MALICIOUS URL DETECTION USING MACHINE LEARNING

1.N. VENKATESH, 2.V. TEJASWINI, 3.G. SOUMYA,4. T. SHIVA PRIYA

1.ASSISTANT PROFESSOR, 2,3&4.UG SCHOLAR

DEPARTMENT OF ECE, MALLA REDDY ENGINEERING COLLEGE FOR WOMEN, HYDERABAD

ABSTRACT In recent years, with the increasing use of mobile devices, there is a growing trend to move almost all real-world operations to the cyberworld. Although this makes easy our daily lives, it also brings many security breaches due to the anonymous structure of the Internet. Used antivirus programs and firewall systems can prevent most of the attacks. However, experienced attackers target on the weakness of the computer users by trying to phish them with bogus webpages. These pages imitate some popular banking, social media, e-commerce, etc. sites to steal some sensitive information such as, user-ids, passwords, bank account, credit card numbers, etc. Phishing detection is a challenging problem, and many different solutions are proposed in the market as a blacklist, rule-based detection, anomaly-based detection, etc. In the literature, it is seen that current works tend on the use of machine learning-based anomaly detection due to its dynamic structure, especially for catching the “zero-day” attacks. In this paper, we proposed a machine learning-based phishing detection system by using eight different algorithms to analyze the URLs, and three different datasets to compare the results with other works. The experimental results depict that the proposed models have an outstanding performance with a success rate.

INTRODUCTION Phishing is the most commonly used social engineering and cyber attack. Through such attacks, the phisher targets naïve online users by tricking them into revealing confidential information, with the purpose of using it fraudulently. In order to avoid getting phished, users should have awareness of phishing websites. Have a blacklist of phishing websites which requires the knowledge of website being detected as phishing. Detect them in their early appearance, using machine learning and deep neural network algorithms of the above three, the machine learning based method is proven to be most effective than the other methods. Even then, online users are still being trapped into revealing sensitive information in phishing websites. A phishing website is a common social engineering method that mimics trustful uniform resource locators (URLs) and webpages. The objective of this project is to train machine learning models and deep neural nets on the dataset created to predict phishing websites. Both phishing and benign URLs of websites are gathered to form a dataset and from them required URL and website content-based features are extracted. The performance level of each model is measures and compared. The phishing website has evolved as a major cybersecurity threat in recent times. The phishing websites host spam, malware, ransomware, drive-by exploits, etc. A phishing website many a time look-alike a very popular website and lure an unsuspecting user to fall victim to the trap. The victim of the scams incurs a monetary loss, loss of private information and loss of reputation. Hence, it is imperative to find a solution that could mitigate such security threats in a timely manner. Traditionally, the detection of phishing websites is done using blacklists. There are many popular websites which host a list of blacklisted websites, e. g. PhisTank. The blacklisting technique lack in two aspects, blacklists might not be exhaustive and do not detect a newly generated phishing website. In recent times machine learning techniques have been used in the classification and detection of phishing websites. In, this paper we have compared different machine learning techniques for the phishing website. In our daily life, we carry out most of our work on digital platforms. Using a computer and the internet in many areas facilitates our business and private life. It allows us to complete our transaction and operations quickly in areas such as trade, health, education, communication, banking, aviation, research, engineering, entertainment, and public services. The users who need to access a local network have been able to easily connect to the Internet anywhere and anytime with the development of mobile and wireless technologies. Although this situation provides great convenience, it has revealed serious deficits in terms of information security. Thus, the need for users in cyberspace to take measures against possible cyber-attacks has emerged. Attacks can be carried out by people such as cybercriminals, pirates, or non-malicious (white-capped) attackers and hacktivists. The aim is to reach the computer or the information it contains or to capture personal information in different ways. The attacks, as internet worms (Morris Worm), started in 1988, and they have been carried out until today. These attacks are mainly targeted in the following areas: fraud, forgery, force, shakedown, hacking, service blocking, malware applications, illegal digital contents and social engineering. Reaching with a wide range of target users, attackers aim to get a lot of information and/or money. According to Kaspersky's data, the average cost of an attack in 2019 (depending on the size of the attack) is between $ 108K and $ 1.4 billion. In addition, the money spent on global security products and services is around $ 124 billion. Among these attacks, the most widespread and also critical one is “phishing attacks”. In this type of attack, cybercriminals especially use an email or other social networking communication channels. Attackers reach the victim users by giving the impression that the post was sent from a reliable source, such as a bank, e-commerce site, or similar. Thus, they try to access sensitive information of them. Attackers then access their victims’ accounts by using this information. Thus, it causes pecuniary loss and intangible damages.The method of reaching target users in phishing attacks has continuously increased since the last decade. This method has been carried out in the 1990s as an algorithm-based, in the early 2000s based on e-mail, then as Domain Spoofing and in recent years via HTTPs. Due to the size of the mass attacked in recent years, the cost and effect of the attacks on the users have been high. The average financial cost of the data breach as part of the phishing attacks in 2019 is $ 3.86 million, and the approximate cost of the BEC (Business Email Compromise) phrases is estimated to be around $ 12 billion. Also, it is known that about 15% of people who are attacked are at least one more target. With this result, it can be said that phishing attacks will continue to being carried out in the ongoing years. Figure 1 also supports this idea and show the number of phishing sites in 2019, and as can be seen from it, there is an increasing trend in this type of attack. In this regard, regular reports published by APWG (Anti Phishing Working Group) are an important guide for the researchers. According to the reports, the number of phishing sites is reached to approximately 640,000 sites were determined in 2018, and in the first three quarters of 2019, this number was reported as 629,611 [6]. Reports for the last quarter of 2019 have not been published yet. However, it can be said that the phishing attacks not only continue, but also there will be an increase in the number of attack types compared to the previous year. This increase indicates that phishing attacks are used more by attackers. Because they are easy to design. Phishing attacks are based on the attacker’s creation of a fake website, as depicted in Figure 2. First, a phisher makes fake websites, including a phishing kit.

Then, the victim is directed to the fake website with the prepared email. Believing that the e-mail and URL are secure, the victim uses the fake website by clicking on the URL. After this moment, the Phishing kit receives the victim's credentials and sends it to the phisher. Finally, Phisher makes fake earning from the legitimate website using the victim's credentials. These sites generally have very similar or even identical visuals. In an e-mail that is thought to be sent from a trusted source, the target is directed to this fake website. The target accesses the website at the relevant URL via e-mail, which she/he finds reliable and writes the information that the attacker wants to obtain. The attacker receives the necessary information and uses it in the real system.In this way, the attacker gets information and / or earnings. Reliable e-mail contents are created in different ways for the victim to believe. Previously, e-mails with low probability offers, urgent texts, links, or attachments that may be relevant and unusual senders were used. Today, reliable organizations or similar links to these organizations are preferred. Attackers prefer reaching to victims by using a secure communication protocol, and the real URL is served by changing in a way that is close to the original. At this stage, if the victim knows the website is fake, he can protect himself from the attack. It is very difficult for the victim to detect the attack by himself, because mainly this type of messages gave some alert messages to the users, and aims to make panic for entering his confidential data to the forwarded page. Therefore, different decision support or detection systems have been developed to protect the end user against phishing attacks. Different approaches are used in these systems, such as Blacklists, Rule-based systems, Similarity-based systems, and Machine Learning based systems, etc. The literature was reviewed in detail, and the studies in this context were examined carefully. Currently, machine learning-based systems are especially preferred for its protection mechanism to the zero-day attacks. Therefore, in this paper, it is aimed to implement a phishing detection system based on a machine learning algorithm for investigating the URL address of the target web page. With the idea of existing improvable ways of the designed system, it is aimed at the detection of phishing attacks in a short time, without the need for third-party services, and also without waiting for the blacklists to be updated.The project is organized as follows: in the next section, the literature review is included. In the third section, the details of the designed system are explained. In the fourth section and fifth section, the results obtained in the experiments are shared, and conclusion and future studies are drawn, respectively.

LITERATURE SURVEY 1. ALTYEB TAHA," INTELLIGENT ENSEMBLE LEARNING APPROACH FOR PHISHING WEBSITE DETECTION BASED ON WEIGHTED SOFT VOTING" The continuous development of network technologies plays a major role in increasing the utilization of these technologies in many aspects of our lives, including e-commerce, electronic banking, social media, e-health, and e-learning. In recent times, phishing websites have emerged as a major cybersecurity threat. Phishing websites are fake web pages that are created by hackers to mimic the web pages of real websites to deceive people and steal their private information, such as account usernames and passwords. Accurate detection of phishing websites is a challenging problem because it depends on several dynamic factors. Ensemble methods are considered the state-of-theart solution for many classification tasks. Ensemble learning combines the predictions of several separate classifiers to obtain a higher performance than a single classifier. This paper proposes an intelligent ensemble learning approach for phishing website detection based on weighted soft voting to enhance the detection of phishing websites. First, a base classifier consisting of four heterogeneous machine-learning algorithms was utilized to classify the websites as phishing or legitimate websites. Second, a novel weighted soft voting method based on Kappa statistics was employed to assign greater weights of influence to stronger base learners and lower weights of influence to weaker base learners, and then integrate the results of each classifier based on the soft weighted voting to differentiate between phishing websites and legitimate websites. The experiments were conducted using the publicly available phishing website dataset from the UCI Machine Learning Repository, which consists of 4898 phishing websites and 6157 legitimate websites. The experimental results showed that the suggested intelligent approach for phishing website detection outperformed the base classifiers and soft voting method and achieved the highest accuracy of 95% and an Area Under the Curve (AUC) of 98.8%. Due to their flexibility, convenience, and simplicity of use, the number of web users who utilize online services, e-banking, and online shopping has increased rapidly in recent years. This massive increase in the use of online services and e-commerce has encouraged phishers and cyber attackers to create misleading and phishing websites in order to obtain financial and other sensitive information. Online phishing sites typically utilize similar page layouts, fonts, and blocks to imitate official web pages in order to persuade web visitors to provide personal information, such as login credentials. Due to the evolution of online hacking techniques and a lack of public awareness, internet users are frequently exposed to cyber dangers, such as phishing, spam, trojans, and adware. Phishing has grown in popularity as a means of collecting users' private information, such as login details, credit card information, and social security numbers, via fraudulent websites. Therefore, phishing attacks represent a serious cybersecurity problem that significantly affects commercial websites and the users of the web. Personal information collected in this way can be used to steal money via stolen credit cards, debit cards, bank account fraud, and gaining illegal access to people's social media profiles. Phishing attacks have already resulted in significant losses and may have a negative impact on the victim, not just financially, but also in terms of reputation and national security. In comparison to 2018 and 2019, in 2020, there was a 15% increase in the number of phishing attacks. In addition, Kaspersky Lab's anti-phishing security systems stopped over 482 million phishing threats in 2018, a twofold increase over 2017. Based on the Anti Phishing Working Group's (APWG) report (APWG 2020), the number of phishing attacks is rising continually, with 146,994 phishing websites discovered in the second quarter of 2020. In 2020, the anticipated average cost of a business breach caused by phishing attacks was 2.8 million USD. It is important to utilize anti-phishing methods to avoid such significant losses

2. Ye Cao, Weili Han," Anti-phishing Based on Automated Individual White-List" In phishing and pharming, users could be easily tricked into submitting their username/passwords into fraudulent web sites whose appearances look similar as the genuine ones. The traditional blacklist approach for anti-phishing is partially effective due to its partial list of global phishing sites. In this paper, we present a novel anti-phishing approach named Automated Individual WhiteList (AIWL). AIWL automatically tries to maintain a white-list of user's all familiar Login User Interfaces (LUIs) of web sites. Once a user tries to submit his/her confidential information to an LUI that is not in the white-list, AIWL will alert the user to the possible attack. Next, AIWL can efficiently defend against pharming attacks, because AIWL will alert the user when the legitimate IP is maliciously changed; the legitimate IP addresses, as one of the contents of LUI, are recorded in the white-list and our experiment shows that popular web sites' IP addresses are basically stable. Furthermore, we use Naïve Bayesian classifier to automatically maintain the white-list in AIWL. Finally, we conclude through experiments that AIWL is an efficient automated tool specializing in detecting phishing and pharming Most of the techniques for phishing detection are based on blacklist [30]. In the blacklist approaches, once the user visits a web site that is in the blacklist, he/she will be warned of the potential attack. But maintaining a blacklist requires a great deal of resources for reporting and verification of the suspicious web sites. In addition, phishing sites emerge endlessly, so it is difficult to keep a global blacklist up to date. Contrary to blacklist, white-list approach maintains a list containing all legitimate web sites. But a global white-list approach is likewise hardly used because it is impossible for a white-list to cover all legitimate web sites in the entire cyber world.In this paper, we present a novel approach, named Automated Individual White-List (AIWL). AIWL uses a white list that records all familiar Login User Interfaces (LUIs) of web sites for a user. A familiar LUI of a web site refers to the characteristic information of a legitimate login page on which the user wants to input his/her username/password. Every time a user tries to submit his/her sensitive information into an LUI that is not included in the white-list, the user will be alerted to the possible attack. Here, LUI refers to the user interface where user inputs his/her username/passwords. For instance, a typical LUI is composed of URL address, page feature, DNS-IP mapping. Once the user tries to submit the confidential information into a web site that is in the white-list, LUI information of current web site will be collected and compared with the pre-stored one in the white-list. Any mismatch will also cause warning to the user. To conveniently set up the white-list in AIWL, we use the Naïve Bayesian classifier [8, 9] to identify a successful login process. After a web site has been logged in successfully several times, it is believed to be a familiar one of the user and the LUI information of the web site can be added to the white-list automatically after user's confirmation. The rest of our paper is organized as follows: in section 2, we introduce background and motivation of the paper; section 3 introduces the overall approach of AIWL and discusses some important issues in the approach; section 4 describes the experiments for evaluation; section discusses the advantages of AIWL on the basis of its comparison with other solutions and consider the limitations of AIWL; section 6 introduces the related work; and section 7 summarizes our paper and introduces future work. Phishing attackers use both social engineering and technical subterfuge to steal user's identity data as well as financial account information. By sending "spoofed" e-mails, social-engineering schemes lead users to counterfeit web sites that are designed to trick recipients into divulging financial data such as credit card numbers, account usernames, passwords and social security numbers. In order to persuade the recipients to respond, phishers often hijack brand names of banks, e-retailers and credit card companies. Furthermore, technical subterfuge schemes often plant crimewares, such as Trojan, keylogger spyware, into victims' machines to steal user's credentials. Phishing attack not only leads to great loss to users but also influences the expansion of ecommerce. Rampant phishing attacks would cause the whole e-commerce environment to be dangerous and aggressive. Furthermore, it is difficult for common users to distinguish fraudulent web site from the genuine one. Thus, users would feel hesitant to use e-banking and online shopping services in such an environment.

3. Arathi Krishna V , Anusree A," Phishing Detection using Machine Learning based URL Analysis: A Survey" As we have moved most of our financial, work related and other daily activities to the internet, we are exposed to greater risks in the form of cybercrimes. URL based phishing attacks are one of the most common threats to the internet users. In this type of attack, the attacker exploits the human vulnerability rather than software flaws. It targets both individuals and organizations, induces them to click on URLs that look secure, and steal confidential information or inject malware on our system. Different machine learning algorithms are being used for the detection of phishing URLs, that is, to classify a URL as phishing or legitimate. Researchers are constantly trying to improve the performance of existing models and increase their accuracy. In this work we aim to review various machine learning methods used for this purpose, along with datasets and URL features used to train the machine learning models. The performance of different machine learning algorithms and the methods used to increase their accuracy measures are discussed and analysed. The goal is to create a survey resource for researchers to learn the current developments in the field and contribute in making phishing detection models that yield more accurate results.The year 2020 saw peoples' life being completely dependent on technology due to the global pandemic. Since digitalization became significant in this scenario, cyber criminals went on an internet crime spree. Recent reports and researches point to an increased number of security breaches that costs the victims a huge sum of money or disclosure of confidential data. Phishing is a cybercrime that employs both social engineering and technical subterfuge in order to steal personal identity data or financial account credentials of victims[1]. In phishing, attackers counterfeit trusted websites and misdirect people to these websites, where they are tricked into sharing usernames, passwords, banking or credit card details and other sensitive credentials. These phishing URLs may be sent to the consumers through email, instant message or text message. According to the FBI crime report 2020, phishing was the most common type of cyber attack in 2020 and phishing incidents nearly doubled from 114,702 in 2019 to 241,342 in 2020. The Verizon 2020 Data Breach Investigation Report states that 22% of data breaches in 2020 involved phishing[3]. The number of phishing attacks as observed by the AntiPhishing Work Group (APWG) grew through 2020, doubling over the course of the year. In the 4th quarter of 2020, it was found that phishing attacks against financial institutions were the most prevalent. Phishing attacks against SaaS and Webmail sites were down and attacks against E-commerce sites escalated, while attacks against media companies decreased slightly from 12.6% to 11.8%[1]. In light of the prevailing pandemic situation, there have been many phishing attacks that exploit the global focus on Covid-19. According to WHO, many hackers and cyber scammers are sending fraudulent emails and WhatsApp messages to people, taking advantage of the coronavirus disease[4]. These attacks are coming in the form of fake job offers, fabricated messages from health organizations, covid vaccine themed phishing and brand impersonation. A URL based phishing attack is carried out by sending malicious links, that seems legitimate to the users, and tricking them into clicking on it. In phishing detection, an incoming URL is identified as phishing or not by analysing the different features of the URL and is classified accordingly. Different machine learning algorithms are trained on various datasets of URL features to classify a given URL as phishing or legitimate

4. Mohsen Sharifi, Seyed Hossein Siadati," A Phishing Sites Blacklist Generator" Phishing is an increasing web attack both in volume and techniques sophistication. Blacklists are used to resist this type of attack, but fail to make their lists upto-date. This paper proposes a new technique and architecture for a blacklist generator that maintains an up-to-date blacklist of phishing sites. When a page claims that it belongs to a given company, the company's name is searched in a powerful search engine like Google. The domain of the page is then compared with the domain of each of the Google's top10 searched results. If a matching domain is found, the page is considered as a legitimate page, and otherwise as a phishing site. Preliminary evaluation of our technique has shown an accuracy of 91% in detecting legitimate pages and 100% in detecting phishing sites.Phishing attack is a type of identity theft that aims to deceit users into revealing their personal information which could be exploited for illegal financial purposes. A phishing attack begins with an email that claims it is from a legal company like eBay. The content of email motivates the user to click on a malicious link in the email. The link connects the user to an illegitimate page that mimics the outward appearance of original site. The phishing page then requests user's personal information, like online banking passwords and credit card information. The number of phishing attacks has grown rapidly. According to trend reports by Anti-Phishing Working Group (APWG) [1], the number of unique phishing sites has been reported 37,444 sites in October 2006, increased from 4,367 sites in October 2005. Other statistics show the increase in the volume of the Phishing attack and their techniques are becoming much more advanced. A number of techniques have been studied and practiced against phishing and a large number of them use phishing blacklists to battle against phishing. Blacklists of phishing sites are valuable sources that are in use by anti-phishing toolbars to notify users and deny their access to phishing sites, web and email filters to filter spam and phishing emails, and phishing termination communities to terminate the phishing sites. Blacklist indicates whether a URL is good or bad. A bad URL means that it is known to be used by attackers to steel users' information. The blacklist publisher assigns the "goodness" (the URLs that are not in the list) and the "badness" (the URLs that are in the list) to all internet URLs. Many browsers now check blacklist databases to address phishing problem and notify users when they browse phishing pages. Internet Explorer 7, Netscape Browser 8.1[4], Google Safe Browsing (a feature of the Google Toolbar for Firefox) are important browsers which use blacklists to protect users when they navigating phishing sites.

SYSTEM ANALYSIS

EXISTING SYSTEM Existing solutions detect mimicked phishing pages by either text-based features or visual similarities of webpages and it can be easily bypassed and proposed a technique to identify the real domain name of a visiting webpage based on signatures created for web sites, site signatures, including distinctive texts and images, can be generated by analysing common parts from pages of a website. The authors claimed that the method achieves high accuracy and low error rates. Aaron Blum et. Eexplored the possibility of utilizing confidence weighted classification combined with content-based phishing URL detection to produce a dynamic and extensible system for detection of present and emerging types of phishing domains, and authors further claims the system can detect emerging threats and can provide an increased protection against zero-hour threats, unlike traditional blacklisting techniques which function reactively.Exist in phishing attacks in reality and can detect zero-hour phishing attack. But the characteristics are not guaranteed to always exist in such attacks and false positive rate in detection is very high. This tag is used to add another web page into existing main webpage. Phishers can make use of the “iframe” tag and make it invisible i.e. Without frame borders. Since border of inserted webpage is invisible, user seems that the inserted web page is also the part of the main web page and can enter sensitive information.

PROPOSED SYSTEM This is the most common type of phishing attack wherein a cybercriminal impersonates a known popular entity, domain or organization and attempt to steal sensitive private information from the victim such as login, password, bank account detail, credit card detail, etc. This type of attack lacks sophistication as it does not have personalization and customization for the individuals. For an example, emails containing Phishing URL is disseminated in bulk to large users as a volume of mail is very high the cybercriminal would expect that many users will open the emails and visit the malicious URLs or open the infected attachments. The idea behind this type of phishing is deception and impersonation. This type of email mostly creates panic and urgency for the victims to divulge sensitive information. The email subject will be such that it might create urgency such as "Your account has been hacked, change your password immediately!", "Your bill is overdue-pay immediately of pay fine!" or other similar messages, once a user open such messages or visit the URLs the damage is done. The victim of the scams incurs a monetary loss, loss of private information and loss of reputation. Hence, it is imperative to find a solution that could mitigate such security threats in a timely manner. Traditionally, the detection of phishing websites is done using blacklists. There are many popular websites which host a list of blacklisted websites, e. g. PhisTank. The blacklisting technique lack in two aspects, blacklists might not be exhaustive and do not detect a newly generated phishing website. In recent times machine learning techniques have been used in the classification and detection of phishing websites.

OVERVIEW OF THEPROJECT In recent years, with the increasing use of mobile devices, there is a growing trend to move almost all real-world operations to the cyberworld. Although this makes easy our daily lives, it also brings many security breaches due to the anonymous structure of the Internet. Used antivirus programs and firewall systems can prevent most of the attacks. However, experienced attackers target on the weakness of the computer users by trying to phish them with bogus webpages. These pages imitate some popular banking, social media, e-commerce, etc. sites to steal some sensitive information such as, user-ids, passwords, bank account, credit card numbers, etc. Phishing detection is a challenging problem, and many different solutions are proposed in the market as a blacklist, rule-based detection, anomaly-based detection, etc. In the literature, it is seen that current works tend on the use of machine learning-based anomaly detection due to its dynamic structure, especially for catching the “zero-day” attacks. In this paper, we proposed a machine learning-based phishing detection system by using eight different algorithms to analyze the URLs, and three different datasets to compare the results with other works. The experimental results depict that the proposed models have an outstanding performance with a success rate. Phishing is a form type of a cybersecurity attack where an attacker gains control on sensitive website user accounts by learning sensitive information such as login credentials, credit card information by sending a malicious URL in email or masquerading as a reputable person in email or through other communication channels. The victim receives a message from known contacts, persons, entities or organizations and looks very much genuine in its appeal. The received message might contain malicious links, software that might target the user computer or the malicious link might direct the user to some forged website which is similar in look and feel of a popular website, further victim might be tricked to divulge his personal information e.g. credit card information, login and password details and other sensitive information like account id details etc. Phishing is the most popular type of cybersecurity attack and very common among the attackers. Phishing attacks are generally easy as most of the victims are not well aware of the intricacies about the web applications and computer networks and its technologies and are easy prey for getting tricked or spoofed. It is very easy to phishing unsuspecting users using forged websites and luring them for clicking the websites for some prize and offers than targeting the computer defense system. The malicious website is designed in such a way that it has a similar look and feel and it appears very genuine in its appearance as it contains the organization's logos and other copyrighted contents. As many users unwittingly clicking the phishing websites URLs and this results in huge financial and loss of reputation to the person and to the concerned organization In our daily life, we carry out most of our work on digital platforms. Using a computer and the internet in many areas facilitates our business and private life. It allows us to complete our transaction and operations quickly in areas such as trade, health, education, communication, banking, aviation, research, engineering, entertainment, and public services. The users who need to access a local network have been able to easily connect to the Internet anywhere and anytime with the development of mobile and wireless technologies. Although this situation provides great convenience, it has revealed serious deficits in terms of information security. Thus, the need for users in cyberspace to take measures against possible cyber-attacks has emerged. Attacks can be carried out by people such as cybercriminals, pirates, or non-malicious (white-capped) attackers and hacktivists. The aim is to reach the computer or the information it contains or to capture personal information in different ways. The attacks, as internet worms (Morris Worm), started in 1988, and they have been carried out until today. These attacks are mainly targeted in the following areas: fraud, forgery, force, shakedown, hacking, service blocking, malware applications, illegal digital contents and social engineering Reaching with a wide range of target users, attackers aim to get a lot of information and/or money. According to Kaspersky's data, the average cost of an attack in 2019 (depending on the size of the attack) is between $ 108K and $ 1.4 billion. In addition, the money spent on global security products and services is around $ 124 billion . Among these attacks, the most widespread and also critical one is “phishing attacks”. In this type of attack, cybercriminals especially use an email or other social networking communication channels. Attackers reach the victim users by giving the impression that the post was sent from a reliable source, such as a bank, e-commerce site, or similar. Thus, they try to access sensitive information of them. Attackers then access their victims’ accounts by using this information. Thus, it causes pecuniary loss and intangible damages.

CONCLUSION In this project, we have explored how well to classify phishing URLs from the given set of URLs containing benign and phishing URLs. We have also discussed the randomization of the dataset, feature engineering, feature extraction using lexical analysis host-based features and statistical analysis. We have also used different classifiers for the comparative study and found that the findings are almost consistent across the different classifiers. We also observed dataset randomization yielded a great optimization and the accuracy of the classifier improved significantly. We have adopted a simple approach to extract the features from the URLs using simple regular expressions. There could be more features that can be experimented and that might lead to improving further the accuracy of the system. The dataset used in this paper contains the URLs list which may be a little old, hence regular continuous training along with a new dataset would enhance the model accuracy and performance significantly. In our experiment we have not used the content based features as the main problem with the content-based strategy for detecting phishing URLs is the non-availability of phishing web-sites and the life span of the phishing website is small, and it is difficult to train an ML classifier based on its content-based features. In the future, we would like to incorporate a rule-based prediction based on the content analysis of a URL. Hence, the combination of classification based lexical analyzer along with a rule-based URL content analyzer for phishing URL detection would provide a comprehensive solution

REFERENCES

[1]. Khonji, Mahmoud, Youssef Iraqi, and Andrew Jones. "Phishing detection: a literature survey." Communications Surveys & Tutorials, IEEE 15.4 (2013): 2091-2121.

[2]. Anti Phishing Working Group. (2015. March.) APWG Phishing Activity Trend Report 2nd Quarter 2010. [Online]. Available: http://docs.apwg.org/reports/apwg_trends_repo rt_q2_2 014.pdf

[3]. Anti Phishing Working Group. (2015. March.) APWG Phishing Activity Trend Report 2nd Quarter 2014. [Online]. Available: http://docs.apwg.org/reports/apwg_report_q2_2 010.pdf

[4]. Huang, Huajun, Junshan Tan, and Lingxi Liu. "Countermeasure techniques for deceptive phishing attack." New Trends in Information and Service Science, 2009. NISS'09. International Conference on. IEEE, 2009.

[5]. Ma, Justin, et al. "Beyond blacklists: learning to detect malicious web sites from suspicious URLs." Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009.

[6]. Nguyen, Luong Anh Tuan, et al. "A novel approach for phishing detection using URLbased heuristic." Computing, Management and Telecommunications (ComManTel), 2014 International Conference on. IEEE, 2014.

[7]. Wikipedia. (2015. March) Uniform Resource Loactor. Avaliable: http://en.wikipedia.org/wiki/Uniform_resource _locator

[8]. Kausar, Firdous, et al. "Hybrid Client Side Phishing Websites Detection Approach." International Journal of Advanced Computer Science and Applications (IJACSA) 5.7 (2014).

[9]. Sunil, A. Naga Venkata, and Anjali Sardana. "A pagerank based detection technique for phishing web sites." Computers & Informatics (ISCI), 2012 IEEE Symposium on. IEEE, 2012.

[10]. Mohammad, Rami M., Fadi Thabtah, and Lee McCluskey. "Intelligent rule-based phishing websites classification." Information Security, IET 8.3 (2014): 153-160.

[11]. Singh, C., & `Meenu., "Phishing Website Detection Based on Machine

Learning: A Survey", IEEE 6th International Conference on Advanced Computing & Communication Systems, Gorakhpur, India, 2020, 978- 1-7281-5197-7.

[12]. Aydin, M., Butun, I., Bicakci, K., & Baykal, N., "Using Attribute-based Feature Selection Approaches and Machine Learning Algorithms for Detecting Fraudulent Website URLs",IEEE 10th Annual Computing and Communication Workshop and Conference, Ankara, Turkey,Goteborg, Sweden, Guzelyurt, Cyprus, 30-May- 2020, 978-1-7281-3783-4.

[13]. A, A. A., & K, P. “Towards the Detection of Phishing Attacks”, IEEE 4th International Conference on Trends in Electronics and Informatics, Coimbatore,India, July 27-2020, 978-1-7281-5518-0

[14]. Arun Kulkarni & Leonard L. Brown, ” Phishing Websites Detection using Machine Learning ”, International Journal of Advanced Computer Science and Applications , Tyler, TX, 2019.

[15]. El Aassal, A., Baki, S., Das, A., & Verma, R. M.,"An In-Depth Benchmarking and Evaluationof Phishing Detection Research for Security Needs", IEEE Access,Houston, U.S., 5- Feb-2020, 2969780