Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, 2014, pp.1448-1450.

[19] L. Li, E. Berki, M. Helenius, and S. Ovaska, "Towards a contingency approach with whitelist-and blacklist-based anti-phishing applications: what do usability tests indicate?," Behaviour & Information Technology, Vol.33, No.11, pp.1136-1147, 2014.

[20] S. Li and R. Schmitz, A novel anti-phishing framework based on honeypots: IEEE, 2009.

[21] M. A. Mahmood and L. Rajamani, "APD: ARM Deceptive Phishing Detector System Phishing Detection in Instant Messengers Using Data Mining Approach " Global Trends in Computing and Communication Systems, Vol.269, No.1, pp.490-502, 2011.

[22] Miniwatts Marketing Group. (2019, November 11,). Internet World Starts. Available: https://www.internetworldstats.com/stats.htm

[23] M. Moghimi and A. Y. Varjani, "New rule-based phishing detection method," Expert systems with applications, Vol.53, pp.231-242, 2016.

[24] PhishTank. (2019, November 25,). Statistics about phishing activity and PhishTank usage. Available: http://www.phishtank.com/stats.php

[25] G. Ramesh, I. Krishnamurthi, and K. S. S. Kumar, "An efficacious method for detecting phishing webpages through target domain identification," Decision Support Systems, Vol.61, pp.12-22, 2014.

[26] H. N. Security. (2019). Phishing attacks at highest level in three years. Available: https://www.helpnetsecurity.com/2019/11/07/phishing-attacks-levels-rise/

[27] E. Soegoto and M. Rafi, "Internet role in improving business transaction," in Proceedings of the IOP Conference Series: Materials Science and Engineering, 2018, pp.012059.

[28] V. Suganya, "A review on phishing attacks and various anti phishing techniques," International Journal of Computer Applications, Vol.139, No.1, pp.20-23, 2016.

[29] R. Verma and A. Das, "What's in a url: Fast feature extraction and malicious url detection," in Proceedings of the Proceedings of the 3rd ACM on International Workshop on Security and Privacy Analytics, 2017, pp.55-63.

[30] A. Wang, W. Chang, S. Chen, and A. Mohaisen, "Delving into internet ddos attacks by botnets: Characterization and analysis," IEEE/ACM Transactions on Networking (TON), Vol.26, No.6, pp.2843-2855, 2018.

[31] S. Wedyan and F. Wedyan, "An Associative Classification Data Mining Approach for Detecting Phishing Websites," Journal of Emerging Trends in Computing and Information Sciences, Vol.4, No.12, 2013.

[32] D. Zhang, Z. Yan, H. Jiang, and T. Kim, "A domain-feature enhanced classification model for the detection of Chinese phishing e-Business websites," Information & Management, Vol.51, No.7, pp.845-853, 2014.

BIOGRAPHIES

Philip Abidoye received his M.Sc. from the University of Ibadan, Ibadan, Nigeria in 2006 and a Ph.D. degree from the University of the Western Cape, Cape Town, South Africa in 2015, both in Computer Science.

He is currently a Postdoctoral Fellow in the Department of Information Technology, Cape Peninsula University of Technology, Cape Town, South Africa. He has presented conference papers at international conferences, as well as published many papers in reputable international journals.

Dr. Abidoye is a member of the Institute of Electrical and Electronics Engineers (IEEE), South African Institute of Computer Scientists and Information Technologists (SAICSIT), and Computer Professionals Registration Council of Nigeria (CPN).

His research interests include secure wireless sensor networks, Cloud Computing security, security and privacy in the Internet of Things (IoT).

Boniface Kabaso received a Ph.D. degree in Information Technology from the Cape Peninsula University of Technology, Cape Town, South Africa. His research interests include software development, Internet of Things (IoT), soft computing, and Cloud Computing. Dr. Kabaso has published many research articles and conference papers in top-quality journals and conference proceedings.