Development of a Cybernetics Phishing Detection Approach

Onyema Chinazo Juliet, Chidi Ukamaka Betrand, and Douglas Allswell Kelechi

1. Department of Computer Science, School of Information and Communication Technology, Federal University of Technology Owerri, Imo State

Abstract:
Phishing being one of the core problems encountered by online community has led to numerous financial losses, identity theft, kidnappings, deaths and lots more. Phishing detection and response software which is a cybersecurity tool that identifies and rectifies phishing threats before the phishing attack causes harm, is a part of the broader threat detection and online security response. The phishing detection system employed in this study is a one-page web application model aimed at solving phishing problems, with high-rate detection accuracy, low-rate false alarm and consistent database updates. Machine Learning technology was used by the detection system in extracting and analyzing different features of malicious and phishing URLs. Extreme Gradient Booster (XGBoost), Decision trees, Multilayer Perceptron and Random Forest models were compared in detecting phishing websites with the aid of Python, Science-kit learn (sklearn) and Flask. The Frontend of the one-page web application was done using HTML and CSS. The generated dataset from which the models were tested and trained were extracted from various open-source platforms which provided some phishing URLs in various formats like JSON and CSV with hourly updates. Up to 2000 random phishing URLs were collected from this dataset to train and test the Machine Learning models. Out of all these types, the benign URL dataset was considered for this paper. An accuracy of 86.6% with 13.34% false-positive rate was achieved in our proposed approach on our dataset, together with 13.6% false-positive rate and 86.4% accuracy on the benchmark dataset, which performs better than the existing baseline approaches.

Keywords: Machine Learning, Feature Selection, URL, Classification, XGBOOST, Detection, Phishing.

INTRODUCTION

The term Phishing is a cybercrime that involves a perpetrator sending a fraudulent or malicious mail in disguise, making it seem like it comes from a legitimate source thereby requesting for some particular and sensitive information like username, phone number, bank account details, and so forth [1].

Phishing attacks constitute a significant danger to online space along with IT companies that operate with highly sensitive data and this has been on the increase over the years surpassing every effort to curtail it.

As a result of over 33000 worldwide phishing assaults in 2012 a loss of $687 million was recorded likewise in 2004, when over 50 million phished emails caused a loss of $10 billion in financial institutions. According to the Anti-Phishing Working Group (APWG), the top targets in the second quarter are financial bodies (16%), web (15%), cloud storages (9%) and payment systems (45%) [2].

However, it is crucial for all internet and computer users to keep information secure and safe as to reduce the risk of fraud that may occur while accessing various websites by identifying phishing links and emails thus helping protect them against these attacks.

THE CONCEPT OF PHISHING ATTACKS

Phishing attacks generally fall in two classes; social engineering and malware-based attacks. Attackers in the social engineering phishing-base usually try to control the victims' accounts by sending them simulated emails with fake URLs that deliver to phishing websites. Social engineering-base attacks, also called deceptive phishing are further categorized into email-based and website-based phishing. Malware-based phishing on the other side uses a variety of malicious programs that run on the victims' machines. This type of phishing is further classified as; keyloggers/screen loggers, session hijacking, host file poisoning, content injection and DNS phishing [3].

According to Zhang et al. (2012), different regions may use different phishing techniques such as

- Spear Phishing
- Link Manipulation
- Vishing (Voice Phishing)
- Web-Based Delivery
- Smishing (SMS Phishing)
- Payment/delivery scam
- Downloads

According to Yasin et al. (2018), some address bar-based features on how to predict phishing websites include the followings:

The Use of Internet Protocol (IP) Address

Whenever a domain name is substituted with an IP address in the URL, it is certain that a perpetrator is attempting to gather some sensitive information. For instance, the following link illustrates a domain being replaced by an IP address and an IP address converted to hexadecimal code respectively: http://126.78.2.134./fake/html and
http://0x67.0xAA.0xDA.0x63/3/paypal.ca/index.html

Rule: IF (Domain part has an IP Address = phishing
Otherwise = Legitimate

Use of Long Universal Resource Locator (URL)

Phishers can conceal suspicious information using long URLs as illustrated: http://fedacadefifo.com.fr/4/ute/ab56e3e419e51502h318bde47b884e4a/cd=home&receive= 11004d58f5b74f8dc1e7c2e8dd4105e811004d58f5b74f8dc1e7c2e8dd4105e8@malicious.website. HTML

Guarantying the accuracy of the study, the length of the Universal Resource Locator in the dataset was calculated to produce a certain average length. This implies that when URL's length is equal or greater than 54 characters, the URL is suspected to be phishing. Reviewing the dataset, 1220 URL lengths were equal or greater than 54 constituting 48.8% size of the entire dataset.

Rule: IF {$URLlength$ < 54 → $feature$ = Legitimate
$elseifURLlength$ ≥ 54 and ≤ 75 → $feature = Suspicious$
$otherwise$ → $feature$ = Phishing

This feature rule has been updated by using a method based on the frequency, hence improving its accuracy.

The Use of Sub Domain and Multi-Sub Domains

Considering this link: hhtp//www.hud.ac.ng/students/, the domain includes the country-code Top-Level Domain (ccTLD) which is 'ng' in our given link, 'ac' is shortened form of academics, and 'ac.ng' being a Second-Level Domain (SLD) with 'hud' the domain's real name. Creating a rule to extract this feature, the 'www.' in the Universal Resource Locator which is the subdomain must be removed, then the ccTLD deleted if it already exists. The left over dots are then added,if there are more than one dot and only a subdomain the URL is categorized as malicious. However, dots with multiple subdomains are classified as phising otherwise, it will assign 'Legitimate' to the feature.

Rule: IF {Dots in Domain Part = 1 =Legitimate
Dots in Domain Part = 2 = Suspicious
Otherwise →Phishing

The Use of Hypertext Transfer Protocol with Secure Sockets Layer (HTTPS)

Though using HTTPS is crucial in creating a legitimate website, it is deemed insufficient. Kazemian H. B. & Ahmed S. (2021) recommend verifying the HTTPS certificate's validity, the issuer's level of trust, and perhaps the certificates length of existence. Among the most reliable certified authorities are "GeoTrust, GoDaddy, Network Solutions, Thawte, Comodo, Doster, and VeriSign." Furthermore, it was discovered that the minimum age of a trustworthy certificate is two years by putting our datasets to the test.

Rule: IF {Using HTTPS with trusted Issuer and Certificate >=1year=Legitimate
Using HTTPS and Untrusted Issuer = Suspicious
Otherwise=Phishing

Favicon

This is visual representation (icon) connected to a certain website. However, Favion is shown in the address bar, including newsreaders and graphical browsers as visual reminder of the website identifier. If the favicon loads from a different domain than the one displayed in the URL bar, the webpage is probably a phishing attempt.

Rule: IF {Favicon Loads from External Domain → Phishing
Otherwise → Legitimate

Use of Non-Standard Port

This feature is useful for confirming if a particular service like the HTTP is available or not on a particular server. It is far preferable to only open the ports you need to do what you need to do to control intrusions. Many firewalls, proxy servers, and Network Address Translation (NAT) servers will by default block all or the majority of the ports and only open the ones that are chosen. User information is at risk if all ports are open because phishers can operate almost any service they want.

The most important ports and their preferred status are shown in Table 1

Rule: IF {Port # is of the Preferred Status → Phishing
Otherwise → Legitimate

Table I: Important ports and their most preferred status

PORT	SERVICE	MEANING	PREFERRED STATUS
21	FTP	Transfer files from one host to another	Close
22	SSH	Secure File Transfer Protocol	Close
23	Telnet	Provides a bidirectional interactive text-oriented communication.	Close
80	HTTP	Hypertext transfer protocol	Open
443	HTTPS	Hypertext transfer protocol secured.	Open
445	SMB	Providing shared access to files, printers, serial ports	Close
1433	MSSQL	Store and retrieve data as requested by other software applications.	Close
1521	ORACLE	Access the Oracle database from the web.	Close
3306	MySQL	Access MySQL database from the web.	Close
3389	Remote Desktop	Allow remote access and remote collaboration	Close

The Existence of 'https' in the Domain of an URL.
Phishers may add the "HTTPS" token to the domain part of a URL to trick users. For example, http://https-www-paypal-it-webapps-mpp-home.soft-hair.com/.

Rule: IF {Using HTTP Token in Domain Part of The URL → Phishing
Otherwise → Legitimate

In the existing system, the detection processes include:

- The use of blacklist database which has all the phishing URLs.
- The use of IP address
- The use of mail/mail-to attributes

Meanwhile, that an address is on a blacklist does not mean it is malicious. Legitimate addresses can be blacklisted though this is not very flexible and can be time-consuming to maintain. However, an effective blacklist has to be kept up-to-date with new threats and this takes extra time and effort.

Thus, a real-time detection system is developed to help reduce the stress undergone in observing and analyzing the physical features of URLs alone in other to tell if it is phishing or not.

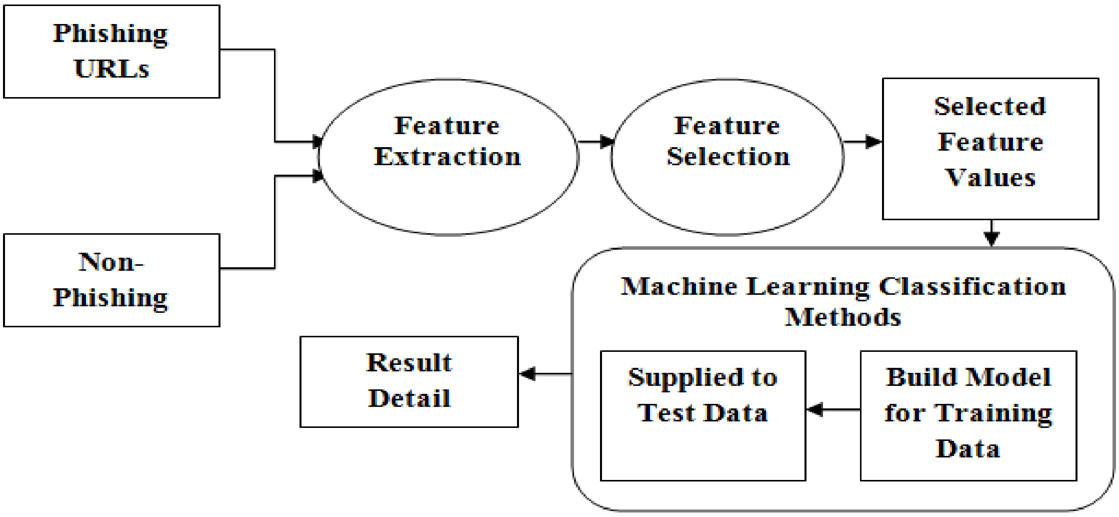
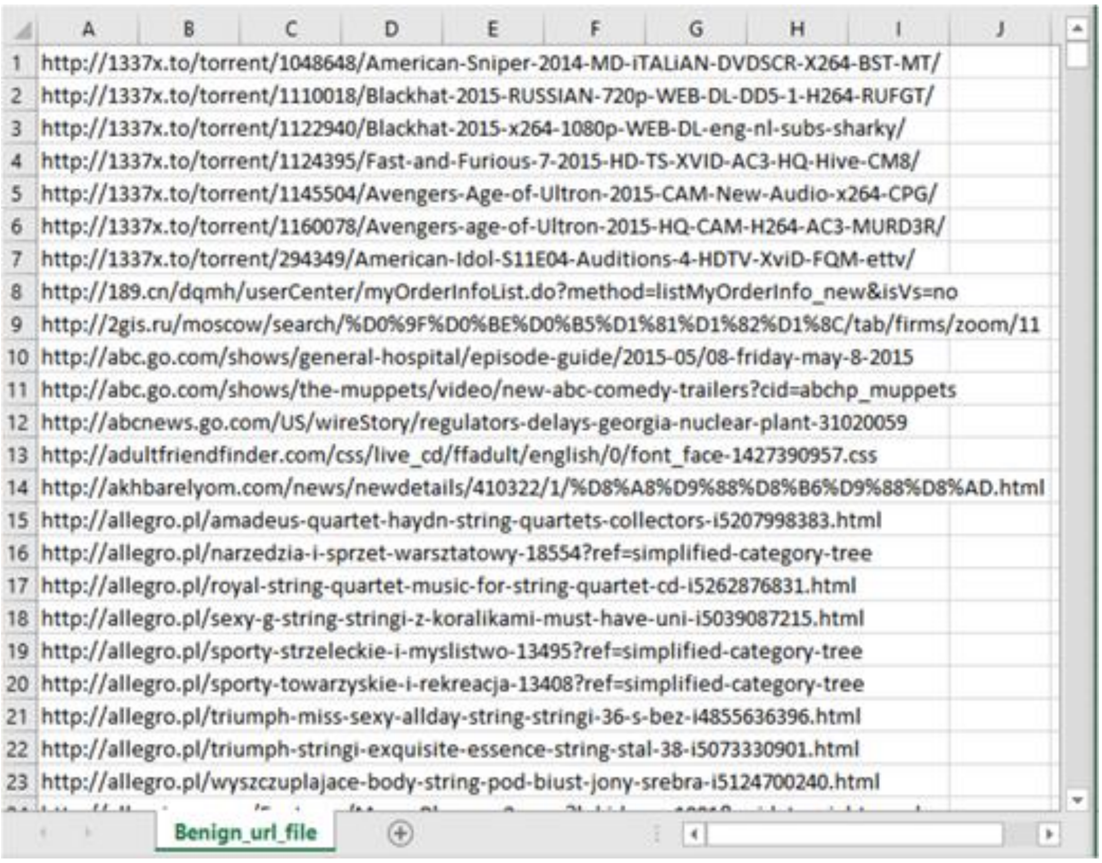


Figure I.: Architecture of Our Proposed System

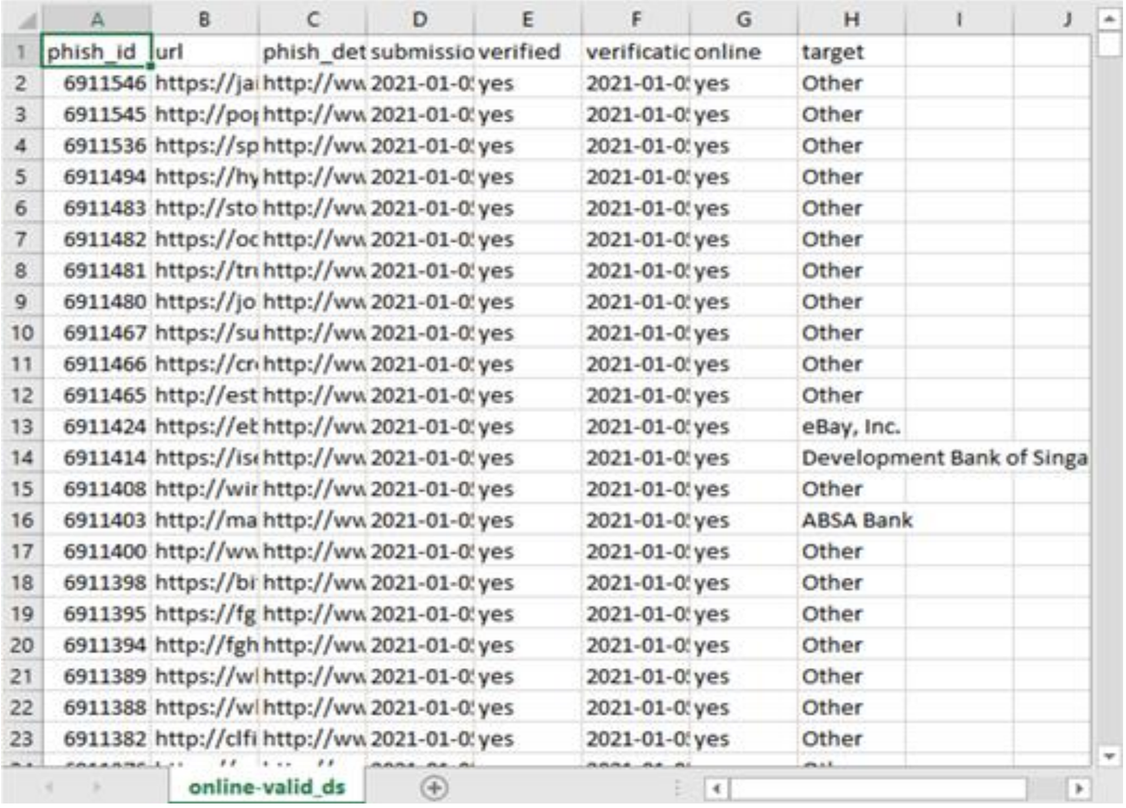
Data Set Collection

The data that generated the datasets from which the models were trained and tested were extracted from various open-source platforms known as Phish Tank. These datasets comprise of both legitimate and phishing URLs in multiple formats like CSV, JSON, and so on that get updated hourly.



	A
1	http://1337x.to/torrent/1048648/American-Sniper-2014-MD-iTALiAN-DVDSCR-X264-BST-MT/
2	http://1337x.to/torrent/1110018/Blackhat-2015-RUSSIAN-720p-WEB-DL-DD5-1-H264-RUFGT/
3	http://1337x.to/torrent/1122940/Blackhat-2015-x264-1080p-WEB-DL-eng-nl-subs-sharky/
4	http://1337x.to/torrent/1124395/Fast-and-Furious-7-2015-HD-TS-XVID-AC3-HQ-Hive-CM8/
5	http://1337x.to/torrent/1145504/Avengers-Age-of-Ultron-2015-CAM-New-Audio-x264-CPG/
6	http://1337x.to/torrent/1160078/Avengers-age-of-Ultron-2015-HQ-CAM-H264-AC3-MURD3R/
7	http://1337x.to/torrent/294349/American-Idol-S11E04-Auditions-4-HDTV-XviD-FQM-ettv/
8	http://189.cn/dqmh/userCenter/myOrderInfoList.do?method=listMyOrderInfo_new&isVs=no
9	http://2gis.ru/moscow/search/%D0%9F%D0%BE%D0%B5%D1%81%D1%82%D1%8C/tab/firms/zoom/11
10	http://abc.go.com/shows/general-hospital/episode-guide/2015-05/08-friday-may-8-2015
11	http://abc.go.com/shows/the-muppets/video/new-abc-comedy-trailers?cid=abchp_muppets
12	http://abcnews.go.com/US/wireStory/regulators-delays-georgia-nuclear-plant-31020059
13	http://adultfriendfinder.com/css/live_cd/ffadult/english/0/font_face-1427390957.css
14	http://akhbarelyom.com/news/newdetails/410322/1/%D8%A8%D9%88%D8%B6%D9%88%D8%AD.html
15	http://allegro.pl/amadeus-quartet-haydn-string-quartets-collectors-i5207998383.html
16	http://allegro.pl/narzedzia-i-sprzet-warsztatowy-18554?ref=simplified-category-tree
17	http://allegro.pl/royal-string-quartet-music-for-string-quartet-cd-i5262876831.html
18	http://allegro.pl/sexy-g-string-stringi-z-koralikami-must-have-uni-i5039087215.html
19	http://allegro.pl/sporty-strzeleckie-i-myslistwo-13495?ref=simplified-category-tree
20	http://allegro.pl/sporty-towarzyskie-i-rekreacja-13408?ref=simplified-category-tree
21	http://allegro.pl/triumph-miss-sexy-allday-string-stringi-36-s-bez-i4855636396.html
22	http://allegro.pl/triumph-stringi-exquisite-essence-string-stal-38-i5073330901.html
23	http://allegro.pl/wyszczuplajace-body-string-pod-biust-jony-srebra-i5124700240.html

Figure II: Phishing URLs Dataset



	A	B	C	D	E	F	G	H	I	J
1	phish_id	url	phish_det	submissio	verified	verificatic	online	target		
2	6911546	https://ja	http://ww	2021-01-0:	yes	2021-01-0:	yes	Other		
3	6911545	http://po	http://ww	2021-01-0:	yes	2021-01-0:	yes	Other		
4	6911536	https://sp	http://ww	2021-01-0:	yes	2021-01-0:	yes	Other		
5	6911494	https://hy	http://ww	2021-01-0:	yes	2021-01-0:	yes	Other		
6	6911483	http://sto	http://ww	2021-01-0:	yes	2021-01-0:	yes	Other		
7	6911482	https://oc	http://ww	2021-01-0:	yes	2021-01-0:	yes	Other		
8	6911481	https://tr	http://ww	2021-01-0:	yes	2021-01-0:	yes	Other		
9	6911480	https://jo	http://ww	2021-01-0:	yes	2021-01-0:	yes	Other		
10	6911467	https://su	http://ww	2021-01-0:	yes	2021-01-0:	yes	Other		
11	6911466	https://cr	http://ww	2021-01-0:	yes	2021-01-0:	yes	Other		
12	6911465	http://est	http://ww	2021-01-0:	yes	2021-01-0:	yes	Other		
13	6911424	https://et	http://ww	2021-01-0:	yes	2021-01-0:	yes	eBay, Inc.		
14	6911414	https://is	http://ww	2021-01-0:	yes	2021-01-0:	yes	Development Bank of Singa		
15	6911408	http://wir	http://ww	2021-01-0:	yes	2021-01-0:	yes	Other		
16	6911403	http://ma	http://ww	2021-01-0:	yes	2021-01-0:	yes	ABSA Bank		
17	6911400	http://ww	http://ww	2021-01-0:	yes	2021-01-0:	yes	Other		
18	6911398	https://bi	http://ww	2021-01-0:	yes	2021-01-0:	yes	Other		
19	6911395	https://fg	http://ww	2021-01-0:	yes	2021-01-0:	yes	Other		
20	6911394	http://fgh	http://ww	2021-01-0:	yes	2021-01-0:	yes	Other		
21	6911389	https://wl	http://ww	2021-01-0:	yes	2021-01-0:	yes	Other		
22	6911388	https://wl	http://ww	2021-01-0:	yes	2021-01-0:	yes	Other		
23	6911382	http://clfi	http://ww	2021-01-0:	yes	2021-01-0:	yes	Other		

Figure III: Legitimate URL Dataset

REVIEW OF RELATED WORKS

Phishing attacks severely affect national security, intellectual properties and the economy at large in a negative way as online businesses, banks, Internet users and government are the primary target [4]. An algorithm that would generate random credit card numbers was developed in the early 1990s, in an attempt to create fake American Online service provider (AOL) accounts.

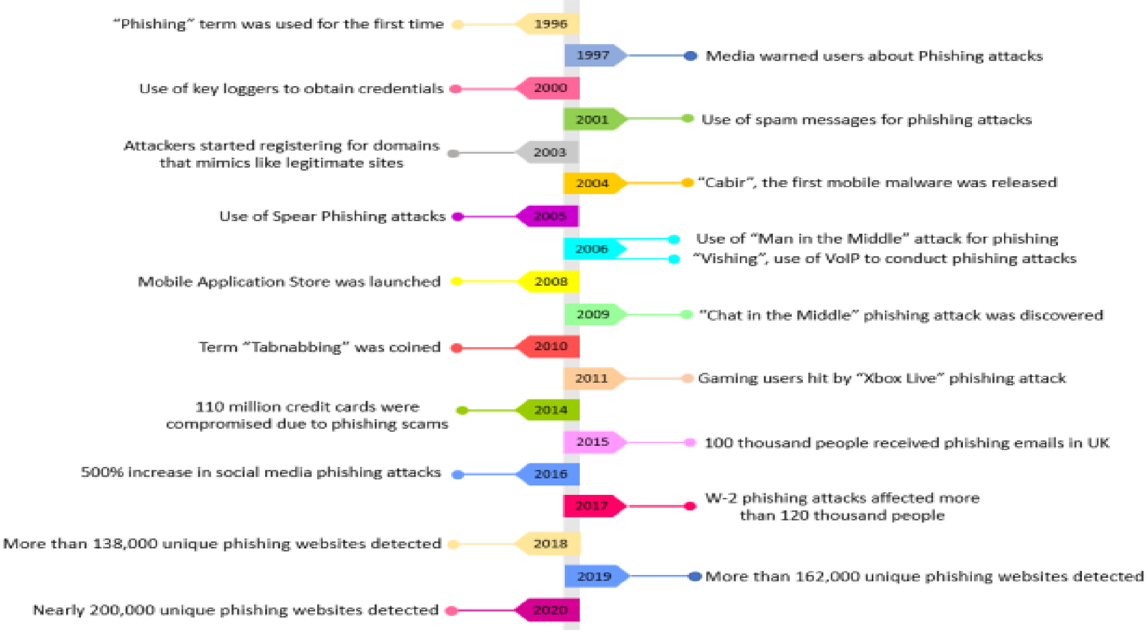

"Phishing" term was used for the first time
1996
1997
Media warned users about Phishing attacks
Use of key loggers to obtain credentials
2000
2001
Use of spam messages for phishing attacks
Attackers started registering for domains that mimics like legitimate sites
2003
2004
"Cabir", the first mobile malware was released
Use of Spear Phishing attacks
2005
2006
Use of "Man in the Middle" attack for phishing
"Vishing", use of VoIP to conduct phishing attacks
Mobile Application Store was launched
2008
2009
"Chat in the Middle" phishing attack was discovered
Term "Tabnabbing" was coined
2010
2011
Gaming users hit by "Xbox Live" phishing attack
110 million credit cards were compromised due to phishing scams
2014
2015
100 thousand people received phishing emails in UK
500% increase in social media phishing attacks
2016
2017
W-2 phishing attacks affected more than 120 thousand people
More than 138,000 unique phishing websites detected
2018
2019
More than 162,000 unique phishing websites detected
Nearly 200,000 unique phishing websites detected
2020

Figure IV: Evolution of phishing attacks

Phishing attack progression as shown in Figure 4[6] started in 1996 when the word ''phishing'' was first introduced in 1996, however, it spread via different information channels as time went on starting with spam messages, mobile malware to spear-phishing and so on. and This became an eye opener to researchers when a huge financial loss was made in 2014.

Due to the emergence of the Internet and its prevalence on media platforms, phishing has risen tremendously and has continued to rise [7] hence, becoming increasingly sophisticated, giving the phisher access to observe the activities of his victims while navigating the web, however, transversing any security boundaries with the victim. Phishing being a form of social engineering where attackers deceive their victims into installing malwares like ransomware thus revealing sensitive information can be averted when users protect themselves from phishing assaults using a variety of methods, such as the heuristic approach, rule-based approach, Visual Similarity-Based Phishing Detection Systems and a supervised machine learning (ML) approach [10].

Supervised Machine Learning algorithm is shown to be more extensively used for classification when compared to other ways of identifying phishing websites, hence giving a high accuracy in phishing detection and in a very short while.

Detection being a process of identifying an attempted computer intrusion is particularly based on the recurrence of the carrier signal, just like the radio broadcasting frequencies, but requires separating background noises from the weak signals just as in radio astronomy or constructing a hidden signal as seen in steganography.

Stegnographic analysis which refers to detection of concealed or hidden messages is in contrast to the detection of simply encrypted signals where the cipher text is usually identified even though it cannot be decoded. Steganalysis simply determines the probability of the existence of hidden messages making it an interesting distinction from other forms of detection.

In conclusion, the art of detection, also called 'following clues', is an attempt to reconstruct a sequence of events by identifying the relevant information concerning the situation.

Phishing Websites

A domain that duplicates an official website in appearance and name in other to deceive people into believing they are legitimate is referred to as phishing website [12].

Phishing had more changes in implementation in the early 2000s, in which the "love bug of 2000" is a typical example where victims were sent 'I LOVE YOU' email with an attachment that overwrites files on the victim's computer and copies itself to the user's contact list. That same year, different phishers began to register phishing websites.

In recent years, phishing websites appear frequently and poses as a new cyber security threat which has caused a great harm in data security and online financial services [13].

It has been assumed that the creation of most phishing websites is attributed to the vulnerability of most web servers, allowing phishers to host websites without the owner's knowledge or even host a new and independent web server just for phishing activities.

According to Shapiro (1992), the ability of the computer system to acquire knowledge and be able to use the acquired knowledge in self-improvement rather than being programmed with the knowledge is known as Machine Learning (ML) which is a branch of AI that enables machines to automatically gain knowledge and improve on the knowledge with minimal human intervention. Machine Learning also cuts across other scientific disciplines like cognitive science and statistics.

Machine Learning is vast following its production of basic statistical theories of learning processes, designed learning algorithms like speech recognition used in commercial systems [14].

EXPERIMENTAL RESULT AND ANALYSIS

Comparing the four leading ML models (Random Forest, Decision Tree, Multilayer Perceptrons and XGBOOST), XGBOOST shows to be more accurate as shown in Table II and figure V below.

Table II. Accuracy and Performance of the four leading Models

ML Model	Train Accuracy	Test Accuracy
XGBOOST	0.866	0.864
Multilayer Perceptrons	0.865	0.864
Decision Tree	0.814	0.812
Random Forest	0.818	0.811

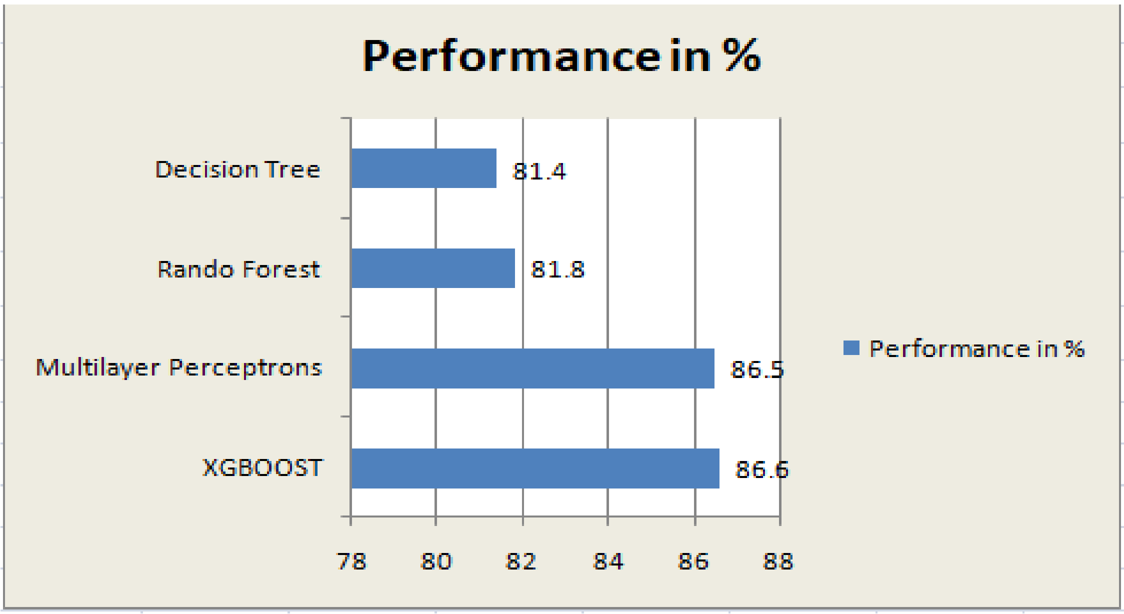


Figure V: Chart Accuracy Comparison of the Four Leading Models

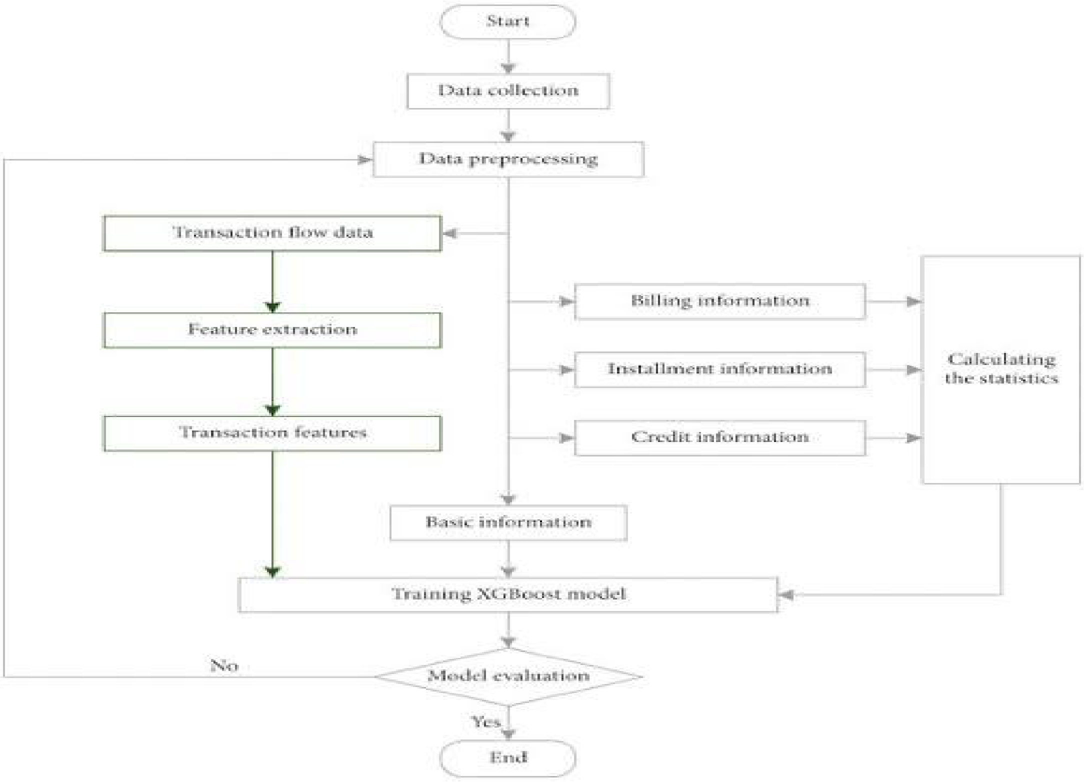
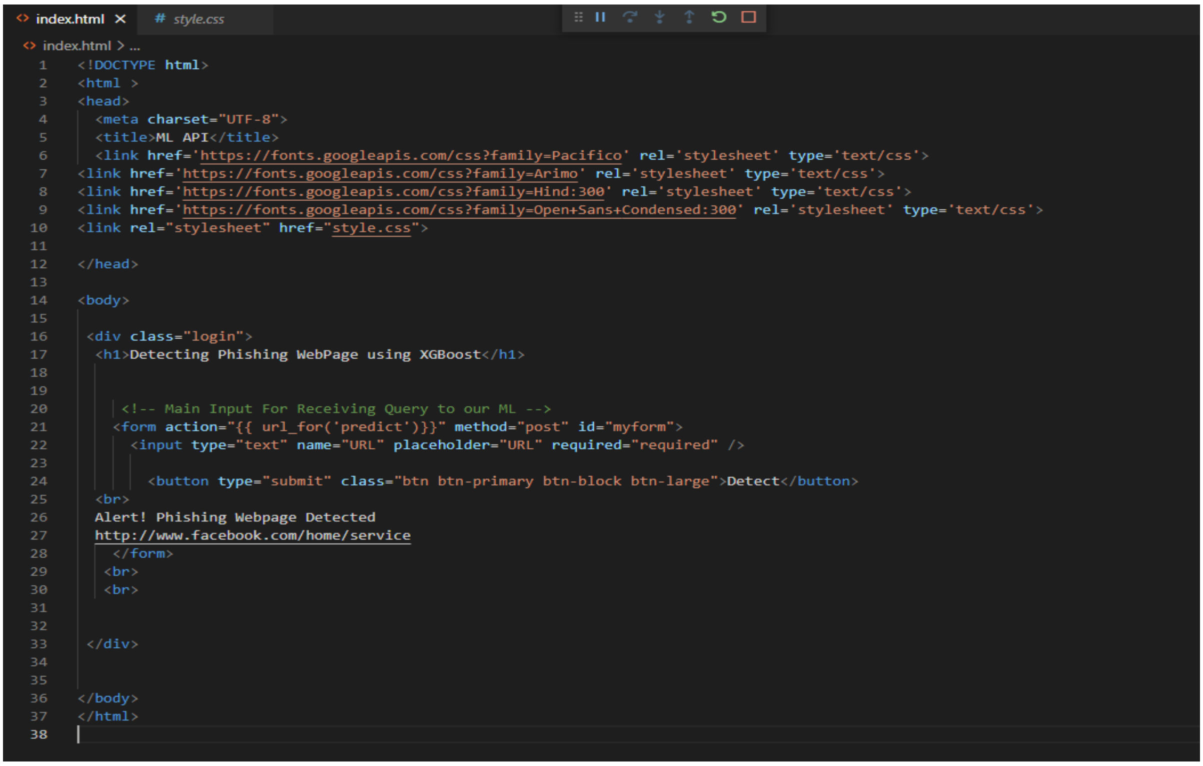
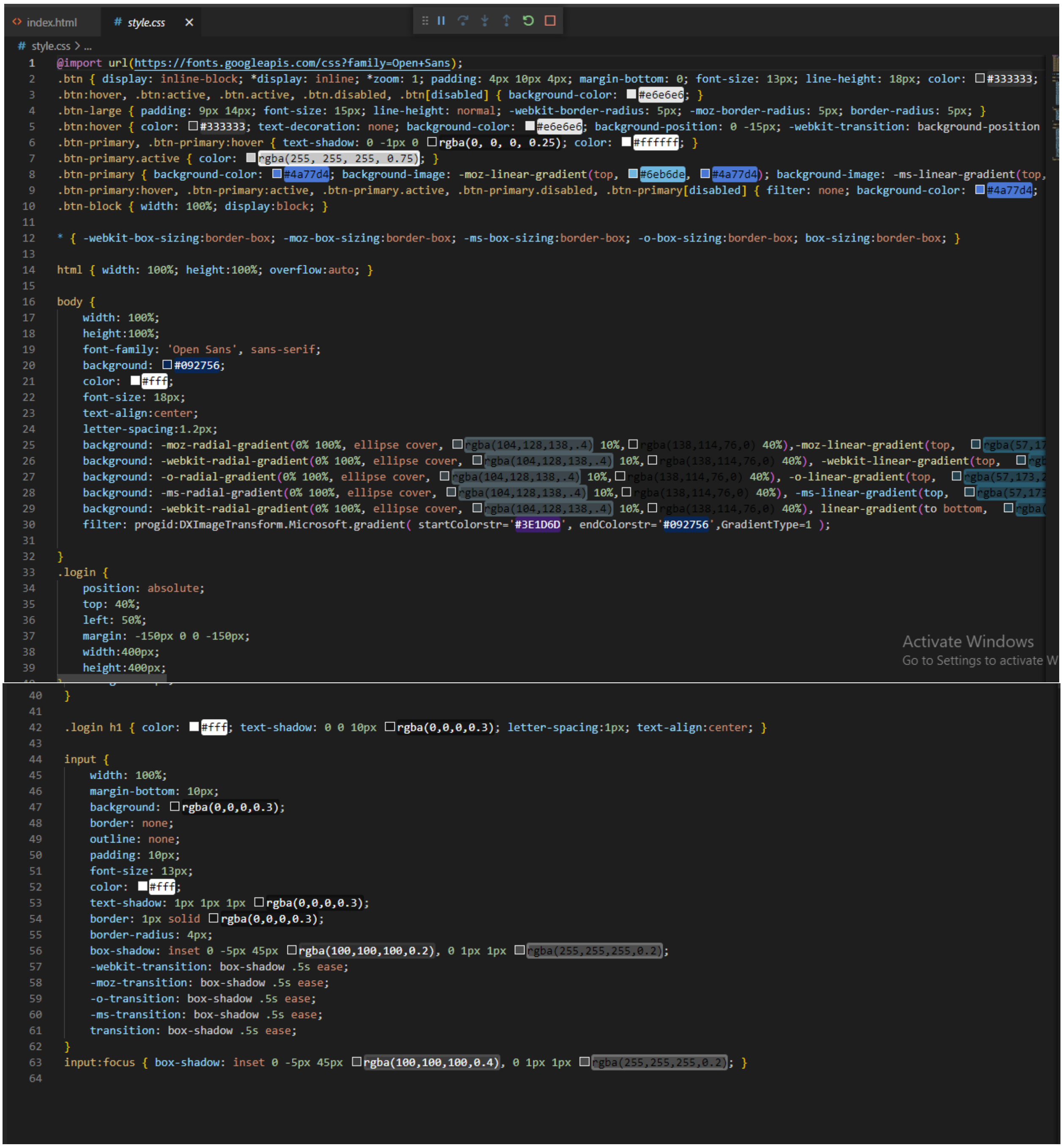


Figure VI: Visualization of XGBOOST Model



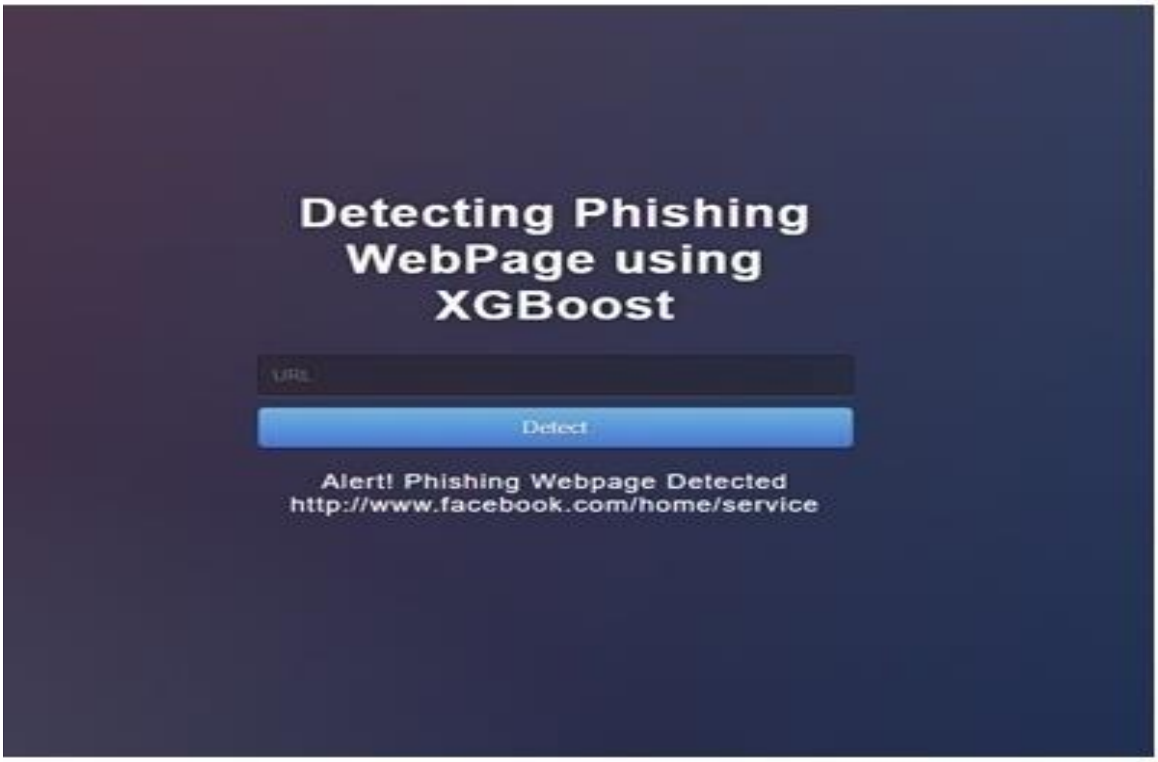
```html
<!DOCTYPE html>
<html >
<head>
  <meta charset="UTF-8">
  <title>ML API</title>
  <link href='https://fonts.googleapis.com/css?family=Pacifico' rel='stylesheet' type='text/css'>
<link href='https://fonts.googleapis.com/css?family=Arimo' rel='stylesheet' type='text/css'>
<link href='https://fonts.googleapis.com/css?family=Hind:300' rel='stylesheet' type='text/css'>
<link href='https://fonts.googleapis.com/css?family=Open+Sans+Condensed:300' rel='stylesheet' type='text/css'>
<link rel="stylesheet" href="style.css">

</head>

<body>

 <div class="login">
  <h1>Detecting Phishing WebPage using XGBoost</h1>


    <!-- Main Input For Receiving Query to our ML -->
    <form action="{{ url_for('predict')}}" method="post" id="myform">
      <input type="text" name="URL" placeholder="URL" required="required" />

        <button type="submit" class="btn btn-primary btn-block btn-large">Detect</button>
  <br>
  Alert! Phishing Webpage Detected
  http://www.facebook.com/home/service
    </form>
   <br>
   <br>


 </div>


</body>
</html>
```

Figure VII: A HTML webpage application



```css
@import url(https://fonts.googleapis.com/css?family=Open+Sans);
.btn { display: inline-block; *display: inline; *zoom: 1; padding: 4px 10px 4px; margin-bottom: 0; font-size: 13px; line-height: 18px; color: #333333;
.btn:hover, .btn:active, .btn.active, .btn.disabled, .btn[disabled] { background-color: #e6e6e6; }
.btn-large { padding: 9px 14px; font-size: 15px; line-height: normal; -webkit-border-radius: 5px; -moz-border-radius: 5px; border-radius: 5px; }
.btn:hover { color: #333333; text-decoration: none; background-color: #e6e6e6; background-position: 0 -15px; -webkit-transition: background-position
.btn-primary, .btn-primary:hover { text-shadow: 0 -1px 0 rgba(0, 0, 0, 0.25); color: #ffffff; }
.btn-primary.active { color: rgba(255, 255, 255, 0.75); }
.btn-primary { background-color: #4a77d4; background-image: -moz-linear-gradient(top, #6eb6de, #4a77d4); background-image: -ms-linear-gradient(top,
.btn-primary:hover, .btn-primary:active, .btn-primary.active, .btn-primary.disabled, .btn-primary[disabled] { filter: none; background-color: #4a77d4;
.btn-block { width: 100%; display:block; }

* { -webkit-box-sizing:border-box; -moz-box-sizing:border-box; -ms-box-sizing:border-box; -o-box-sizing:border-box; box-sizing:border-box; }

html { width: 100%; height:100%; overflow:auto; }

body {
    width: 100%;
    height:100%;
    font-family: 'Open Sans', sans-serif;
    background: #092756;
    color: #fff;
    font-size: 18px;
    text-align:center;
    letter-spacing:1.2px;
    background: -moz-radial-gradient(0% 100%, ellipse cover, rgba(104,128,138,.4) 10%,rgba(138,114,76,0) 40%),-moz-linear-gradient(top,  rgba(57,17
    background: -webkit-radial-gradient(0% 100%, ellipse cover, rgba(104,128,138,.4) 10%,rgba(138,114,76,0) 40%), -webkit-linear-gradient(top,  rgb
    background: -o-radial-gradient(0% 100%, ellipse cover, rgba(104,128,138,.4) 10%,rgba(138,114,76,0) 40%), -o-linear-gradient(top,  rgba(57,173,2
    background: -ms-radial-gradient(0% 100%, ellipse cover, rgba(104,128,138,.4) 10%,rgba(138,114,76,0) 40%), -ms-linear-gradient(top,  rgba(57,173
    background: -webkit-radial-gradient(0% 100%, ellipse cover, rgba(104,128,138,.4) 10%,rgba(138,114,76,0) 40%), linear-gradient(to bottom,  rgba(
    filter: progid:DXImageTransform.Microsoft.gradient( startColorstr='#3E1D6D', endColorstr='#092756',GradientType=1 );

}
.login {
    position: absolute;
    top: 40%;
    left: 50%;
    margin: -150px 0 0 -150px;
    width:400px;
    height:400px;
}

.login h1 { color: #fff; text-shadow: 0 0 10px rgba(0,0,0,0.3); letter-spacing:1px; text-align:center; }

input {
    width: 100%;
    margin-bottom: 10px;
    background: rgba(0,0,0,0.3);
    border: none;
    outline: none;
    padding: 10px;
    font-size: 13px;
    color: #fff;
    text-shadow: 1px 1px 1px rgba(0,0,0,0.3);
    border: 1px solid rgba(0,0,0,0.3);
    border-radius: 4px;
    box-shadow: inset 0 -5px 45px rgba(100,100,100,0.2), 0 1px 1px rgba(255,255,255,0.2);
    -webkit-transition: box-shadow .5s ease;
    -moz-transition: box-shadow .5s ease;
    -o-transition: box-shadow .5s ease;
    -ms-transition: box-shadow .5s ease;
    transition: box-shadow .5s ease;
}
input:focus { box-shadow: inset 0 -5px 45px rgba(100,100,100,0.4), 0 1px 1px rgba(255,255,255,0.2); }
```

Figure VIII: A CSS webpage application



Detecting Phishing WebPage using XGBoost

Alert! Phishing Webpage Detected
http://www.facebook.com/home/service

Figure X: XGBOOST Webpage design

As indicated in Fig.V, XGBOOST model leads in performance, hence used in our phishing detection web-page design as shown in Figure IX.

Using a web application that integrates the model with the highest accuracy based on the feature and algorithm used in distinguishing phishing URL from legitimate URL links, users can enter website URL links to determine whether they are legitimate or phishing.

CONCLUSION

From the above study, XGBOOST model shows more performance accuracy compared with Decision Tree, Random Forest and Multilayer Perceptrons. Hence, it is used in the design of a URL phishing detection webpage which detects if a URL link is authentic or not.

The Phishing detection approach implemented in this study is imperative in other to avoid and drastically reduce the chances of data theft and other cyber frauds achieved either through name or brand impersonation and subdomain attacks.

Required content-based features of both phishing and benign URLs of websites were extracted and phishing websites were easily predicted using trained machine learning models.

REFERENCES

[1] B. B. Gupta, A. Tewari, A. K. Jain, and D. P. Agrawal, "Fighting against phishing attacks: state of the art and future challenges," Neural Comput. Appl., vol. 28, no. 12, pp. 3629–3654, 2017, doi: 10.1007/s00521-016-2275-y.

[2] Anti-Phishing Working Group, "Phishing Activity Trends Report 3rd Quarter," no. November, pp. 1–9, 2021.

[3]. M. Jakobsson and S. Myers, "Phishing and Countermeasures: Understanding the Increasing Problem of Electronic Identity Theft," John Wiley Sons, 2006.

[4] Subasi, A., Molah, E., Almkallawi, F. and Chaudhery, T.J. (2017), *Intelligent phishing website detection using random Forest classifier*, International Conference on Electrical and ComputingTechnologies and Applications (ICECTA '17), IEEE, pp. 1-5.

[5] Kuyama, M., Kakizaki, Y., ... Sasaki,R.(2016).*Method for detecting a malicious domain by using whois and dns features*, The Third International Conference on Digital Security and Forensics (DigitalSec2016).

[6] G. Diksha and J. A. Kumar, ''Mobile phishing attacks and defence mechanisms: State of art andopen research challenges,''Comput. Secur., vol. 73, pp. 519–544, Mar. 2018, doi: 10.1016/j.cose.2017.12.006.

[7] Shrivas, A.K. and Suryawanshi, R. (2017). *Decision Tree Classifier for Classification of Phishing Website with Info Gain Feature.* Int. J. for Res. Appl. Sci. Eng. Technol.

[8] Verified Phishing URL, Available at: https://www.phishtank.com. Last accessed on September 22, 2017.

[9] Ankit,K.J and Gupta B.B. Phishing Detection: Analysis of Visual Similarity Based Approaches. National Institute of Technology, Kurukshetra, India, 10 January2017, ticle ID 5421046, 20 pages.

[10] Zhang L. and Zhan C. *Machine Learning in Rock Facies Classification: An Application of XGBOOST*. In International Geophysical Conference, Qingdao, China, 17-20 April 2017, pp. 1371-1374.

[11] Buber, E., Demir O., Sahingoz O. K. (2017). *Feature selections for the machine learning based detection of phishing websites*: International Artificial Intelligence and Data Processing Symposium (IDAP).

[12] Jain, A. (2016) *Complete Guide to Parameter Tuning in XGBOOST (with code in python)*: Complete guide to parameter tuning in XGBOOST (with code in Python).

[13] Kazemian,H. B. and Ahmed S. (2021)*Comparisons of machine learning techniques for detecting malicious webpages:*Expert Systems with Applications.

[14] Khonj M., Iraqi, Jones,A. (2013).*Phishing detection: a literature survey*. IEEEcommunications surveys & tutorials.

[15] Lord, N. (2018). *What is a Phishing Attack? Defining and Identifying Different Types of Phishing Attacks:*https://digitalguardian.com/blog/what-phishing-attack-defining-and-identifying-different-types-phishing-attacks.

[16] M. Jakobsson and S. Myers, "Phishing and Countermeasures: Understanding the Increasing Problem of Electronic Identity Theft," John Wiley Sons, 2006.

[17] Mustafa,A. and Nazife, B. (2015).Feature Extraction and Classification Phishing Websites Based on URL: IEEE.

[18] Pradeepthi, K.V. and Kannan, A. (2014). Performance Study of Classification Techniques for Phishing URL Detection: Sixth International Conference on Advanced Computing (ICoAC) IEEE.

[19] Rao,R. S. and Pais,A.R. (2018).*Detection of phishing websites using an efficient feature-based machine learning framework:*Neural Computing and Applications.

[20] Routhu S.R. and Alwyn, R.P. (2018). *Detection of phishing websites using an efficient feature-based machine learning framework*: In Springer.

[21] Sahoo D., Liu C. and Hoi S.C. (2017). *Malicious URL detection using machine learning: A survey*. arXiv:1701.07179.

[22] H. Zhang, G. Liu, T. W. S. Chow, and W. Liu, "Textual and visual content-based anti-phishing: a Bayesian approach," *IEEE Transactions on Neural Networks*, vol. 22, no. 10, pp. 1532–1546,2011

[23] S. Sheng, M. Holbrook, P. Kumaraguru, L. F. Cranor and J. Downs, "Who falls for phish?: a demographic analysis of phishing susceptibility and effectiveness of interventions", *Proceedings of the 28th international conference on Human factors in computing systems ser. CHI'10. New York NY USA:ACM*, pp. 373-382, 2010.

[24] C. H. Hsu, P. Wang, and S. Pu, "Identify fixed-path phishing attack by STC," in *Proceedings of the 8th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference (CEAS '11)*, pp. 172–175, ACM, Perth, Australia, September 2011.

[25] Ahmad, A., Anazida, Z., ... Oluwatobi, A. (2013). Feature Extraction Process: *A Phishing Detection Approach*: In IEEE communications and surveys.

[26] Anti-Phishing Working Group (APWG) (2019), *Phishing activity trends report, 1st quarter https://apwg.org/trendsreports_q3_2019*.