DETECTION OF PHISHING WEBSITES BY USING MACHINE LEARNING-BASED URL ANALYSIS

ACKNOWLEDGEMENT

I am very grateful to Dr.N.SUDHA, M.Tech.,PhD., Professor and Principal, CMS College of Engineering and Technology, Coimbatore, for providing us with an environment to complete our project successfully.I am deeply indebted to r.G.CHITHRA GANAPATHI, ME.,PhD., Associate Professor and Head, Department of Computer Science and Engineering, CMS College of Engineering and Technology,Coimbatore,whomodeled us both technically and morally for achieving greater success in life. I express my sincere thanks to P.RESHMA.,ME., Assistant Professor ,Department of Computer Science and Engineering, CMS College of Engineering and Technology, Coimbatore, for her constant encouragement and support throughout the course of the project period. I am also thankful to all other Faculties, Staff Members, Parents, Friends and almighty for their valuable assistance in carrying out this project work.

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.

CHAPTER 1 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.

CHAPTER 2 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 White-List (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 e-commerce. 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.

Due to the wide use of blacklists of phishing sites against phishing, it is very important to introduce techniques that generate the updated blacklists of phishing sites. The problem of the blacklist is that it is hard to keep the list up-to-date since it is easy to register new domains in the Internet. In this paper we propose a technique to detect deceptive phishing pages, as well as our proposed architecture for a blacklist of phishing sites generator. The rest of paper is organized as follows. Section 2 discusses related works. Section 3 presents our proposed algorithm and the architecture of our blacklist generator. The evaluation of the

approach is given in section 4.
Our proposed technique tries to generate an updated blacklist of phishing sites. Each web page belongs to a web site and most of them show this relation using the site's logo. Phishing pages also use legitimate site's logo to make their pages credible and claim that they belong to that site. Thus we can find which site a page claims to belong, using its logo. On the other hand, the domain of a legal site can be found by searching its name in a search engine like Google.
Our technique is based on these two properties of the web pages and search engines to detect phishing pages. Figure 1 demonstrates our algorithm which gets a URL as input and returns True if the page is phishing and False if the page is a legitimate one.

5. Shwetha. Kavitha," Detection Of Phishing Websites Using Machine Learning"
Phishing is a social manipulation assault aimed at leveraging the vulnerability found in the program at the end of the user. For example, a program may be technically secure enough for password theft, but an unrecognized user can leak his / her password when an attacker sends a request for a false password update via a fake website. To resolve this problem, a layer of security must be added for use. As of late, there have been a few examinations that attempted to tackle the phishing issue. A few analysts utilized the URL furthermore, contrasted it and, existing boycotts that contain arrangements of vindictive sites, which they have been making, and others have utilized the URL in a contrary way, to be specifically contrasting the URL and a whitelist of real sites. The latter approach uses heuristics, which is used Database of signatures for any known attacks that match the Signature of the heuristic template to determine whether it's a phishing This is the platform. Also besides tracking traffic on Alexa 's website is another way in which researchers have been applied to detect websites for phishing.
Phishing is a social engineering attack that aims at exploiting the weakness found in system processes as caused by system users. For example, a system can be technically secure enough against password theft, however unaware end users may leak their passwords if an attacker asked them to update their passwords via a given Hypertext Transfer Protocol (HTTP) link, which ultimately threatens the overall security of the system. Moreover, technical vulnerabilities (e.g. Domain Name System (DNS) cache poisoning) can be used by attackers to construct far more persuading socially-engineered messages (i.e. use of legitimate, but spoofed, domain names can be far more persuading than using different domain names). This makes phishing attacks a layered problem, and an effective mitigation would require addressing issues at the technical and human layers. Since phishing attacks aim at exploiting weaknesses found in humans (i.e. system end-users), it is difficult to mitigate them. For example, as evaluated in [1], end-users failed to detect 29% of phishing attacks even when trained with the best performing user awareness program. On the other hand, software phishing detection techniques are evaluated against bulk phishing attacks, which makes their performance practically unknown with regards to targeted forms of phishing attacks. These limitations in phishing mitigation techniques have practically resulted in security breaches against several organizations including leading information security providers.
The definition by Colin Whittaker et. al. aims to be broader than PhishTank's definition in a sense that attackers goals are no longer restricted to stealing personal information from victims. On the other hand, the definition still restricts phishing attacks to ones that act on behalf of third parties, which is not always true. For example phishing attacks may communicate socially engineered messages to lure victims into installing MITB malware by attracting the victims to websites that are supposed to deliver safe content (e.g. video streaming). Once the malware (or crimeware as often named by Anti-Phishing Working Group (APWG)2) is installed, it may log the victim's keystrokes to steal their passwords. Note that the attacker in this scenario did not claim the identity of any third party in the phishing process, but merely communicated messages with links (or attachments) to lure victims to view videos or multimedia content.In order to address the limitations of the previous definitions above, we consider phishing attacks as semantic attacks which use electronic communication channels (such as EMails, HTTP, SMS, VoIP, etc. . .) to communicate socially engineered messages to persuade victims to perform certain actions (without restricting the actions) for an attacker's benefit (without restricting the benefits). See Definition 1.
Definition 1: Phishing is a type of computer attack that communicates socially engineered messages to humans via electronic communication channels in order to persuade them to perform certain actions for the attacker's benefit. For example, the performed action (which the attacker persuades the victim to perform it) for a PayPal user is submitting his/her login credentials to a fake website that looks similar to PayPal. As a perquisite, this also implies that the attack should create a need for the end-user to perform such action, such as informing him that his/her account would be suspended unless he logs in to update certain pieces of information.

6. Pratik Patil , Devale," A Literature Survey of Phishing Attack Technique"
It is a crime to practice phishing by employing technical tricks and social engineering to exploit the innocence of unaware users. This methodology usually covers up a trustworthy entity so as to influence a consumer to execute an action if asked by the imitated entity. Most of the times, phishing attacks are being noticed by the practiced users but security is a main motive for the basic users as they are not aware of such circumstances. However, some methodologies are limited to look after the phishing attacks only and the delay in detection is mandatory. In this paper we emphasize the various techniques used for the detection of phishing attacks. We have also discovered various techniques for detection and prevention of phishing. Apart from that, we have introduced a new model for detection and prevention of phishing attacks.
The purpose or goal behind phishing is data, money or personal information stealing through the fake website. The best strategy for avoiding the contact with the phishing web site is to detect real time malicious URL. Phishing websites can be determined on the basis of their domains. They usually are related to URL which needs to be registered (low-level domain and upper-level domain, path, query).

Recently acquired status of intra-URL relationship is used to evaluate it using distinctive properties extracted from words that compose a URL based on query data from various search engines such as Google and Yahoo.
These properties are further led to the machine-learningbased classification for the identification of phishing URLs from a real dataset. This paper focus on real time URL phishing against phishing content by using phish-STORM. For this a few relationship between the register domain rest of the URL are consider also intra URL relentless is consider which help to dusting wish between phishing or non phishing URL. For detecting a phishing website certain typical blacklisted urls are used, but this technique is unproductive as the duration of phishing websites is very short. Phishing is the name of avenue. It can be defined as the manner of deception of an organization's customer to communicate with their confidential information in an unacceptable behaviour. It can also be defined as intentionally using harsh weapons such as Spasm to automatically target the victims and targeting their private information. As many of the failures being occurred in the SMTP are exploiting vectors for the phishing websites, there is a greater availability of communication for malicious message deliveries.
Along with the various criminal enterprises, if there is enough amount of money generated through the mode of phishing, hunting of various other systems of message delivery can be done, even though the errors are closed eventually in SMTP. Along with the ever increasing dishonesty through phishing scams, organizations are getting more attention from their customers regarding the security of their personal information.AntiPhish is used to avoid users from using fraudulent web sites which in turn may lead to phishing attack. Here, AntiPhish traces the sensitive information to be filled by the user and alerts the user whenever he/she is attempting to share his/her information to a untrusted web site. The much effective elucidation for this is cultivating the users to approach only for trusted websites.
However, this approach is unrealistic. Anyhow, the user may get tricked. Hence, it becomes mandatory for the associates to present such explanations to overcome the problem of phishing. Widely accepted alternatives are based on the creepy websites for the identification of “clones” and maintenance of records of phishing websites which are in hit list.
An alternative for detecting these attacks is a relevant process of reliability of machine on a trait intended for the reflection of the besieged deception of user by means of electronic communication. This approach can be used in the detection of phishing websites, or the text messages sent through emails that are used for trapping the victims. Approximately, 800 phishing mails and 7,000 nonphishing mails are traced till date and are detected accurately over 95% of them along with the categorization on the basis of 0.09% of the genuine emails. We can just wrap up with the methods for identifying the deception, along with the progressing nature of attacks. Very complex and dynamic to be identified and classified. Due to the involvement of various ambiguities in the detection, certain crucial data mining techniques may prove an effective means in keeping the e-commerce websites safe since it deals with considering various quality factors rather than exact values. In this paper, an effective approach to overcome the “fuzziness” in the e-banking phishing website assessment is used an intelligent resilient and effective model for detecting e-banking phishing websites is put forth. The applied model is based on fuzzy logics along with data mining algorithms to consider various effective factors of the e-banking phishing website.

CHAPTER 3 SYSTEM ANALYSIS

3.1 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.

3.2 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.

3.3 FEASIBILITYSTUDY

A feasibility study is concerned to select the best system that meets performance requirements. These entities are an identification description, an evaluation of candidate systems and the selection of the best for the job.

- Economic Feasibility
- Technical Feasibility
- Behavioral Feasibility

3.3.1 Economic Feasibility

Economic analysis is the most frequently used method for evaluating the effectiveness of the candidate system. More commonly known as cost/benefit analysis, the procedure is to determine the benefits and savings that benefits outweigh costs, and then the decision is made to design and implement the system. Otherwise, further justification or alterations in the proposed system will have to be made if it is to have an enhancement toapprove.

3.3.2 Technical Feasibility

Technical analysis centre on the existing computer system (Hardware, Software, etc) and to what extend it can support the proposed addition. This involves financial considerations to accommodate technical enhancement. If the budget is a serious constraint then the project is judged not feasible.

3.3.3 Behavioral Feasibility

An estimate should be made of how strong a reaction the user staff is likely to have toward the development of a computerized system. It is common knowledge the computer installations have something to do understandable that the introduction of a candidate system requires special effort to educate, sell and train the stay on now ways of considering business.

3.3.2 Advantage of Making Feasibility Study

There are many advantage of making feasibility study some of which are summarized below

- This study being made as the initial step of software development life cycle has all the analysis part in it which helps in analyzing the system requirements completely.
- Helps in identifying the risk factors involved in developing and deploying the system.
- The feasibility study helps in planning for riskanalysis.
- Feasibility study helps in making cost/benefit analysis which helps the organization and system to run efficiently.
- Feasibility study helps in making plans for training developers for implementing the system.

CHAPTER 4 SYSTEM SPECIFICATION

All computer software needs certain hardware components or other software resources to be present on a computer to be used efficiently. These prerequisites are known as (computer) system requirements and are often used as a guideline as opposed to an absolute rule.

4.1 HARDWAREREQUIREMENTS

- Processor : i3
- RAM : 4GB OR MORE
- Hard Disk Drive : 500 GB

4.2 SOFTWARE REQUIREMENTS

- Development Platform : Windows 10
- Front End : Python,Pandas,Numpy,Matplot,Sklearn
- Back-End : Dataset

CHAPTER 5 SYSTEM DESCRIPTION

5.1 FRONTEND

PYTHON

Python is a general purpose, dynamic, high level, and interpreted programming language. It supports Object Oriented programming approach to develop applications. It is simple and easy to learn and provides lots of high-level data structures. Python is easy to learn yet powerful and versatile scripting language, which makes it attractive for Application Development. Python's syntax and dynamic typing with its interpreted nature make it an ideal language for scripting and rapid application development.Python supports multiple programming pattern, including object-oriented, imperative, and functional or procedural programming styles.Python is not intended to work in a particular area, such as web programming. That is why it is known as multipurpose programming language because it can be used with web, enterprise, 3D CAD, etc.We don't need to use data types to declare variable because it is dynamically typed so we can write a=10 to assign an integer value in an integer variable.Python makes the development and debugging fast because there is no compilation step included in Python development, and edit-test-debug cycle is very fast. Python is known for its general-purpose nature that makes it applicable in almost every domain of software development. Python makes its presence in every emerging field. It is the fastest-growing programming language and can develop any application.

Pandas

Pandas is an open-source Python Library providing high-performance data manipulation and analysis tool using its powerful data structures. The name Pandas is derived from the word Panel Data – an Econometrics from Multidimensional data.In 2008, developer Wes McKinney started developing pandas when in need of high performance, flexible tool for analysis of data. Prior to Pandas, Python was majorly used for data munging and preparation. It had very little contribution towards data analysis. Pandas solved this problem. Using Pandas, we can accomplish five typical steps in the processing and analysis of data, regardless of the origin of data — load, prepare, manipulate, model, and analyze.

Python with Pandas is used in a wide range of fields including academic and commercial domains including finance, economics, Statistics, analytics, etc.

- Fast and efficient DataFrame object with default and customized indexing.
- Tools for loading data into in-memory data objects from different file formats.
- Data alignment and integrated handling of missing data.
- Reshaping and pivoting of date sets.
- Label-based slicing, indexing and subsetting of large data sets.
- Columns from a data structure can be deleted or inserted.
- Group by data for aggregation and transformations.
- High performance merging and joining of data.
- Time Series functionality.

Numpy

NumPy is a Python package. It stands for 'Numerical Python'. It is a library consisting of multidimensional array objects and a collection of routines for processing of array.

Numeric, the ancestor of NumPy, was developed by Jim Hugunin. Another package Numarray was also developed, having some additional functionalities. In 2005, Travis Oliphant created NumPy package by incorporating the features of Numarray into Numeric package. There are many contributors to this open source project.Operations using NumPy

Using NumPy, a developer can perform the following operations −

- Mathematical and logical operations on arrays.
- Fourier transforms and routines for shape manipulation.
- Operations related to linear algebra. NumPy has in-built functions for linear algebra and random number generation.

NumPy – A Replacement for MatLab

NumPy is often used along with packages like **SciPy** (Scientific Python) and **Mat−plotlib** (plotting library). This combination is widely used as a replacement for MatLab, a popular platform for technical computing. However, Python alternative to MatLab is now seen as a more modern and complete programming language.

Matplot

Matplotlib is one of the most popular Python packages used for data visualization. It is a cross-platform library for making 2D plots from data in arrays.
Matplotlib is written in Python and makes use of NumPy, the numerical mathematics extension of Python.
It provides an object-oriented API that helps in embedding plots in applications using Python GUI toolkits such as PyQt, WxPythonotTkinter.
It can be used in Python and IPython shells, Jupyter notebook and web application servers also.
Matplotlib has a procedural interface named the Pylab, which is designed to resemble MATLAB, a proprietary programming language developed by MathWorks.
Matplotlib along with NumPy can be considered as the open source equivalent of MATLAB.
Matplotlib was originally written by John D. Hunter in 2003. The current stable version is 2.2.0 released in January 2018.
Matplotlib and its dependency packages are available in the form of wheel packages on the standard Python package repositories and can be installed on Windows, Linux as well as MacOS systems using the pip package manager.

Sklearn

What is Scikit-Learn (Sklearn)
Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib.
Origin of Scikit-Learn It was originally called scikits.learn and was initially developed by David Cournapeau as a Google summer of code project in 2007. Later, in 2010, Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, and Vincent Michel, from FIRCA (French Institute for Research in Computer Science and Automation), took this project at another level and made the first public release (v0.1 beta) on 1st Feb. 2010.
Dataset Loading
A collection of data is called dataset. It is having the following two components −
Features − The variables of data are called its features. They are also known as predictors, inputs or attributes.
Feature matrix − It is the collection of features, in case there are more than one.
Feature Names − It is the list of all the names of the features.
Response − It is the output variable that basically depends upon the feature variables. They are also known as target, label or output.
Response Vector − It is used to represent response column. Generally, we have just one response column.
Target Names − It represent the possible values taken by a response vector.
Scikit-learn have few example datasets like iris and digits for classification and the Boston house prices for regression.

DATASET

The key to success in the field of machine learning or to become a great data scientist is to practice with different types of datasets. But discovering a suitable dataset for each kind of machine learning project is a difficult task. So, in this topic, we will provide the detail of the sources from where you can easily get the dataset according to your project.

Before knowing the sources of the machine learning dataset, let's discuss datasets.

What is a dataset?

A dataset is a collection of data in which data is arranged in some order. A dataset can contain any data from a series of an array to a database table. Below table shows an example of the dataset:

A tabular dataset can be understood as a database table or matrix, where each column corresponds to a **particular variable,** and each row corresponds to the **fields of the dataset.** The most supported file type for a tabular dataset is **"Comma Separated File,"** or **CSV.** But to store a "tree-like data," we can use the JSON file more efficiently.

Types of data in datasets

- **Numerical data:**Such as house price, temperature, etc.

- **Categorical data:**Such as Yes/No, True/False, Blue/green, etc.
- **Ordinal data:**These data are similar to categorical data but can be measured on the basis of comparison.

Note: A real-world dataset is of huge size, which is difficult to manage and process at the initial level. Therefore, to practice machine learning algorithms, we can use any dummy dataset.

Need of Dataset

To work with machine learning projects, we need a huge amount of data, because, without the data, one cannot train ML/AI models. Collecting and preparing the dataset is one of the most crucial parts while creating an ML/AI project.

The technology applied behind any ML projects cannot work properly if the dataset is not well prepared and pre-processed.

During the development of the ML project, the developers completely rely on the datasets. In building ML applications, datasets are divided into two parts training and testing:

1. Kaggle Datasets

Kaggle is one of the best sources for providing datasets for Data Scientists and Machine Learners. It allows users to find, download, and publish datasets in an easy way. It also provides the opportunity to work with other machine learning engineers and solve difficult Data Science related tasks.

CHAPTER 6 PROJECT DESCRIPTION

6.1 PROBLEMDEFINITION

Phishing attack is a simplest way to obtain sensitive information from innocent users. Aim of the phishers is to acquire critical information like username, password and bank account details. Cyber security persons are now looking for trustworthy and steady detection techniques for phishing websites detection. This paper deals with machine learning technology for detection of phishing URLs by extracting and analyzing various features of legitimate and phishing URLs. Decision Tree, random forest and Support vector machine algorithms are used to detect phishing websites. Aim of the paper is to detect phishing URLs as well as narrow down to best machine learning algorithm by comparing accuracy rate, false positive and false negative rate of each algorithm. Nowadays Phishing becomes a main area of concern for security researchers because it is not difficult to create the fake website which looks so close to legitimate website. Experts can identify fake websites but not all the users can identify the fake website and such users become the victim of phishing attack. Main aim of the attacker is to steal banks account credentials. In United States businesses, there is a loss of US$2billion per year because their clients become victim to phishing. In 3rd Microsoft Computing Safer Index Report released in February 2014, it was estimated that the annual worldwide impact of phishing could be as high as $5 billion . Phishing attacks are becoming successful because lack of user awareness. Since phishing attack exploits the weaknesses found in users, it is very difficult to mitigate them but it is very important to enhance phishing detection techniques.

.

6.2 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.

6.3 MODULE

- Presence of IP address in URL
- Presence of @ symbol in URL
- Number of dots in Hostname
- Prefix or Suffix separated by (-) to domain
- URL redirection
- Information submission to Email
- URL Shortening Services “TinyURL”
- Length of Host name
- Presence of sensitive words in URL

6.3.1 Presence of IP address in URL: If IP address present in URL then the feature is set to 1 else set to 0. Most of the benign sites do not use IP address as an URL to download a webpage. Use of IP address in URL indicates that attacker is trying to steal sensitive information.

6.3.2 Presence of @ symbol in URL: If @ symbol present in URL then the feature is set to 1 else set to 0. Phishers add special symbol @ in the URL leads the browser to ignore everything preceding the "@" symbol and the real address often follows the "@" symbol [4].

6.3.3 Number of dots in Hostname: Phishing URLs have many dots in URL. For example http://shop.fun.amazon.phishing.com, in this URL phishing.com is an actual domain name, whereas use of "amazon" word is to trick users to click on it. Average number of dots in benign URLs is 3. If the number of dots in URLs is more than 3 then the feature is set to 1 else to 0.

6.3.4 Prefix or Suffix separated by (-) to domain: If domain name separated by dash (-) symbol then feature is set to 1 else to 0. The dash symbol is rarely used in legitimate URLs. Phishers add dash symbol (-) to the domain name so that users feel that they are dealing with a legitimate webpage. For example Actual site is http://www.onlineamazon.com but phisher can create another fake website like http://www.online-amazon.com to confuse the innocent users.

6.3.5 URL redirection: If "//" present in URL path then feature is set to 1 else to 0. The existence of "//" within the URL path means that the user will be redirected to another website

6.3.6 Information submission to Email: Phisher might use "mail()" or "mailto:" functions to redirect the user's information to his personal email[4]. If such functions are present in the URL then feature is set to 1 else to 0.

6.3.7 URL Shortening Services "TinyURL": TinyURL service allows phisher to hide long phishing URL by making it short. The goal is to redirect user to phishing websites. If the URL is crafted using shortening services (like bit.ly) then feature is set to 1 else 0

6.3.8 Length of Host name: Average length of the benign URLs is found to be a 25, If URL's length is greater than 25 then the feature is set to 1 else to 0

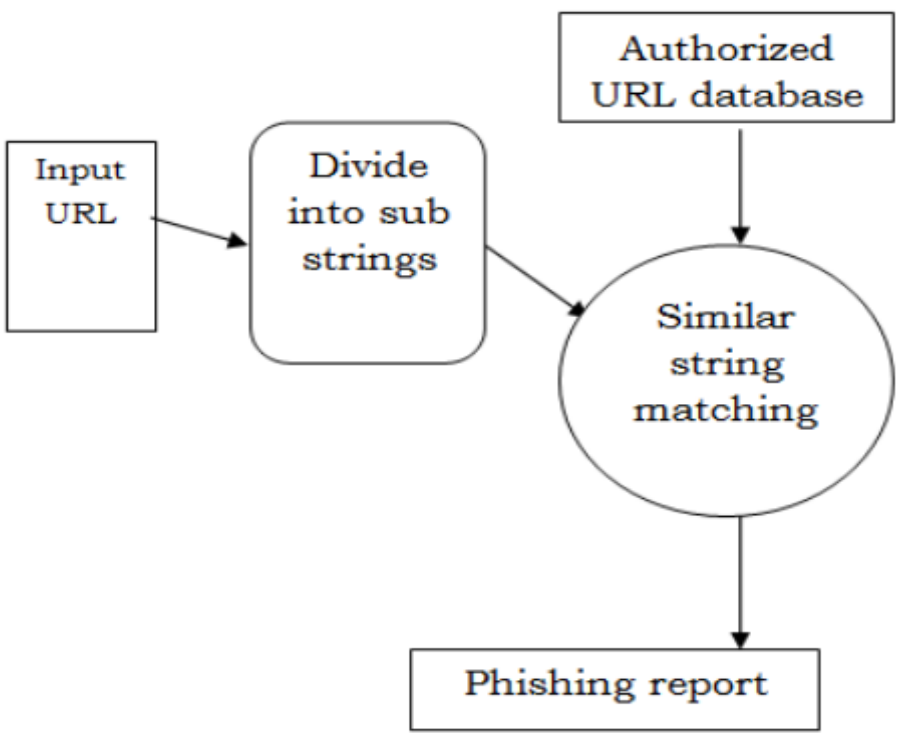
6.3.9 Presence of sensitive words in URL: Phishing sites use sensitive words in its URL so that users feel that they are dealing with a legitimate webpage. Below are the words that found in many phishing URLs :- 'confirm', 'account', 'banking', 'secure', 'ebyisapi', 'webscr', 'signin', 'mail', 'install', 'toolbar', 'backup', 'paypal', 'password', 'username', etc;

DATA FLOW DIAGRAM

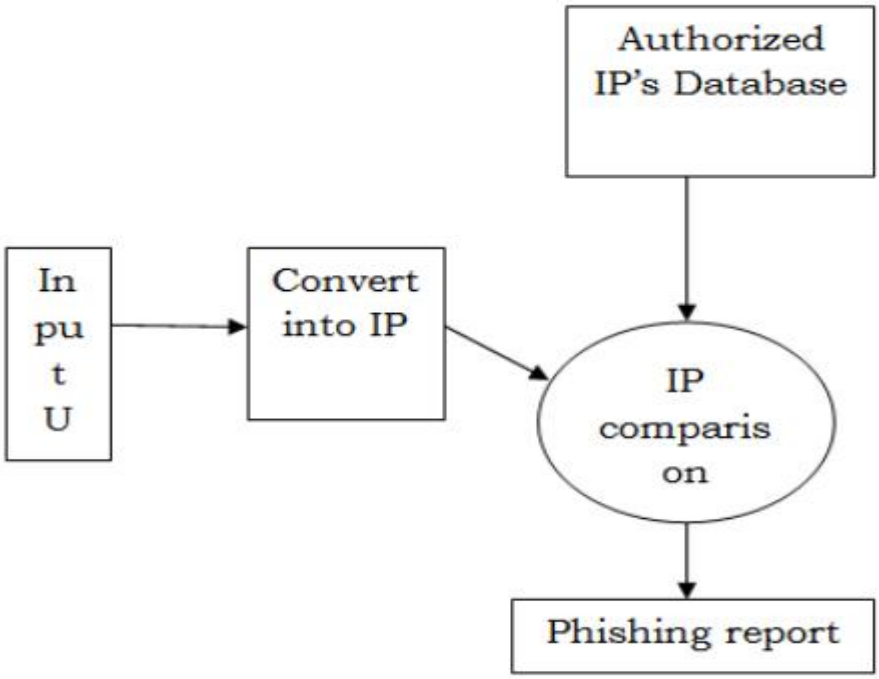
The database may be defined as an organized collection of related information. The organized information serves as a base from which further recognizing can be retrieved desired information or processing the data. The most important aspect of building an application system is the design of tables.

The data flow diagram is used for classifying system requirements to major transformation that will become programs in system design. This is starting point of the design phase that functionally decomposes the required specifications down to the lower level of details. It consists of a series of bubbles joined together by lines.

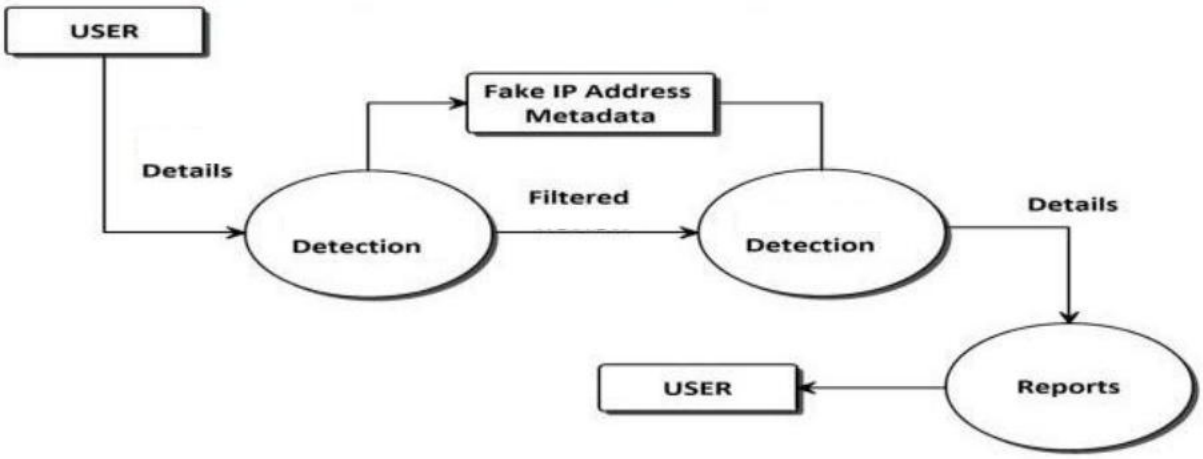
Level 0 Data Flow Daiagram



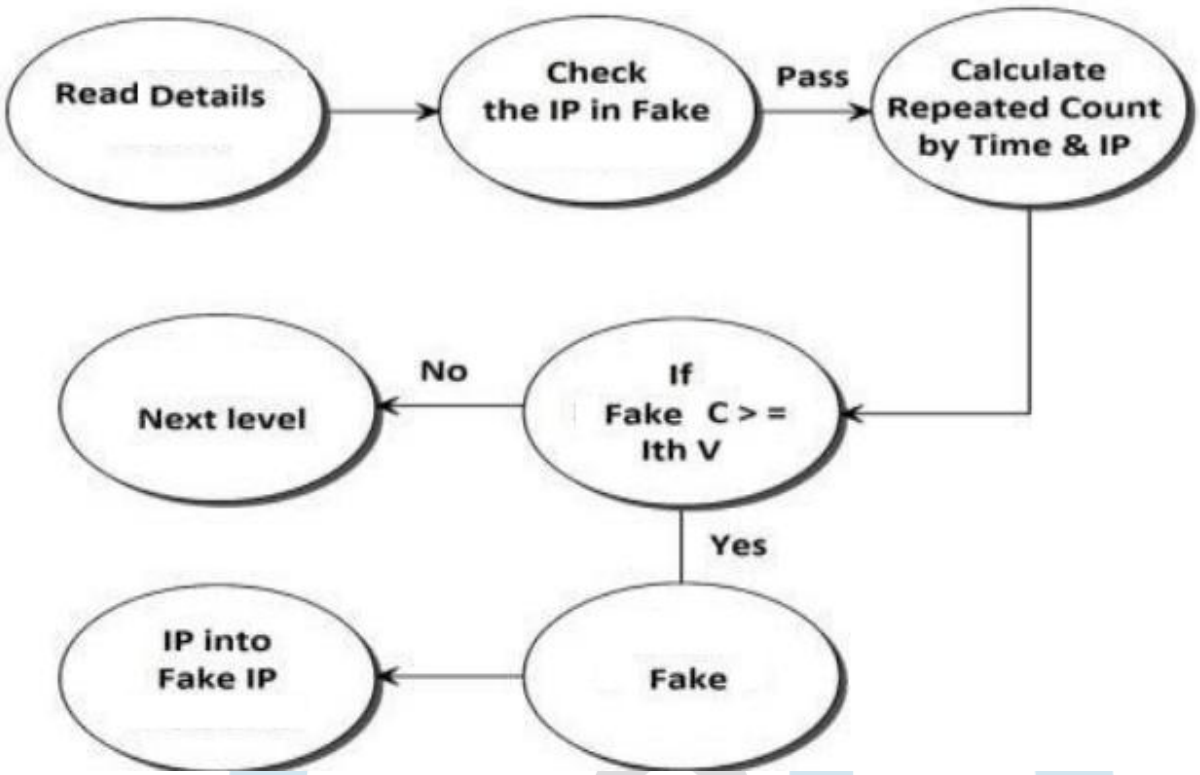
Level 1 Data Flow Daiagram



Level 1 Data Flow Daiagram



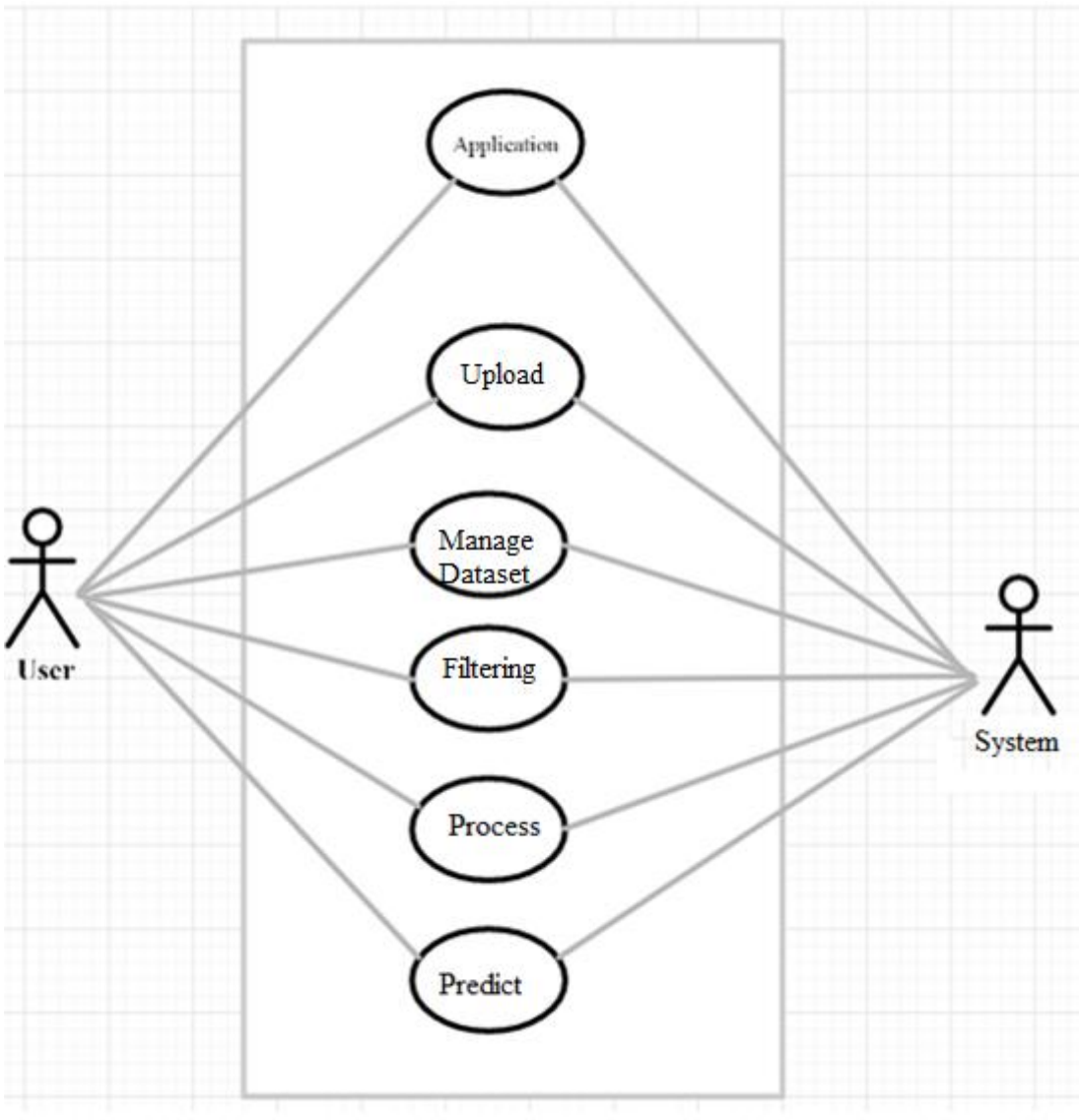
Level 2 Data Flow Daiagram



Use Case Diagram

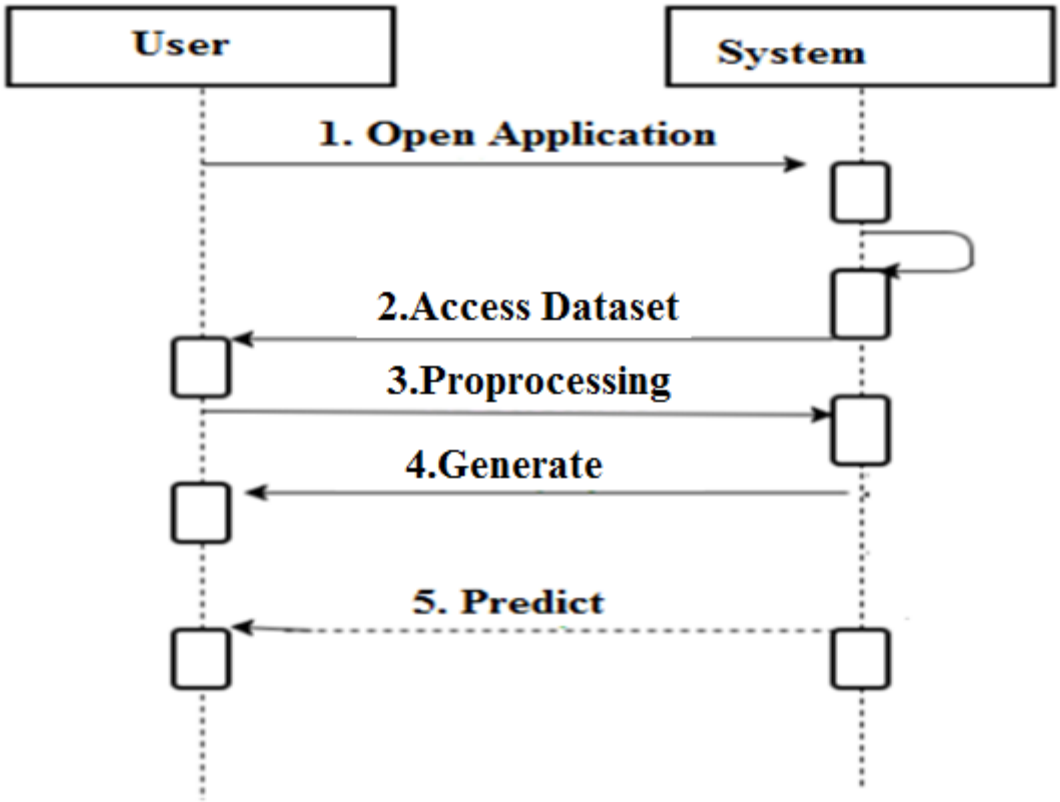
A use case diagram at its simplest is a representation of a user's interaction with the system that shows the relationship between the user and the different use cases in which the user is involved. A use case diagram can identify the different types of users of a system and the different use cases and will often be accompanied by other types of diagrams as well. The figure shows the use case diagram for the system.

The following figure shows the use case diagram:



System Sequence Diagram

A system sequence diagram is, as the name suggests, a type of sequence diagram in UML. These charts show the details of events that are generated by actors from outside the system. Standard sequence diagrams show the progression of events over a certain amount of time, while system sequence diagrams go a step further and present sequences for specific use cases. Use case diagrams are simply another diagram type which represents a user's interaction with the system. An SSD shows – for one particular scenario of a use case –



CHAPTER 7

SYSTEM IMPLEMENTATION

7.1 SYSTEMMAINTENANCE

System implementation is the important stage of project when the theoretical design is tuned into practical system. The main stages in the implementation are as follows:

- Planning
- Training
- System testingand
- Changeover Planning

Planning is the first task in the system implementation. At the time of implementation of any system people from different departments and system analysis involve. They are confirmed to practical problem of controlling various activities of people outside their own data processing departments. The line managers controlled through an implementation coordinating committee.

The committee considers ideas, problems and complaints of user department, it must also consider

- The implication of system environment;
- Self selection and allocation for implementation tasks;
- Consultation with union sand resources available;
- Standby facilities and channels of communication.

7.2 TRAINING

To achieve the objectives and benefits from computer based system, it is essential for the people who will be involved to be confident of their role in new system. These involve them in understanding overall system and its effect on the organization and in being able to carry out effectively their specified task. So training must take place at an early stage. Training session must give user staff, the skills required in their new jobs.

7.3 SYSTEMTESTING

It is the stage of implementation, which ensures that system works accurately and effectively before the live operation commences. It is a confirmation that all are correct and opportunity to show the users that the system must be tested with text data and show that the system will operate successfully and produce expected results under expected conditions.Before implementation, the proposed system must be tested with raw data to ensure that the modules of the system work correctly and satisfactorily. The system must be tested with valid data to achieve its objective.The purpose of system testing is to identify and correct errors in the candidate system. As important this phase is, it is one that is frequently compromised. Typically, the project schedule or the user is eager to go directly to conversion. Actually, testing is done to achieve the system goal. Testing is vital to the parts of the system are correct; the goal will be successfully achieved.

This creates two problems

- The time lag between the cause and appearance of theproblem
- The effect of system errors on files and records within the system, a small system error can conceivably exploded into much larger problem. Effectively early in the process translates directly into long term cost savings from a reduced number of errors.

7.4 CHANGEOVER

Changeover is the process where the existing system is converted into the new system. The changeover from old to new system takes place when:

- 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.

CHAPTER 8

CONCLUSION

8.1 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.