PhishSpy – A Phishing Detection Tool and Defensive Approaches

Vinayak Chaturvedi
Department of CEA,
GLA University,
Mathura, India
vinayak.chaturvedi_cs.csf19@gla.ac.in

Rajnish Kumar Gupta
Department of CEA,
GLA University,
Mathura, India
rajnish.gupta_cs.ccf19@gla.ac.in

Priyanshu Upadhyay
Department of CEA,
GLA University,
Mathura, India
priyanshu.upadhyay_cs.ccf19@gla.ac.in,

Asheesh Tiwari
Department of CEA,
GLA University,
Mathura, India
asheesh.tiwari@gla.ac.in

Abstract- The art to trick the victim into believing the fake scenarios as legitimate with the intention of getting the target to either download malware or take over personal information. Phishing has become the ultimate fashionable cybercrime among cybercriminals. The impact of phishing is adverse and can lead to unfavorable scenarios, indeed cybercrime. Since these attacks are increasing exponentially and cause huge damage and financial losses, the detection of phishing is of great importance and has also become an area of great interest. In this paper will discuss a few new tactics for detecting this phishing technique. However, there are already contrasting explanations in many papers, but phishing is very active and in action with new masks that were just discovered in 2022, and for detecting them, there is no algorithm or approach yet. This document describes the most frequent phishing tactics as well as the PhishSpy algorithmic (heuristics-based) tool created to detect phishing, which can discover phishing URLs and provide a suspect score as an output to the user. The PhishSpy algorithmic program features a catch rate of 95%. This paper discusses the methodology used to create this algorithmic program, as well as the implementation details and testing results.

Keywords- Suspectscore, Cybercrimes, Phishspy algorithm, Heuristics-based approach, Malware, Cybercriminals

I. Introduction

Nowadays, we go to the internet to know every little thing, to comprehend new things or skills, to communicate with anyone, for remittance, for receiving money from someone, for booking tickets, hotels, cabs for your ride, entertainment, Everything. We have become eminently reliant on the internet. In the era where the internet has become very crucial and a basic necessity for us, many crimes are also happening on the same internet, and we call them internet crimes or cybercrimes. Cyber-crime continues to generate new threats, attacks, tools, and strategies in a social network-powered world, allowing attackers to discover a flaw in complicated or well-controlled systems, inflicting greater harm, and sometimes remaining unexpected [15]. When we talk about Cybercrimes, the name of the Phishing technique comes to the top.

We've witnessed an exponential increase in the number of individuals performing transactions online in recent years, from making simple purchases to paying bills to banks, and even acquiring a mortgage or vehicle loan or paying their taxes. This increase in online transactions has unfortunately coincided with an increase in cyber-attacks. Phishing is an attack where the attacker makes use of a social engineering approach to perform identity theft. It focused on accumulating sensitive and confidential data such as usernames, passwords, credit card numbers, and even money by faking a legitimate entity in cyberspace [7]. This information is then used by criminals to steal the identity of the victim and hence commit further crimes by using this stolen identity. It has been observed that more than 90% of the time, phishing attacks originate from an email embedded with a malicious link that victims receive and make the victim trust that mail and redirect them to visit the attacker-controlled website at which they are tricked to enter their sensitive information.

There are four steps to a phishing attack. [1]:

- First, a criminal constructs a website that looks just like the official one.

- The attacker poses as a reputable organization or firm and provides a link to the people he intends to target using an URL website.
- The attacker is attempting to trick the victims into visiting a bogus website.
- Following the pleading process, the victims will go on to the website and provide the attacker with the essential personal information, with which the attacker will commence fraudulent operations with the victim.

Some technical equipment has been automated and streamlined to the point where non-technical cybercriminals can use it. Because these technology capabilities are widely accessible, the frequency of phishing attacks has increased. Social engineering schemes have been around since the Internet's beginning. Before the Internet was available to the general public, hackers used the telephone to impersonate trustworthy agents and get sensitive information, a practice known as vishing. [6]

Phishing attacks are becoming more common, and they are remarkably rising in both sophistication and frequency. As per the articles by the Anti-Phishing Working Group, approximately around 214345 distinctive phishing sites had been detected in total in the monthly report of September 2021. While the number of recent phishing attacks has more than quadrupled since early 2020, the numbers in September 2021 were lower than those in July 2021, when APWG reported 260642 attacks – the highest monthly attack account recorded by APWG. [2]

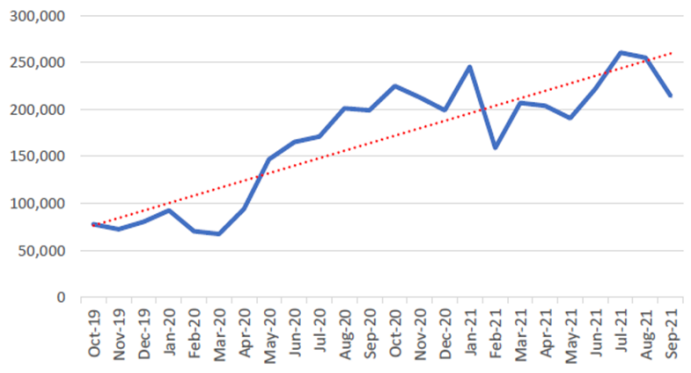

Figure 1. **Unique Phishing Attacks, Q4 2019 -3Q 2021**

From the above stats, we can analyze that there is a massive hike in the covid pandemic period from May 2020 to July 2021, the COVID-19 epidemic altered our lifestyles and prompted a widespread migration to digital platforms, making individuals more susceptible to cyber-crime. As a result, hackers launched enormous phishing attempts, taking advantage of the Covid-19 event's unusually unusual circumstances. According to the Phishing and Fraud Report's fourth edition, global phishing incidences risen by 220 percent above the yearly average during the pandemic.[3]

About 90% of data breaches occur as a result of Phishing. The frequency of phishing attacks might surge by 400% every year, according to the US Federal Bureau of Investigation (FBI) [8]. In terms of phishing scams, the most targeted sectors have long included financial institutions, social media businesses, SaaS / webmail services, and retailers, as shown in the pie chart.

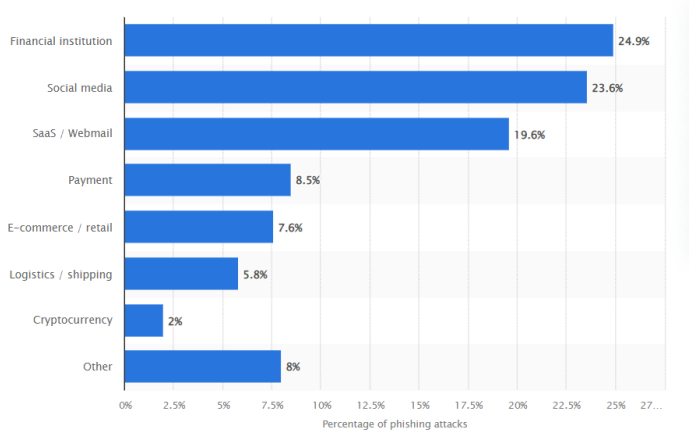

Figure 2. **Most- Targets Industries 3Q 2021**

Financial institution and payment provider cyber attacks continued to remain ordinary, accounting for 24.9 percent of all attacks. Social media platforms are the daily communication medium of today's world, and discoveries continue to grow. Because of this, it's like how threat actors started targeting mobile users, they've already started harassing social media, too. A survey by Google found that email phishing is on average 13.7 percent efficacious. In contrast, a later study by Blackhat found that social media phishing attacks were up to 66% effective [9]. The top-targeted sectors were rounded out by social media, according to the Swiss Cyber Institute, LinkedIn phishing messages accounts for 47% of all social media phishing attempts [8]. Logistics/Shipping is now the second most targeted category, with intimidating actors continuing to take advantage of regular e-commerce promotions by targeting buyers and shipping companies directly. DHL ranks second on LinkedIn, accounting for 14% of criminal attempts to steal sensitive information [8]. Phishing attempts targeting cryptocurrency targets, such as cryptocurrency exchanges and wallet providers, accounted for around 2% of total cyberattacks. Phishing and cryptocurrency scams have taken advantage of the Russian invasion of Ukraine, setting up donation schemes and utilizing the crisis to harvest data and bitcoin from victims, according to the Q1 2022 Cofense Phishing Intelligence

Trends. To target prospective phishing victims, threat actors utilized email subject lines like "Ukraine Donations" and "Help save children from Ukraine." [4]
In this paper, we examine some of the most frequent phishing tactics as well as various anti-phishing strategies. We concentrate on an automated method for identifying phishing sites that we have created. This paper provides a step that uses a combination of phishing detection techniques to identify ongoing phishing attacks with an accuracy higher than 96%. The remaining part of the paper is classified as follows. The section 2 represents the related work in phishing. The suggested methodology's elements are defined in section 3. The conclusion is presented in section 4.

II. Related work

Today, various anti-phishing methods have been proposed to minimize the phishing sites. After extensive research on this problem, we came up with the result that these strategies can be categorized into five approaches mainly-

i. Lists Based Approach: This strategy includes two types of lists, namely the blacklist, the list containing all the phishing URLs, and the whitelist, which is the list containing all the legitimate URLs. When a browser opens a page, it requests a blacklist to see if the URL presently being viewed is on it. If so, appropriate action may be taken against it. If not, the page is considered as legitimate [11]. PhishTank and OpenPhish are the widely used blacklist databases containing all the phishing sites.

ii. Heuristics-Based Approach: This methodology is based on using numerous discriminative features extracted from phishing web pages after comprehending and evaluating their structure. The approach for processing these characteristics has a big impact on how well and precisely web pages are classified. This covers features such as IP address in the domain section, domain age, picture & link source, right-click blocked, and so on.[12]

iii. Hybrid-Based Approach: This approach incorporates both the list-based and heuristic-based approaches. A combination of machine learning algorithms is another hybrid model [10]. In this type of model, the first algorithm is used to train a collection of data, and the results are then passed to the second training method.

iv. Fuzzy Logic Based Approach: This technique provided a unique methodology based on comparing a few rules experimentally after gathering different characteristics from a list of websites, as disclosed. Those factors vary between the three uncertain values “Legitimate”, “Doubtful”, and “Phishing” [13].

v. Content Based Approach: This strategy is useful for analyzing page content in detail. Creating classifiers and extracting features from page content and third-party services like search engines and DNS servers [14]. TF-IDF (term frequency-inverse document frequency) is a well-known information retrieval technique for comparing and distinguishing documents as well as retrieving documents from a large entity.

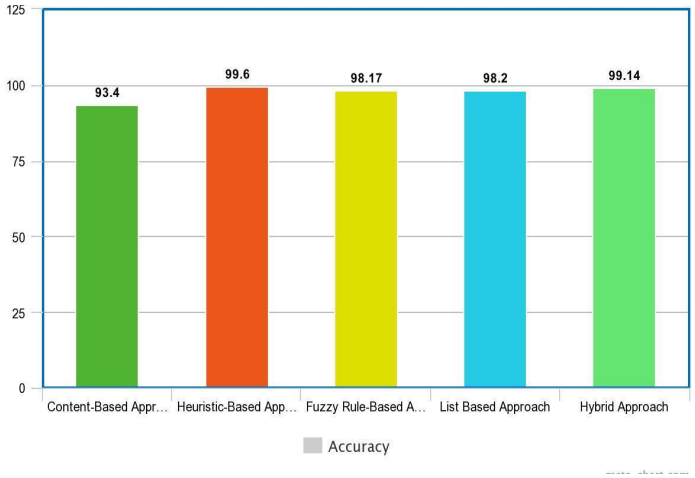

Figure 3. **Accuracy Percentage**

Many researchers have given many heuristic-based approaches. Based on those studies, we developed our larger set of Heuristic-based approaches. Some of these are:

- **Long Subdomain**:

  An attacker uses a long subdomain phishing attack in which attackers set up the subdomain similar to any legit company/organization and since the subdomain is too long many browsers on Android show the subdomain part of the domain, which tricks users into believing the website they are on is legit. Example below :
  long-subdomain.google.communication.abc.com

  In this Heuristics Based Approach, we detect this type of phishing attack by using algorithms like *checkUrlLength(url)*. This function checks the length of the given URL. If greater than 54, there will be a higher chance of Phishing.

- **Domain Age:**

  The majority of attackers create a newly registered domain as a phishing page because by hosting and domain policy, phishing pages domain and hosting do not last longer than a month, and hence age of the domain of the phishing is very small.

  In this Heuristics Based Approach, we detect this type of phishing attack by using algorithms like *checkDomainAge(url)*. This function checks the age of the domain. If less than 182, it is considered as Phishing.

- **Shorten URLs:**

  Attackers use shortened URL services to hide suspicious URLs.

  In this Heuristics Based Approach, we detect this type of phishing attack by using an algorithm like *checkShortenUrl()*. This function checks the URL in the shortened URL provider database.

- **Too Many Redirections:**

  In the many phishing pages, we have analyzed that the phishing page redirects so many times.

  In this Heuristics Based Approach, we detect this type of phishing attack by using algorithms like *checkRedirection(url)*. This function checks for the redirection history of the page.

- **Content Ratio of Own Domain To Another Domain:**

  When making or creating a phishing page, the attacker just clones the site and leaves all the images, and javascript URL pointed to the original website.

  In this Heuristics Based Approach, we detect this type of phishing attack by using algorithms like *checkRequestURL(url)*. This function checks for the ratio of content in the web page with its own domain vs. another pointed domain.

- **Checking the Phishing Databases:**

  There are various phishing databases available which are regularly being updated. These databases contain all the lists of phishing websites. OpenPhish, Phishing.Database, PhishStats are the most famous databases available on the internet.

  In this Heuristics Based Approach, we detect this type of phishing attack by using algorithms like *checkPhishDatabase(url)*. This function checks for the URL in the database. If found in the database, the URL is Phishing.

- **Open Ports in Domain:**

  If a domain is providing a service on the web, only port 80 or 443 should be open but for increasing the attack surface, attackers usually open non-standard ports.

  In this Heuristics Based Approach, we detect this type of phishing attack by using an algorithm like *checkOpenPorts(url)*. This function checks the URL for the open ports and compares it with the desired list of open ports. Hence predicts the website to be legitimate or not.

- **Check Subdomain level Order:**

  Usually, attackers create multi-level subdomains to create the desired name of the domain.

  Example: flipkart.shopping.shopz.com

  Here, the order of subdomains is 3, and hence it is suspicious.

- **Missing/Suspicious DNS Record:**

  The claimed identity of the domain looks suspicious and sometimes has no DNS record.

  In this Heuristics Based Approach, we detect this type of phishing attack by using an algorithm like *checkDNSRecord(url)*. This function checks the DNS record of the URL. Having no DNS records signifies that the provided domain is Phishing.

III. PROPOSED WORK

So far, proposed approaches to phishing detection are good to go and ready to prevent any user from phishing to a large extent, but there are some flaws. To fill this gap, we added some more features to it so that the user/environment/internet can be a safer place for users who are not aware of the fact that internet cybercriminals can destroy or vandalize anyone's life. We added features like the Browser in the Browser attack (BITB) detector, Full-Screen API Phishing detector, and Totally Blank Address bar (Work on Android Devices) detector.

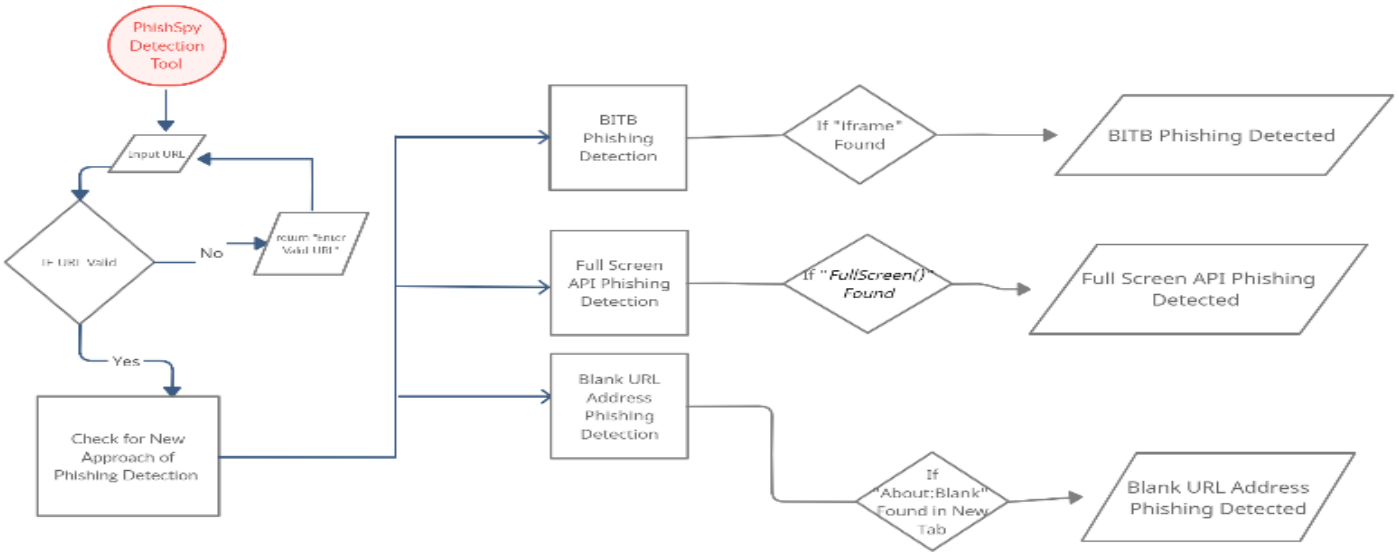

Figure 4. **PhishSpy: Process Flow Diagram**

1. BITB (Browser in the browser) Attack Scenario:

BITB attack is an advanced and more sophisticated phishing attack in which the attacker uses a phishing technique that simulates a browser window within the browser to spoof a legitimate domain.

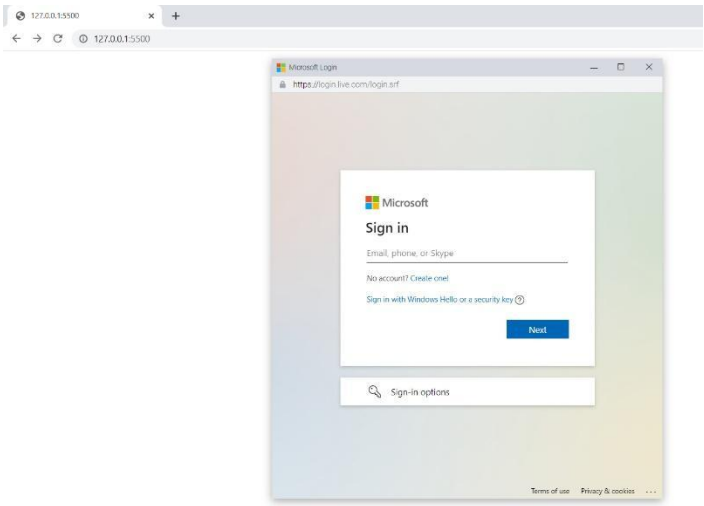

Figure 5. **BITB Phishing Example**

1a. Proposed Mitigation Approach:

Nowadays, there are many phishing pages that are in trend. After analyzing and researching (over the Phishtank database), we found there is something common in them which is a browser in the browser attack. After looking at various phishing pages, what we got, in conclusion, is that they all are using the "iframe" tag. So, after detecting them, we came up with an algorithm that is typewritten below.

```
def detectBITB(URL):
    if valid(URL):
    checkSourceCode(URL):
        if "iframe" in checkSourceCodeOutput:
          suspectScore = suspectScore + 1
             return suspectScore
        else:
           call next functions
    else:
     return "Enter Valid URL/IP"
```

Figure 6. **Pseudo Code for BITB Detection**

1b. Explanation: As you can see in the above pseudo-code, we divided the detection into some functions like the *valid(URL)* function and *checkSourceCode(URL)* function. The *valid(URL)* function is checking whether the URL given by the user is active or not i.e. whether the phishing website is live or not, or if the URL given by the user is even valid or not. Then we move on to the *checkSourceCode()* function in which we have applied the if condition; if there is an "iframe" present in the output of *checksourcecode()* function, then there might be a chance of BITB attack.

2. Full-Screen API Phishing Attack Scenario:

Another method is using the Full-Screen API of the browser, which opens a new tab with Full-Screen containing a phishing page. In this method, attacker creates a fake page with a button or link pointing to any legit site such as facebook.com. After victim clicks on the button from the attacker's page, the browser goes into the Full-Screen mode with a phishing page, which doesn't show any address bar, the victims think it's the legit site, and they fall into entering their credentials or sensitive information on that page which is controlled by the attacker's.

The Fullscreen API allows web developers to display web content in a full-screen mode completely.

Note that most browsers have a full-screen function that activates the user for some time. The HTML5 Fullscreen API allows the web developer to access this same functionality, and more importantly, the developer can launch it systematically. This is good because an attacker can design a full-screen button that looks like part of their site. [5]

Attacker can trigger full screen mode with this code:

elementToMakeFullscreen.requestFullScreen();

2a. Proposed Mitigation Approach:

```
def detectFullScreenAPI(URL):
    if valid(URL):
    checkSourceCode(URL):
        if "FullScreen()" in checkSourceCodeOutput:
          suspectScore = suspectScore + 1
             return suspectScore
        else:
           call next functions
    else:
     return "Enter Valid URL/IP"
```

Figure 7. **Pseudo Code for Full Screen API Phishing Detection**

2b. Explanation: Likewise, we go to *checkSourceCode(URL)* function in which if there is If *RequestFullScreen()* present in the output of *checkSourceCode()* function, then it might be Full-Screen API Phishing.

3. Totally Blank Address Bar Phishing Scenario:

This method works mainly on Android browsers in which a fake page executes a javascript code on click, which then opens a new tab with a phishing page having no URL or any words in the address bar, we can say totally blank. This doesn't look suspicious since there's nothing in the address bar which makes users believe they are on a legit site and end up providing their data.

3a. Proposed Mitigation Approach:

```
def detectBlankAddress(URL):
    if valid(URL):
    checkSourceCode(URL):
        if SuspeciousRedirectiontoNewTab() in checkSourceCodeOutput:
            if NoURL() in NewTab:
                  suspectScore = suspectScore + 1
                  return suspectScore
        else:
           call next functions
    else:
     return "Enter Valid URL/IP"
```

Figure 8. **Pseudo Code for Totally Blank Address Bar Phishing Detection**

3b. Explanation: As you can see in the above pseudo-code, we divided the detection into some functions like the *valid(URL)* function, *SuspeciousRedirectionToNewTab()*, *NoURL()* function, and *checkSourceCode(URL). The valid(URL) function* checks whether the URL given by the user is active or not i.e. whether the phishing website is live or not, or if the URL given by the user is even valid or not, then we move on to *checkSourceCode()* function. *SuspeciousRedirectionToNewTab()* checks every link that redirects to a new tab. Similarly, *NoURL()* function checks if the url bar is empty or not. If it is empty, then it might be a Totally Blank Address Bar Phishing.

IV. CONCLUSION

PhishSpy is a heuristic-based tool that comprises a new phishing detection technique or algorithm. In this algorithm, we encountered 3 new phishing trends that are more harmful and in action nowadays. It is critical to identify these kinds of fashionable attacks and inform users so that they may defend themselves. Various effective methodologies have been developed till now with the goal of detecting and blocking phishing sites, but somehow attackers are still able to bypass current tools and methodologies and are capable of accessing victims. Phishing is a persistent and complicated issue, and it is constantly changing its ways of attacking victims. One Single approach is not able to recognize all phases of phishing attempts, as phishing uses a wide variety of variants, and their attack studies are different as well. So an advanced tool with some new methodology like PhishSpy is very necessary. In the future, we will be adding newer algorithms to detect new masks of phishing methodologies, and we are also aiming to make a browser extension for easier phishing alerts.

V. References

[1] *Hicham Tout, William Hafner "Phishpin: An identity-based anti-phishing approach" in proceedings of international conference on computational science and engineering, Vancouver, BC, pages 347-352, 2009.*

[2] *Anti-Phishing Working Group (2021). Phishing Activity Trends Report (3 th Quarter 2021).*

[3] Article by F5, *Phishing Attacks Soar 220% During COVID-19 Peak as Cybercriminal Opportunism Intensifies*

[4] *Report by Security Magazine*, *4 Phishing Trends Observed in Q1 2022*

[5] *Using the HTML5 Fullscreen API for Phishing Attacks*

[6] *Journal of Computing::Phishing, SMiShing & Vishing*

[7] *Handbook of Information and Communication Security (Phishing Attacks and Countermeasures)*

[8] *Phishing-attack-statistics trends 2022*

[9] *Social media phishing and its affects article by waterstons*

[10] *P. Singh, Y. P. S. Maravi and S. Sharma, "Phishing Websites Detection through Supervised Learning Networks",* 2015 *International Conference on Computing and Communications Technologies (ICCCT), pp. 61-65, 2015.*

[11] *Xiang, G., Pendleton, B. A., Hong, J. I., and Rose, C. P.A hierarchical adaptive probabilistic approach for zero hour phish detection. In Proceedings of the 15th European Symposium on Research in Computer Security (ESORICS"10). 268–285, 2010.*

[12] *Rami M. Mohammad., Fadi Thabtah., Lee McCluskey: "Phishing Website Features".*

[13] *Phinding Phish: Evaluating Anti-Phishing Tools, Yue Zhang, Serge Egelman, Lorrie Cranor, and Jason Hong, In Proceedings of the 14th Annual Network & Distributed System Security Symposium (NDSS 2007).*

[14] *Almaha Abuzuraiq., Mouhammd Alkasassbeh., Mohammad Almseidin: Intelligent Methods for Accurately Detecting Phishing Websites*

[15] Asheesh Tiwari; Vibhu Mehrotra; Shubh Goel; Kumar Naman; Shashank Maurya; Ritik Agarwal, "Developing Trends and Challenges of Digital Forensics"