Applied Information Technology And Computer Science Vol. 2 No. 2 (2021) 287-301

AITCS

Homepage: http://publisher.uthm.edu.my/periodicals/index.php/aitcs
e-ISSN :2773-5141

Covid-19 Phishing Detection Based on Hyperlink Using K-Nearest Neighbor (KNN) Algorithm

Nurul Ainatasha Afandi, Isredza Rahmi A Hamid*

Faculty Computer Science & Information Technology,
Universiti Tun Hussein Onn Malaysia, Batu Pahat, 86400, Malaysia

DOI: https://doi.org/10.30880/aitcs.2021.02.02.020
Received 15 June 2021; Accepted 09 September 2021; Available online 30 November 2021

Abstract: Phishing scam grow bigger during COVID-19 pandemic as the victim easily being deceived to click on the hyperlink that include latest information related to COVID-19. This link is sent by unknown user through email claimed to be from trusted organization. Although various way has been proposed to overcome this issue, number of phishing attack keep increasing. Our work focused on COVID-19 phishing detection based on hyperlink using KNN Algorithm. There are eight phases involved in this phishing detection model; Raw data, pre-processing, features extraction, training and testing the data using 10-fold cross validations, and classification of algorithm to detect phishing URL or legitimate URL. We consider using Uniform Resource Locator (URL) features such as Generic_TLD, URL_Length, Having_Sub_Domain, Prefix_Suffix and Having_Slash where the dataset is taken from Phishtank, SpyCloud, DomainTool and Kaggle. The phishing URL detection model will be tested on KNN Algorithm in terms of accuracy rate. This research produces promising results using 5 features with 97.80% accuracy for Dataset 1 and 99.60% accuracy for Dataset 2.

Keywords: Hyperlink, KNN algorithm, Phishing, URL

1. Introduction

Phishing is derived from the notion that Internet fraudsters use fake email to steal passwords and personal financial information. Phishing attack by creating email messages and web pages that look similar like an actual website in order to deceive users into submitting personal or financial information to the fraudsters' fake websites. The term 'phishing' first appeared in the 1990s. Hackers frequently substituted the letter 'ph' to replace the letter 'f' to create new phrases in the hacker community[1].

One of the most common threat is phishing email attack. Phishing email attack is a type of social engineering attack used by attacker to steal the victim's sensitive data such as bank credentials, health report, and home address[2]. This occur when an attacker pretends to be from trusted organization, deceiving the victim into opening fraudulent email which look legitimate as the original company. The

*Corresponding author: rahmi@uthm.edu.my

cyber criminals looked at this situation as an opportunity to trick users into revealing their personal information. Normally, they disguise themselves as reliable sources or pretend to be a government organization, health ministries or public health centers.

Since the Covid-19 pandemic began, working from home has become the new normal. The phishing attack are adapting to the pandemic and maps out the trend of COVID-related phishing emails identified as donation to fake charities, malware delivery or credential. Even prior to the pandemic, credential theft and phishing are more than 67% of breaches and 22% of all data breaches in 2020 involved in phishing attacks. While people still struggled to learn the real facts about the pandemic, the unethical cybercriminal community saw this chance as their opportunity. Most users are more likely to click on a phishing link. This is because, the phishing emails used "COVID" or coronavirus" as subject lines which attract the victim to click the email.

During COVID-19 pandemic, the front-line health care workers at University of Washington Medical Centre (UWMC) relied significantly on telemedicine to support patients remotely. They noticed a significant spike in phishing emails that persuade them to download malware via dangerous links. The propagation of malware or ransomware on healthcare networks can delay the diagnosis and treatment of COVID-19 patients[3]. Phishing attacks have continued to be a serious problem since it has become more complex and continually change their ways to defeat the anti-phishing techniques. Phishing websites or emails consists of fake URLs that look similar like the popular and legal websites. The fake websites have distinct Uniform Resource Locator (URL) than the original page but similar graphical user interfaces. By checking the URLs thoroughly, the user may spot the fake website. However, phishing attacker can hide the fake URL to obfuscate the user to click the link. Therefore, existing anti-phishing techniques such as content-based and keyword approach are not able to curb phishing attacks.

To overcome this problem, we proposed Covid-19 phishing detection based on hyperlink using machine learning algorithm. The objectives of this research are as follows:

- To design a Covid-19 detection model for phishing URL.
- To detect phishing URL related to COVID-19 based on hyperlink approach using KNN Algorithm.
- To evaluate the performance of phishing URL detection model tested on KNN Algorithm based on Accuracy rate, True Positive (TP), False Positive (FP), and Precision.

The rest of the paper is organized as follows: Section 2 discussed about related work of phishing email detection. Then, Section 3 described about detection methodology which has been used in this paper. Section 4 explained about the result from the experiment and finally, Section 5 discussed about the conclusion and future work.

2. Related Work

This section explains about phishing email detection, phisher, where it shows how and why phisher would like to do this attack. Phishing websites are created using new techniques which can allow them bypass most anti-phishing tool undetected. Current features do not have ability to give high accuracy since they are not common features shared between phishing and non-phishing websites. Therefore, the aim of this research is to analyze the URL of the phishing detection hat related to COVID-19. This research is using KNN algorithm to evaluate the features of the datasets. This research might also create an awareness to the employees who are currently work from home during this pandemic to work without feeling scared or unsecure or threatened when received an email.

2.1 Phishing

Various definitions for the term "phishing" have been proposed and discussed by experts, researchers, and other cybersecurity institution. The word "phishing" has been defined in a variety of ways based on its use and context, even though there is no definition owing to its ongoing evolution. The study [4] defines phishing as "a fraudulent activity that involves the creation of a replica of an existing web page to fool a user into submitting personal, password data or financial". This describes

phishing as an attempt to scam the user into revealing sensitive information such as bank credentials by sending malicious links to the user that leads to the fake websites. Phishing is becoming more common and sophisticated as a way of illegally compromising personal information and facilitating its misuse. Phisher is the attacker who attempts to trick people into giving information over the internet or by using email to take personal or private information which can results to monetary loss[5].

2.2 Phishing Attack

Phishing attack is a cybercrime where users are fooled by the phisher to access their personal data. To stop these threats, there are numerous methods and numbers of solutions available today. For multiple browsers, several toolbars are available that aim to warn users of possible phishing pages, trying to open them further. This attack is now known as spear phishing[6]. It makes it more difficult to discriminate between genuine and spoofed addresses. Starting phishing attacks with spoofed emails does considerable damage to user authentication.

Phishing URL will typically direct the user to visit a website where they are asked to update personal information, such as username, password, bank account numbers, or credit card numbers that the legitimate organization already has. The method used by phishers is usually to make fraudulent websites, similar to genuine website by mimicking the HTML code containing the same images and text. The most common method used by phishers is by forms, for an example, the internet banking login page. The phishing attack lifecycle are Planning, Setup, Attack, Collection and Fraud.

2.2.1 Planning

Phishers will identify the firm or user who would be their victim. Then, they will figure out how to acquire personal information from their victim, such as passwords, account numbers, and e-mail addresses. For personal information from the victim, mass-mailing and address gathering tactics are routinely be used. Spammers use these two strategies the most.

2.2.2 Setup

Phishers will prepare for the assault after deciding which firm to impersonate and who their intended victims are. Phishers will build strategies for sending the phishing message and capturing the vital data during the setup phase.

2.2.3 Attack

During this phase, a malicious payload is sent by one of three main dissemination vectors by spam email, fake message, or the creation of a fake website. In most cases, the phishing messages look to come from a reliable source. The victim may then take action that puts them exposed to a data breach. The user is requested for private information, either via a distant website or a Web Trojan installed locally.

2.3.4 Collection

When the confidential information is compromised, it is sent from a phishing server to the phisher. Phishers will keep track of every information that victims submit into webpages or popup windows.

2.3.5 Fraud

Finally, the phishers obtained the victim's personal information. This information is sued to mimic the victim in order to make illegal transactions. The success scam and failings are the assessed. If the phisher wants to plan another attack, they must complete this stage. They must then begin the phishing procedure all over again.

Cyber attackers use social engineering to manipulative victims into performing specific actions such as clicking on a malicious link or attachment or willfully divulging the confidential information by posing a legitimate individual or institution via phone and email[7]. Therefore, both individuals and organizations are at risk because almost any kind of their personal or organizational data can be valuable, whether it be to commit fraud or access an organization's network[8]. Besides that, phishing

scams can also target organizational data to support espionage efforts or state-backed spying on opposition groups.

2.3 Structure of Uniform Resource Locator (URL)

URL is used to identify the address of documents or other resources on the internet. URLs consists of multiple parts that include a protocol and a domain name. It tells a browser about how and from where to retrieve a resource. URL structure is the anatomy of how a particular URL looks like. Its either starts with HTTP or HTTPS (Hypertext Transfer Protocol Secure) which shows the secured versions of websites[9]. The structure of URL is as shown in Figure 1.

Figure 1: Structure of URL

Table 1 is the description of URL structure. URL consists of four parts which are protocol, subdomain, domain name and top-level domain.

Table 1: URL Structure

Parts	Description
Protocol	Protocol determines how data is transferred between the host and a web browser. HTTP and HTTPS are two of the most common protocols in URL. The protocol is located before the subdomain and is followed by a colon and two forward slashes.
Subdomain	Part of a domain that comes before the main domain name and domain extension. It is used to logically separate a website into sections.
Domain name	Domain name is the text that a user types into a browser to reach to a particular website. It typically broken up into two or three parts, each separated by a dot.
Top Level Domain	A final component of a domain name. TLD often serves as a clue to the purpose, ownership, or nationally of a website.

URL-based phishing attacks are mainly performed by embedding sensitive words or characters in a link that mimic similar but misspelling words, contain special characters for redirecting, use shortened URLs, use sensitive keywords which seem reliable and add a malicious file in the link and so on.

2.4 Phishing Detection Technique

Current research detects phishing email using various techniques such as image based, hyperlink based, content based, and keyword based. Each technique has its own advantages and disadvantage as explained in Table 2.

Table 2: Phishing detection technique

Technique	Advantage	Disadvantage
Image Based	Can quickly detect such embedded objects present in phishing webpage.	Quite complex than other techniques which need to be developed to make the solution viable.
Hyperlink Based	Real-time execution	Use location specific data
Content Based	It is easy to evaluate and easy to manage	High probability of false and failed alarm.
Keyword Based	Easy to download, manage, and update	Creates false alarm rate and the update is insignificant.

This work used the hyperlink based as a phishing detection technique because by analyzing the knowledge accessible on phishing URL and considering confidence as an indicator, the features like the top-level domain in the URL and Covid-19 keyword within the path portion of the URL were found to be sensible indicators for phishing URL.

2.5 Phishing Detection Approach

Work by [10] focused on detecting malicious URLs of COVID-19 pandemic using machine learning technique. They proposed a framework to detect malicious domain names that contain COVID related keywords. To achieve their aims, they trained and tested their model using 7849 datasets from WhoisDS and DomanTools with a 94.22% accuracy rate. The model offers a promising solution to minimizing COVID related phishing and malware attacks by detecting malicious domain names, early in the attack lifecycle. This is due to its ability to detect malicious URLs with a high accuracy using only the domain name and minimal number of features.

Work by [11] used machine learning based phishing detection from URLs. It is a real-time anti-phishing system, which use seven different classification algorithms and natural language processing (NLP) based features. They used dataset from PhishTank for phishing URL samples. For legitimate URL, they used the YandexXML dataset. To measure the system performance, they construct a new dataset which is Ebbu Phishing Dataset. This dataset consists of 36400 legitimate URLS and 37175 phishing URLs. According to the experimental and comparative results from the implemented classification algorithms, Random Forest algorithm with NLP based features gives the best performance with the 97.98% accuracy rate for detection of phishing URLs.

Orestis Christou et al. [12] developed a machine learning model to detect fraudulent URLs and used the Splunk platform. It detects phishing URL through the top-level domain. They trained the Support Vector Machine (SVM) and Random Forests algorithms using malicious and benign datasets. The datasets are from Alexa and Phishtank with 3400 data entries where 1700 legitimate URLs and 1700 phishing URLs. They evaluated the algorithms performance with precision and recall reaching up to 85% precision and 87% recall in the case of Random Forest while SVM achieved up to 90% precision and 88% recall using only descriptive features. The comparison between previous works is as Table 3 below.

Table 3: Existing works on phishing detection approach

Work	Dataset	Feature Approach	Algorithm	Sample (URL)	Accuracy Result
Jamil et. al [10]	WhoisDS, Domain Tools	Lexical Features	SVM, KNN, Naïve Bayes, Regression, AdaBoostM1	Legitimate - 1573, Phishing - 6321	94.22%
Ozgur et. al [11]	Phishtank, YandexXML	Natural Language Processing (NLP) based features and Hybrid features	Naïve Bayes, Random Forest, KNN, Adaboost, K-star, SMO, Decision Tree	Legitimate - 36400, Phishing - 37175	97.98%
Orestis et. al [12]	Alexa, Phishtank	Hybrid features	Bayes Net	Legitimate - 1700, Phishing - 1700	96%

Our work differs than the previous research in Table 3 in such a way that we focused based on hyperlink detection where we study the phishing URL structure. Phishers have used several ways to fool victims into believing the link is legitimate. Therefore, it is very important to recognize the pattern of phishing URL so that we did not become the victim. Our work consider to use two datasets collected from SpyCloud[13] and PhishTank[14] with 500 data consist of 250 phishing URL from SpyCloud and 250 legitimate URL from Phishtank. We use KNN algorithm tested in WEKA tool to evaluate performance of the proposed hyperlink-based features. The KNN algorithm will evaluate the performance in terms of accuracy rate, true positive, false positive, and precision.

2.6 Machine Learning Algorithm

Machine learning is primarily a field of artificial intelligence that has gained considerable interest in the digital arena as a core component of digitalization solutions. A computer program is assigned to perform such tasks in machine learning, and it is said that the machine has benefited from its training if its measurable success increases in these tasks as it gains more and more experience in performing these tasks. Machine learning is applied in wide variety of field such as robotics, data mining, traffic prediction, online transportation network, medical diagnosis, and online fraud prediction.

The performance of detection approaches can be enhanced during the learning phase of a classifier (whether the classifier human or software). In the case of end-users, their classification ability can be enhanced by improving their knowledge of phishing attacks by learning individually through their online experience, or by external training programs. In the case of software classifiers, this can be achieved during the learning phase of a Machine Learning-based classifier, or the enhancement of detection rules in a rule-based system.

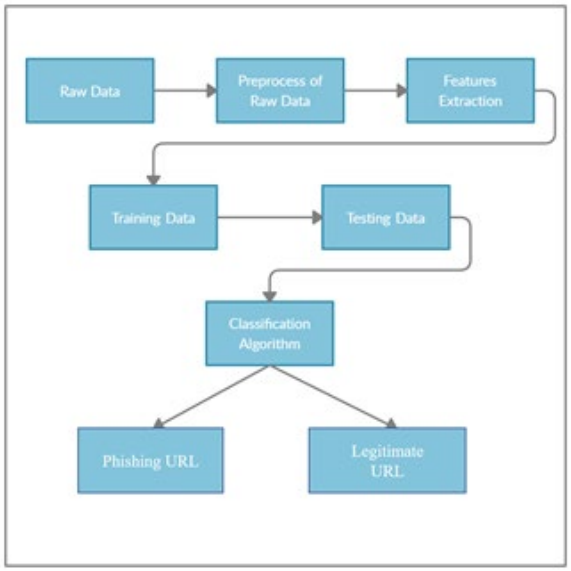
2.6.1 K-Nearest Neighbor (KNN) Algorithm

KNN can be used for classification as the output is a class membership. The object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its K nearest neighbors. The research of [15] states that KNN algorithm is an extensively used classification algorithm owing to its simplicity, ease of implementation and effectiveness as it has large memory requirements as well as high complexity. Work by Jamil et. al [10] detected malicious URLs of COVID-19 pandemic using machine learning technique. One of the machines learning selected was KNN

algorithm. They used lexical based features and achieved the best accuracy rate of 94.22% using KNN algorithm which shows the highest accuracy rate compared to another machine learning algorithm.

3. Covid-19 Phishing URL Detection Model

This section discusses the Covid-19 phishing detection model based on hyperlink using machine learning algorithm. There are eight phases involved in this phishing detection model; Raw data, pre-processing, features extraction, training the data, testing the data using 10-fold cross validations, and classification of algorithm to detect phishing URL or legitimate URL as shown in Figure 2. We used Waikato Environment for Knowledge Analysis (WEKA) tool to analyze the performance of hyperlink features selected in term of Accuracy rate, True Positive (TP), False Positive (FP), and Precision[16].



3.1 Raw Data

We consider using two datasets collected from SpyCloud[13] and PhishTank[14] for Dataset 1, and DominTools[17] and Kaggle[18] for Dataset 2. We constructed two datasets where each dataset has 500 data consist of 250 phishing URL and 250 legitimate URL respectively as shown in Table 4.

Table 4: Dataset summary

Dataset	Phishing	Legitimate	Total
1	Spycloud	PhishTank	500
2	DomainTools	Kaggle	500

3.2 Preprocessing Data

Pre-processing data is the process of transforming raw data into understandable format. In pre-processing procedure, we generate components of a feature by analyzing the dataset. The information gain (IG) values of the extracted features are calculated. In this step, the most informative features are selected using classification algorithms. After locating COVID-19 related domains names, all inbound URLs must be structured for feature collection. Malicious connections containing characters that are not usable at the time of registration of the domain name are then taken away. In preprocessing data, we used Microsoft Excel Spreadsheet to export the data and make sure the data is already clean before later analysis in the feature extraction process.

3.3 Feature Extraction

Work by [10] used lexical based features such as host length, hyphens, and numeric characters. While Ozgur et. al [11] used hybrid features which is similar with our work. We used combination of previous work features which are 5 hyperlink-based features that are Generic_TLD[12], URL_Length[11], Having_Sub_Domain[10], and Prefix_Suffix[19]. We come out with new feature which is having slash in URL as shown in Table 5.

Table 5: Phishing URL features

Features	Descriptions
Length of URL	Length of the domain name string
GenericTop-Level Domain	Domains at the highest level of the domain name system (DNS)
Having Sub Domain	The increased number of dots in URL.
Prefix_Suffix	The URL contains many dashes symbol.
Having_Slash	The URL contains many slashes.

3.4 Training and Testing Data

Training and testing the dataset is important process to perform the suitable data that might help in the end of the research. Testing the dataset will also be using WEKA to choose the most suitable features. In this phase, we used 10-Fold Cross Validation. This method is called rotation estimation or out-of-sample testing for assessing how the results of a statistical analysis will generalize to an independent dataset. This technique will divide the dataset into 10 parts which are called "folds" to hold out each part in turn and average the results. So, each data point in the dataset is used once for testing and 9 times for training. This statistical method used to estimate the skill of machine learning models. Cross-validation goals are to overcome the problem of overfitting and making the predictions more general.

3.5 Classification Algorithm

Classification algorithm is used to classify each data in a dataset into a group of predefined categories. Our work used K-Nearest neighbor (KNN) Classification Algorithm to evaluate the performance of hyperlink features selected for both datasets. We evaluate the hyperlink feature for both dataset performance pertaining Accuracy rate, True Positive (TP), False Positive (FP), and Precision. We run the experiment using WEKA tool. We use KNN algorithm because it uses feature similarity to estimate the value of new datapoints which means that the new data point will be assigned a value based on how the data matches the points in the training set. KNN algorithm was proposed by Cover and Hart [20] where it was frequently used to classify new data as it is the simplest algorithm among the others. We used KNN algorithm because it is a simple algorithm, easy to understand and very powerful. Moreover, KNN algorithm can handle classification problems, and can naturally handle multi-classification problems.

KNN algorithm uses standard Euclidean distance to measure the difference or similarity between training and test instance. KNN considers the most common class of k-nearest neighbors to estimate the class of test instance. The standard Euclidean distance d(xi,xj) is defined in equation as follows:

$$\text{Euclidean} = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2} \qquad Eq.\ 1$$

Where, k is number of nearest neighbors while x_i represent Variable of vector x and y_i is variable of vector y.

3.6 Performance Metric

In order to measure the effectiveness of the proposed Covid-19 phishing detection model, we consider using these performance metrics:

a. Accuracy: The number of phishing URL correctly predicted by the algorithm.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FN+FF} \qquad Eq.\ 2$$

b. True Positive (TP): The number of phishing URL correctly labelled as phishing.

$$\text{True Positive} = \frac{TP}{TP+FP} \qquad Eq.\ 3$$

c. False Positives (FP): The number of legitimate URL incorrectly labelled as phishing.

$$\text{False Positive} = \frac{FP}{TN+FN} \qquad Eq.\ 4$$

d. Precision: Fraction of correctness.

$$\text{Precision} = \frac{TP}{TP+FP} \text{x } 100\% \qquad Eq.\ 5$$

3.7 Software and Hardware Requirement

Table 6 shows the hardware requirement to conduct the experiment. We used Waikato Environment for Knowledge Analysis (WEKA) as a data mining software which use a collection of machine learning algorithms. These algorithms can be applied directly to the data or called from the Java code. WEKA is a collection of tools for Regression, Association, Clustering, Data pre-processing, Visualization and Classification as shown in Figure 3.

Table 6: Hardware Requirement

Hardware		Description
Swift SF314-55G	Processor	Intel® Core ™ i5-8265U CPU @ 1.60GHz 1.80GHz
	Windows Edition	Windows 10 Home Single Language
	System Type	64-bit operating system, x64 based processor
	Installed RAM	8.00GB

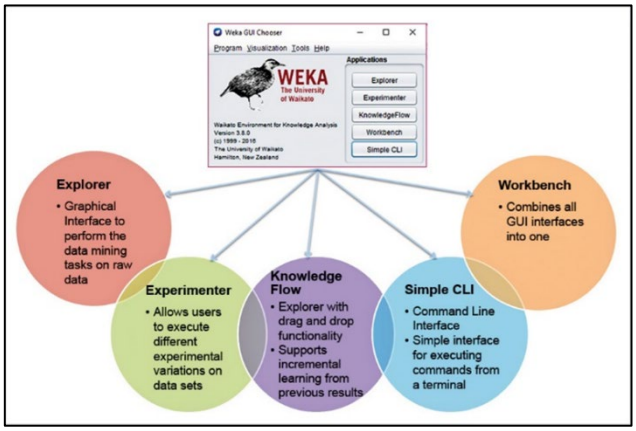


Figure 3: WEKA's application interfaces

4. Performance Analysis

This section describes the collection of malicious and legitimate data and our proposed experimental setup using KNN Algorithm. There are two datasets collected from four sources: Spycloud, Phishtank, Domaintools, and Kaggle.

4.1 Experimental Setup

We conduct the experiment using 2 datasets from four sources as shown in Table 7. Dataset 1 consists of 500 data which are 250 phishing URL and 250 legitimate URL. The phishing URL is

extracted from SpyCloud[13] while the legitimate URL is collected from Phishtank[14]. Dataset 2 consists of 500 data which divided into two parts which are 250 phishing URL and 250 legitimate URL. The phishing URL is collected from DomainTools[21] where the legitimate URL is extracted from Kaggle[18]. A list of legitimate domain names which are publicly available was extracted from Phishtank and Kaggle. Then, each dataset was filtered for coronavirus related keywords such as "Covid19", "coronavirus", "cov-19", "covidtest", "koronavirus", and "corona". Finally, we have 1000 URL for both datasets. Any domain names that do not consists of those coronavirus related keywords in the URL was assumed to be legitimate.

Table 7: Source of datasets

Dataset	Phishing		Legitimate		Total
Dataset 1	SpyCloud	250	PhishTank	250	500
Dataset 2	DomainTools	250	Kaggle	250	500

4.2 Hyperlink Features

We extracted five Uniform Resource Locator (URL) features such as Generic Top-Level Domain (TLD), URL Length, Having Sub Domain, Prefix Suffix and Having Slash. Then, we constructed 2 datasets consists of both legitimate and phishing URLs. The dataset in URL form is saved in comma separated values (csv) file.

4.2.1 Generic Top-Level Domain (TLD)

The URL is processed to split the domain name, path, and TLD portions separately. Phishing occurred on domain names in 182 TLDs [22]. The generic TLDs are used by and are popular with registrants across the world. The highest scoring TLDs from the systematic registration of domain names by phishers are ".ly" for Libya , ".mn" for Mongolia , and ".hk" for Hong Kong. Therefore, in this experiment, URL that have this top three domain names in TLDs are considered as phishing.

if URL have TLDs such as ".ly", ".mn", and ".hk" → Phishing URL, *else*, Legitimate URL.

4.2.2 URL Length

URL is a structured text string that web users used to identify a network resource on the internet. The network protocol, host name, and path are parts of the URL string. The URL length was examined for both legitimate and phishing URL on Dataset 1 and Dataset 2. Phishing URL is found to be greater than 25 characters where legitimate URLs is less than 25 characters. Therefore, the binary data represent URL length. If the character is more than 25 characters, we consider the URL as phishing and set value 1 or else the value is 0 to indicate that the URL is legitimate.

if URL length > 25 Characters → Phishing URL, *else*, Legitimate URL.

4.2.3 Having Sub Domain

Domain name include the country-code top-level domains (ccTLD). Given the following hyperlink: https://www.myhermes.co.uk/. The ccTLD is "uk". The "co" is shorthand for "company", the combination of "co.uk" is called second-level domain (SLD). To produce a rule for extracting this feature, we firstly have to omit the (www.) from the URL which is in fact a sub domain in itself. Then we must remove the (ccTLD) if it exists. Lastly, we count the remaining dots. If the number of dots is greater than two, then the URL is classified as phishing and we set value 1 since it will have multiple sub domains. Otherwise, if the URL has no sub domains, we will assign it as legitimate URL to the feature and set value 0.

if dots in Sub Domain Part > 2 → Phishing URL, *else*, Legitimate URL.

4.2.4 Prefix Suffix

The dash symbol is rarely used in legitimate URLs. Phishers tend to add prefixes or suffixes separated by dash (-) to the domain name so that the victim think that they are dealing with legitimate webpage. We set value 1 if domain name contains dash (-) symbol to classify as phishing URL, otherwise 0.

if Prefix Suffix == "-"→ Phishing URL, *else,* Legitimate URL.

4.2.5 Having Slash

Phishers attempt to deceive users by making a doubtful URL appear real. The insertion of slashes to URLs is a scamming tactic by phishers. Therefore, we proposed new feature that is the number of slash (/) in URLs to identify either the legitimate or phishing URL. If the number of slash (/) is greater than or equal to three, we set as phishing URL, otherwise legitimate URL.

if Having Slash > = 3 → Phishing URL, *else,* Legitimate URL.

4.3 Constructing Feature Matrix

In this section, we construct the feature matrix of 5 features F_i ,i =1,….,5, i for all phishing and legitimate datasets. Note that all features are in binary value. The Ri value for each feature is summarized in Table 8.

Table 8: Features Summarized

Features	Description	Value
F_1	Generic Top-Level Domain	R1 ={0,1}
F_2	URL Length	R1 ={0,1}
F_3	Having Sub Domain	R1 ={0,1}
F_4	Prefix Suffix	R1 ={0,1}
F_5	Having Slash	R1 ={0,1}

Let E = {e_1 ,e_2 ,.....,$e_{|E|}$} and F = {f_1 , f_2 ,.., $f_{|F|}$ } denotes all the URL and feature vector space respectively. So, | E | is a total URL and | F | refer to size of feature vector. Let a_{ik} a be the value of *k*th feature of *i*th URL. Therefore, the presentation of each URL is $A_i = \{a_{i1}, a_{i2}, \ldots . a_{i[e]}\}$, and each URL is $A = \{a_{ik}\}$ where i = 1,2,…,|F| and k =1,2,....,|E| . Where each URL consists of A={Generic_TLD, URL_Length, Having_Sub_Domain, Prefix_Suffix, Having_Slash}. Then, all datasets are converted to arff format to run in the WEKA and tested using KNN Algorithm.

4.4 Result and Discussion

We compare our results with existing work In Table 9. Jamil et. al [10] used 5 features based on URL and achieved 94.22% accuracy rate. Ozgur et. al [11] tested on two different types of features which are 40 NLP features and 1701 word-features. Their results show the best when more features used to classify phishing which is 97.98% accuracy rate. Orestis et. al [12] proposed hybrid features with 18 features. They successfully determine 96% accuracy rate. Our work used 5 features on two datasets and successfully achieved 97.80% for Dataset 1 and 99.60% accuracy rate of Dataset 2 tested using KNN algorithm.

Table 9: Comparison results with existing work

Work	Dataset	Feature Approach	Algorithm	Sample (URL)	Accuracy Result
Jamil et. al [10]	WhoisDS, Domain Tools	Via Lexical Features	SVM, KNN, Naïve Bayes, Regression, AdaBoostM1	Legitimate -1573, Phishing - 6321	94.22%

Table 9: (Cont…)

Work	Dataset	Feature Approach	Algorithm	Sample (URL)	Accuracy Result
Ozgur et. al [11]	Phishtank, YandexXML	Natural Language Processing (NLP) based features and Hybrid features.	Naïve Bayes, Random Forest, KNN, Adaboost	Legitimate - 36400, Phishing - 37175	97.98%
Orestis et. al [12]	Alexa, Phishtank	Hybrid features	Bayes Net	Legitimate - 1700, Phishing - 1700	96%
Our Approach	SpyCloud, Phishtank, DomainTools, Kaggle	Hyperlink based feature	KNN	Legitimate - 500, Phishing - 500	D1: 97.80% D2: 99.60%

4.5 Accuracy Result

Figure 4 shows the accuracy result of Dataset 1 and Dataset 2 tested using KNN Algorithm. Dataset 1 is extracted from SpyCloud and Phishtank while Dataset 2 consists of URL from DomainTools and Kaggle. Accuracy result for Dataset 1 and Dataset 2 is 97.80% and 99.6% respectively. This shows that Dataset 2 detect phishing URL more accurate compared to Dataset 1when tested using KNN Algorithm. The hyperlink features selected show promising result when tested on both datasets.

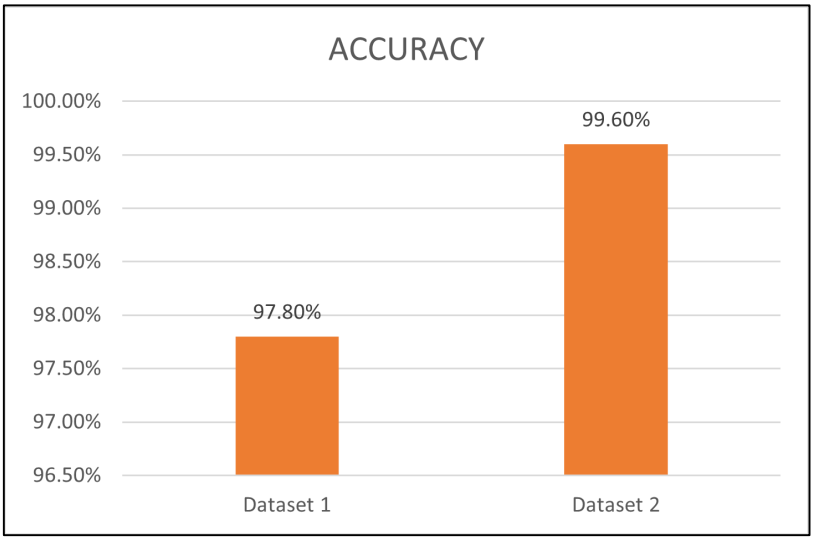


Figure 4: Accuracy result for Dataset 1 and Dataset 2

4.6 True Positive (TP)

Figure 5 shows the TP rate for Dataset 1 and Dataset 2. Datasets 1 achieved 99% result of TP rate of legitimate URLs and 93% TP rate of phishing URL. Besides, Dataset 2 shows exactly 100% TP rate of legitimate URL and 98% TP rate of phishing URL.

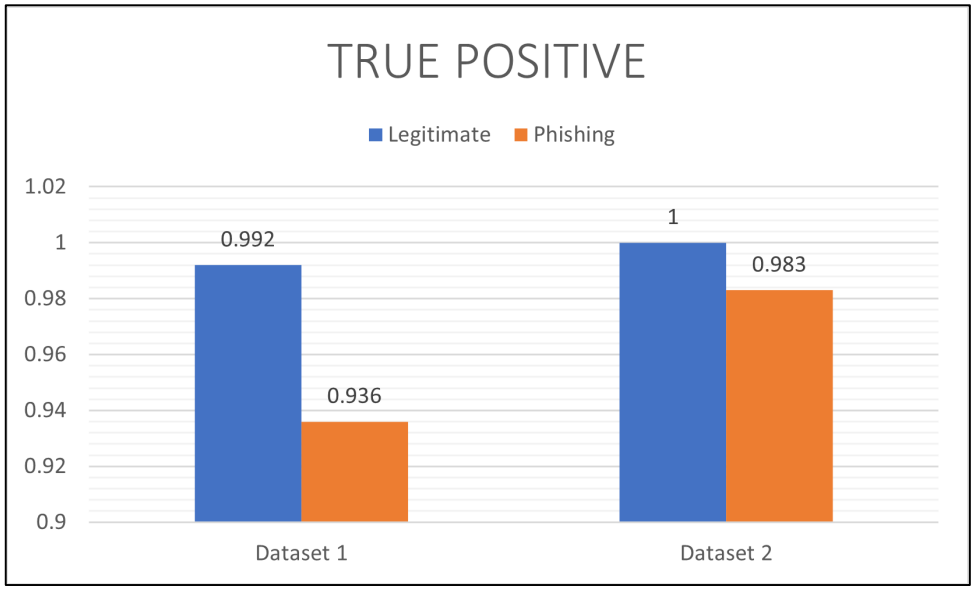


Figure 5: True Positive Rate for Dataset 1 and Dataset 2

4.7 False Positive (FP)

Figure 6 shows FP rate for Dataset 1 and Dataset 2. Datasets 1 achieved of 6.4% FP rate of legitimate URLs and 0.8% FP rate of phishing URL. While Dataset 2 shows 1.7% FP rate of legitimate URL and 0.01% FP rate of phishing URL. FP rate shows that the truth is phishing URL, but the test predicts as legitimate URL. This shows that Dataset 2 better detection as compared to Dataset 1 because it has lowest FP rate.

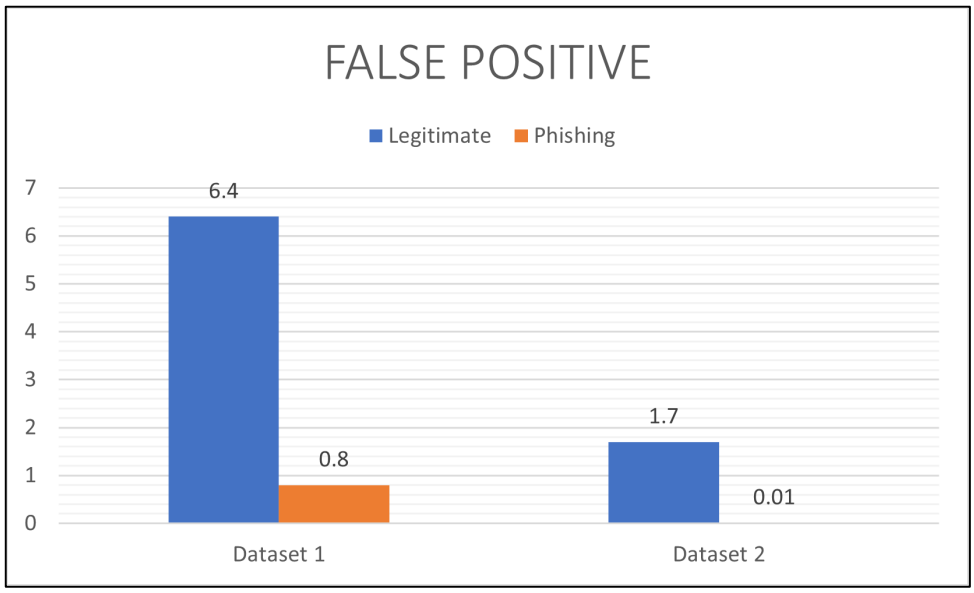
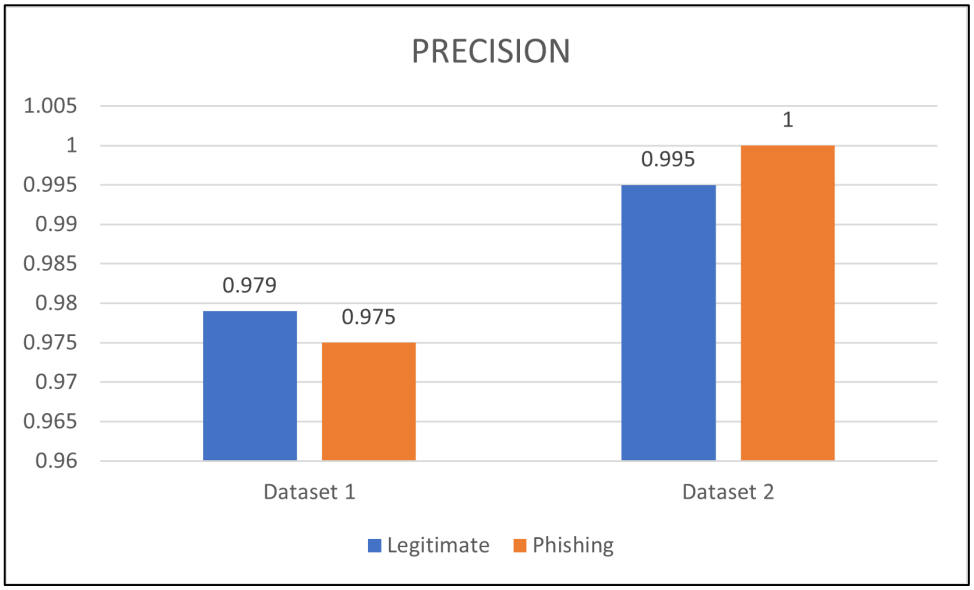


Figure 6: False Positive Rate for Dataset 1 and Dataset 2

4.8 Precision

Precision is the proportion of relevant results in the list of all returned search results. It can also be seen as a measure of "exactness". Figure 7 shows precision of Dataset 1 and Dataset 2. Dataset 1 obtained 97% precision rate of legitimate URL and phishing URL respectively whereas Dataset 2 obtained 99% precision rate of legitimate URL and exactly 100% precision rate of phishing URL.



5. Conclusion and Future Direction

Phishing URLs are challenging threat in cyber security which can steal user's sensitive information. The phishers are using various ways to launch phishing URLs crafting to bypass the detection technique. Therefore, this research objectives are to design a detection model for phishing URL, to detect phishing URL related to COVID-19 based on hyperlink approach using KNN Algorithm and to evaluate the performance of phishing URL detection model tested on KNN Algorithm in terms of Accuracy rate, True Positive (TP) rate, False Positive (FP) rate, and Precision. We proposed a phishing detection model based on hyperlink that contain COVID keywords. We used five hyperlink-based features and achieve promising result with 97.80% and 99.60% accuracy rate for Dataset 1 and Dataset 2 respectively. We consider using two datasets with 1000 data consists of both phishing and legitimate URL. By combining these datasets, we used KNN algorithm to classify the data into phishing or legitimate URLs. The COVID19 phishing detection model offers a promising solution to reduce the COVID related phishing URL received through email. The result motivates future works to explore more about attackers' behavior and profile their modus operandi. We also intend to test the new features with other machine learning algorithms in order to evaluate the performance of this features. Moreover, the proposed approach can be executed on online website as well.

Acknowledgement

The authors would like to thank the Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia for its support and encouragement throughout the process of conducting this study.

References

[1] P. Tiwari and R. Ratan Singh Assistant Professor, "Machine Learning based Phishing Website Detection System.". [Online]. Available: www.ijert.org. [Accessed: 10-Jun-2021].

[2] R. Dhamija and J. D. Tygar, "The battle against phishing: Dynamic security skins," pp. 77–88, 2005.

[3] E. S. Grange *et al.*, "Responding to COVID-19: The UW Medicine Information Technology Services Experience," pp. 265–275, 2020.

[4] Z. Alkhalil, C. Hewage, L. Nawaf, and I. Khan, "Phishing Attacks: A Recent Comprehensive Study and a New Anatomy," pp. 3060–3389, 2021.

[5] F. Quinkert, M. Degeling, J. Blythe, and T. Holz, "Be the Phisher -- Understanding Users' Perception of Malicious Domains," pp. 263–276, 2020.

[6] L. Zhang-Kennedy, E. Fares, S. Chiasson, and R. Biddle, "Geo-Phisher: The design and evaluation of information visualizations about internet phishing trends," pp. 30–41, 2016.

[7] R. A. Hamid, "Phishing Detection and Trackback Mechanism," 2015.

[8] B. Wardman, J. Britt, and G. Warner, "New tackle to catch a phisher," pp. 62, 2014.

[9] K. Arai, "Visualization of Link Structures and URL Retrievals Utilizing Internal Structure of URLs Based on Brunch and Bound Algorithms," pp. 1080–1456, 2012.

[10] J. Ispahany and R. Islam, "Detecting Malicious Urls of COVID-19 Pandemic Using ML Techniques." 2020.

[11] Ozgur, "(PDF) Machine learning based phishing detection from URLs." https://www.researchgate.net/publication/344952543_Machine_learning_based_phishing_detection_from_URLs [Accessed: 13-Jun-2021].

[12] Orestis, "(PDF) Phishing URL Detection Through Top-Level Domain Analysis: A Descriptive Approach." https://www.researchgate.net/publication/338339970_Phishing_URL_Detection_Through_Top-Level_Domain_Analysis_A_Descriptive_Approach [Accessed: 13-Jun-2021].

[13] Team Spycloud, "Analyzing 136,000 New Domains with COVID-19 Themes | SpyCloud." https://spycloud.com/analyzing-136k-new-domains-with-covid-19-themes/ [Accessed: 13-Jun-2021].

[14] Phishtank, "PhishTank > Developer Information." http://phishtank.org/developer_info.php [Accessed: 13-Jun-2021].

[15] S. Taneja, C. Gupta, K. Goyal, and D. Gureja, "An enhanced K-nearest neighbor algorithm using information gain and clustering," pp. 325–329, 2014.

[16] Vidhya, "Weka – Graphical User Interference Way To Learn Machine Learning." https://www.analyticsvidhya.com/learning-paths-data-science-business-analytics-business-intelligence-big-data/weka-gui-learn-machine-learning/ [Accessed: 05-Jan-2021].

[17] Sean, "Free COVID-19 Threat List - Domain Risk Assessments for Coronavirus Threats." https://www.domaintools.com/resources/blog/free-covid-19-threat-list-domain-risk-assessments-for-coronavirus-threats [Accessed: 13-Jun-2021].

[18] Kaggle, "Phishing website dataset | Kaggle." https://www.kaggle.com/akashkr/phishing-website-dataset [Accessed: 13-Jun-2021].

[19] G. Varshney, M. Misra, and P. K. Atrey, "A survey and classification of web phishing detection schemes," pp. 6266–6284, 2016.

[20] Y.-L. Cai, D. Ji, and D.-F. Cai, "A KNN Research Paper Classification Method Based on Shared Nearest Neighbor." 2010.

[21] Proofpoint, "Domain Fraud Report 2019 | Proofpoint." https://www.proofpoint.com/us/resources/white-papers/domain-fraud-report [Accessed: 07-Jun-2021].

[22] G. A. Afilias and R. Rasmussen, "Global Phishing Survey: Domain Name Use and Trends in 2007," 2008. [Online]. Available: http://www.antiphishing.org•info@antiphishing.org. [Accessed: 13-Jun-2021].