URL Based Phishing Detection Using a Hybrid Approach

Author

Muhammad Salman

NUST201362562MCEME35413F

Supervisor

Dr. Usman Qamar

DEPARTMENT OF COMPUTER ENGINEERING

COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY

ISLAMABAD

AUGUST, 2017

URL Based Phishing Detection Using a Hybrid Approach

Author

Muhammad Salman

NUST201362562MCEME35413F

A thesis submitted in partial fulfillment of the requirements for the degree of

MS Computer Software Engineering

Thesis Supervisor:

Dr. Usman Qamar

Thesis Supervisor's Signature:

_____

DEPARTMENT OF COMPUTER ENGINEERING

COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY,

ISLAMABAD

AUGUST, 2017

Declaration

I certify that this research work titled "*URL Based Phishing Detection Using A Hybrid Approach"* is entirely based on my personal efforts under the sincere guidance of my supervisor Dr. Usman Qamar. The work has not been presented elsewhere for assessment. The material that has been used from other sources it has been properly acknowledged / referred.

Signature of Student

Muhammad Salman

NUST201362562MCEME35413F

Language Correctness Certificate

This thesis has been read by an English expert and is free of typing, syntax, semantic, grammatical and spelling mistakes. Thesis is also according to the format given by the university.

Signature of Student

Muhammad Salman

NUST201362562MCEME35413F

Signature of Supervisor

Dr. Usman Qamar

Copyright Statement

Acknowledgements

Dedicated especially to my beloved, exceptional Parents and adored siblings, my respected Supervisor Dr. Usman Qamar, Co-Supervisor Dr. Farhan Hassan Khan, all family and friends whose tremendous support and cooperation led me to this wonderful accomplishment.

Abstract

Since the last decade, the traditional approach of physical business has changed to modern online business approach quite rapidly. The online businesses are accompanied by the online money transactions which exploited by the scammers led to a new type of online scam called phishing. In phishing what a scammer does is he creates a fake web page a visual replica of the original business web page to deceive the user/customer to input their credentials/personal information and in turn this credentials/personal information is then used by the scammer to perform fake money or business transactions. Here we have proposed an approach which combines the Blacklist-Based approach with the Heuristic-Based approach for phishing detection. In the first part of our approach the heuristic-based approach is used to generate rules that are then fed to the second part where based on these rules and the blacklist-based approach the phish is identified. Our approach outperforms the previous approach in terms of better accuracy in the detection of the phishing.

Key Words: *Blacklist-based Phishing Detection, URL Based Phish Detection, Hybrid approach*

Table of Contents

List of Figures

List of Tables

CHAPTER 1: INTRODUCTION

In the recent years, much has been heard about the Phishing attacks. By phishing it is meant to be a scam email, telephone call or text message used to steal the user data and information including important credentials etc. In social engineering, phishing is considered a crime in which a phisher targets and retrieves the confidential or other important credentials. Afterwards, this information is used to gain access to financial accounts, confidential secrets and credentials resulting in the losses at institutional to organizational levels [1] [2]. As Phishing is no longer restricted to sending scamming emails, combatting phishing has become top-priority focus of much work, both to the academia and industry [1] [3]. This research work focuses on the phishing detection based on the URL.

1.1 Background, Scope and Motivation

1.1.1 Background

First phishing attack was observed on America online network systems (AOL) in the early 1990s [4] where many fraudulent users registered on AOL website with fake credit card details. AOL passed these fake accounts with a simple validity test without verifying the legitimacy of the credit card. After activation of the fake account, attackers accessed the resources of America online system. At the time of billing, AOL determined that the accounts were fraudulent, and associated credit cards were also not valid; therefore AOL ceased these accounts immediately. After this incident, AOL took measures to prevent this type of attack by verifying the authenticity of credit card and associated billing identity, which also enabled the attackers to change their way of obtaining AOL accounts. Instead of creating a fake account, attackers would steal the personal information of registered AOL user. Attackers contacted registered AOL users through instant messenger or e-mail and asked them to verify the password for security purposes. E-mail and instant messages appeared to come from an AOL employee. Many users provided their passwords and other personal information to the attackers. The attackers then used the variously billed portions of America online website on behalf of a legitimate user. Moreover, an attacker no longer restricts themselves to masquerading America online website but actively masquerade a large number of financial and electronic commerce websites.

1.1.2 Scope:

The research in this desertion covers the detection of phishing based on two approaches combined; first one is blacklist based and second is heuristic based approach.

1.1.3 Motivation:

The motivations which encourage me for doing this research are:

1. Current anti-phishing techniques don't handles some of the legacy solutions which prove safe heaven for the attackers; blacklisting techniques will not be effective if these rule lists have not been [5].
2. Phishing attacks have been increased over the last decade resulting in great losses to the individuals as well as organizations. In the year 2008, US Internet users Studies show that more than 5 million US Internet customers lost significant volumes of their money on due to such phishing attacks [6].
3. Attackers are using different techniques for gaining access to the users' confidential credentials [7].
4. Security parameters are ignored by most of the internet users because of extra effort that security requires.

1.2 Phishing

In electronic communication phishing is referred to as an attempt to gain access to the users' sensitive information like id, passwords, financial details, confidential information etc. for some illegal activity or harassment [8] [9]. The word phishing is coined from homophone of fishing as it is related to the bait to a victim in a similar way fish is captured capture the fish. Microsoft Safety Index in 2014 has shown that the losses endured by the victims were as high as US$5 billion. [10]

Spamming the emails are common methods to carry out phishing and in that email a guideline in provided to the users to access their financial or other important account by clicking the link which is provide in the email. In fact that link is not a legitimate one but the phishing webpages with same look and feel as of the original websites. Phishing e-mails contain suspicious links to websites that are having been contaminated with viruses.

In social engineering, phishing is used to deceive the users. These attacks have exploited the

security perspectives of the websites [11]. Measures have been taken to overcome the phishing attacks causing great losses to peoples.

1.2.1 Phishing Lifecycle

Huang *et. al.* [12] has defined the five steps for typical phishing attack . The whole process has been also described in the figure 1.1 and steps are also described as follows:

1- Attackers develop a Phishing page.
2- After developing fraudulent website they send scamming emails, to large number of people, containing the link to the attacker's website.
3- Mostly people are lured to visit that fake website and when they visit the fake website they got compromised of their confidential credentials.
4- Confidential information is obtained by the attackers using their fraudulent website.
5- Using this information attacker's access the victims' financial accounts.

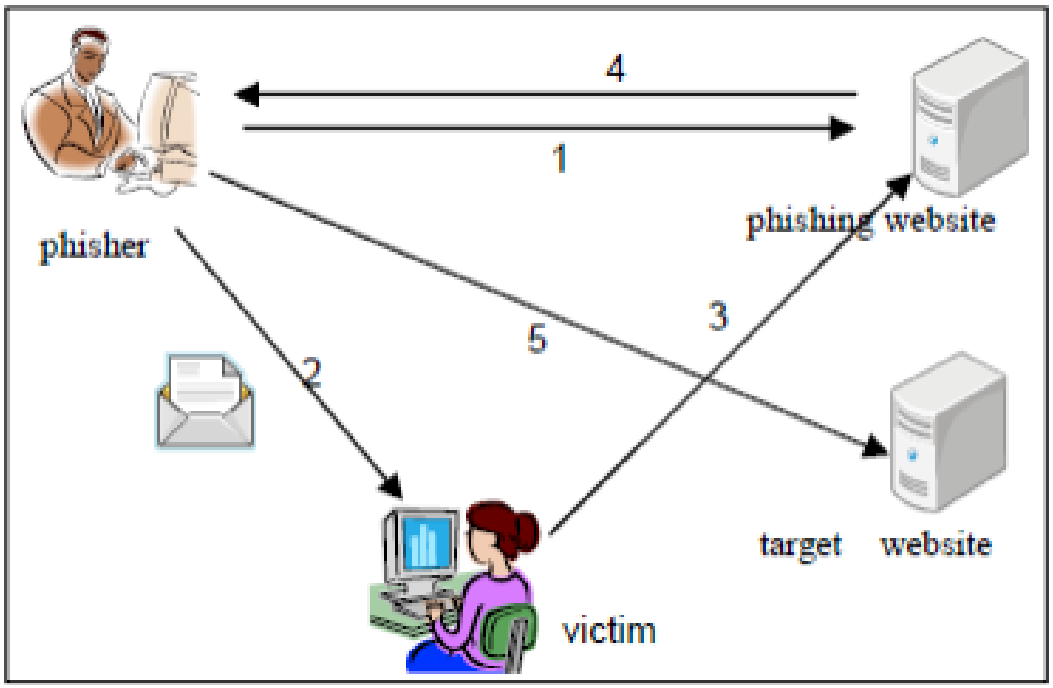

Figure1. 1 : Lifecycle of Phishing Attack

1.3 Phishing types

1.3.1 Spear phishing

Spears phishing refers to an attempt made to access secret information of specific individuals or companies [13]. Hackers use social media or other spying techniques to gather important

information before planning to make an attempt to attack ay user with phishing website. According to the studies almost 91% of the attacks remain successful by these techniques [14].

1.3.2 Clone phishing

Another type of phishing is clone phishing. It works on the concept that attackers made clone email similar to that of the previous legitimate email and embed a malicious or phishing link in it. The email id and its content are similar to or look similar to that of the original email. These phishing emails contains the malicious link claiming to be rechecking the previously sent information or warning to verify the information for security purposes.

1.3.3 Whaling

Another type of phishing attack is Whaling. The term Whaling is coined from idea that senior executives and high profile individuals are targeted within the businesses to gain some personal benefits. [15] For attacking such cases in Whaling, it takes some serious executive level in designing concealed webpages for them. As in this attack, higher management is the target; the content is designed by keeping in mind about the target. For this purpose, Whaling is mostly written as a response to a complaint, executive level query, or some legal issue. For Whaling scam, subterfuge emails are easily designed to be sent to the legitimate executive authority. The content in the email contains some specific country wide concern about the higher management. On such incidence had happed when phishers had sent forged FBI summons email containing some content claiming to be clicked by managers to install special software for viewing specific summon. [16]

1.3.4 Link manipulation

As it has already discussed that in phishing some technical fraud is created that is sent through the email or spoofing website [17]. Mostly links in the emails or the link mentioned on the spoofing websites are misspelled or tiny URLs containing subdomains are sent by the phishers. For example a URL: http://ww.hbl.abc.com appears to be a legitimate URL as it contains the “hbl” subsection of your bank ’hbl’. But in actual it point to a phishing link “hbl.abc” on the domain name of website. In other method of link phishing a text is use to be mentioned in <XYZ> tag that suggests to be a reliable website as it is clicked by the users. In some of the link phishing attacks, links are associated to the original link in the status bar

while a user hovers over it. However in some of the cases this action is overridden by the phishers. In addition, this feature is not available in mobile apps [18].

1.3.5 Filter evasion

Besides some of the common phishing techniques, Phishers have started using images instead of textual methods which have made it harder for anti-phishing filters to be detected in phishing emails [19]. Though this is evolutionary technique but it has made it difficult for anti-phishing filters used for recovering hidden text. OCR based filters are used to scan such types of images to reveal hidden information [20].

With the advancement of technology new filters for anti-phishing have been developed. Some of these filters use intelligent word recognition filters. These filters are useful in detecting hand written, inverted or even distorted (vertically, horizontally or in any direction) texts.

1.3.6 Website forgery

Phishing attack is just not over by just visiting the phishing website because sin some of the phishing scams, attackers uses JavaScripts to change the browsers address bar [21]. And this functionality is achieved by using picture over the address bar for the legitimate URL. In second method this purpose is achieved by closing the original link and by opening the new address bar with the legitimate URL [22].

Some of the time, trusted websites have flaws which are used by the phishers to use their scripts against the victims [23]. Such type of attacks are called cross-site scripting attacks and proves to be problematic as they popup the users to sign in using their phishing page. Cross-site scripts forces the security certificate to appear correct. But in reality these links are originally designed to carry out phishing attacks on the client side making it difficult to filter out the attack without prior expert knowledge. Once during 2006, paypal was attacked by using such flaws [24].

In 2007, a Universal Man-in-the-middle (MITM) Phishing Kit was found. Main purpose of these kits is recreating duplicate copies of legitimate websites which are finally used to retrieve login details using designed phishing websites.

Phishers have started using Flash-Based (also called Phlashing) website for phishing purposes related to textual attacks for preventing anti phishing filters. Such flash-based

website have look and feel very similar to the legitimate websites but in them phishing written text is covered up in a multi-media object [25].

1.3.7 Covert redirect

Covert redirect is a simple technique to execute phishing strikes that creates hyperlinks appear genuine, but actually divert a sufferer to an assailant's web page. The defect is usually masqueraded under a log-in pop-up centered on an impacted site's domain. [26] It can affect OAuth 2.0 and OpenID based on well-known exploit parameters as well. This often makes use of open redirect and XSS vulnerabilities in the third-party application websites. [27]
Regular phishing efforts can be easy to identify because the harmful page's URL will usually be different from the actual website weblink. For secret divert, an opponent could use a actual website instead by corrupting the site with a harmful sign in pop-up conversation box. This creates secret divert different from others [28] [27]. In case the "token" has greater privilege, the attacker could obtain more sensitive information including the mailbox, online presence, and friends list [29]. This could potentially further compromise the victim. [30]
This vulnerability was discovered by Wang Jing in Singapore. [31] Hidden transmit is a distinguished safety fault, though it is not a threat to the Internet worth noteworthy consideration. [32]

1.3.8 Social engineering

Classified contents are forced to be visited by some of the incentivized web links by forcing the users to click them to precede further the websites for some social or technical causes. [33]. Otherwise, some of the link forces users to click the link for some unexpected events or news that had happed based on some fake news. [34]

1.3.9 Phone phishing

As discussed earlier attacks which primarily involves the websites, but fake websites are not the host for phishing all the time. In some of the cases short messaging plays the same purpose as the fake websites do. The message or phone call seems to be from your parent banks financial department asking to verify some of the credentials as a routine security check or updating [35]. The phishers get personal information in the similar ways and uses that information in frauds or financial losses to the people [36]. Similarly the SMS is sent

using some similar phone number and retrieve the internal information using those fake messages resulting in people to disclose their delicate data. [37]

1.3.10 Other techniques

Besides the common techniques there are some of the uncommon yet important techniques used for phishing purpose. Clients are redirected to their bank websites using the phishing pages. These pages look same as the legitimate bank websites. The phishing pages places a popup window which is requesting the user to again enter the credentials [38].

Tabnabbing: This is another way of inducing attacks by taking advantage of tabbed browsing. Multiple new tabs get opened simultaneously. In this method, users are silently redirected to the phishing pages.

Evil twin: This method is not so common and also difficult to be detected. In this method attacker makes bogus wireless communication links that looks same as the legitimate unrestricted network that is publically available.

1.4 Phishing Attacks Statistical Overview

Recent statistical results have shown that during the year 2016 associated phishing attack were recorded to be the highest by the APWG since the time it started monitoring phishing attacks in 2004. AWPG also observed that during the fourth period of 2016 phishing attacks were reported to be higher than any period in 2015.

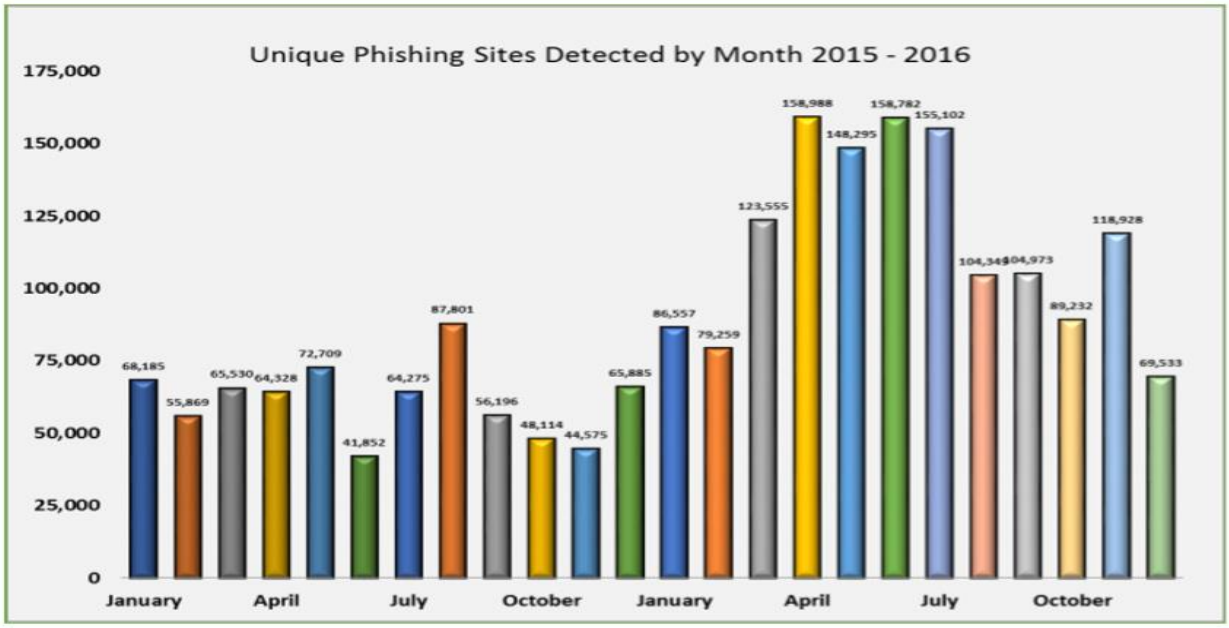

Figure1. 2: APWG 2015-2016 Report

During the year 2016 65% more phishing attacks were reported as compared to the year 2015. In 2016 total attacks were around 1,220,523. Similarly, 1609 attack were reported during every moth of fourth period during 2014. But as compared to the year 2014, APWG has recorded an average of 92,564 phishing attacks/month during fourth period of year 2016. [39]

The table below shows the number of phishing reports received from 2010 to 2016 in each month. The number has increased tremendously over the past years as it is also given in the table 1.1. It is clearly observed from the given table that the total reports received in the year 2010 are 3,13,527 which has increased to 13,80,432 in year 2016, this is approximately 10 times increase in the phishing attacks in the last 6 years.

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total
2010	29499	26909	30577	24664	26781	33617	26353	25273	22188	23619	23017	21020	313517
2011	23535	25018	26402	20908	22195	22273	24129	23327	18388	19606	25685	32979	284445
2012	25444	30237	29762	25850	33464	24811	30955	21751	21684	23365	24563	28195	320081
2013	28850	25385	19892	20086	18297	38100	61453	61792	56767	55241	53047	52489	491399
2014	53984	56883	60925	57733	60809	53259	55282	54390	53661	68270	66217	62765	704178
2015	49608	55795	115808	142099	149616	125757	142155	146439	106421	194499	105233	80548	1413978
2016	99384	229315	229265	121028	96490	98006	93160	66166	69925	89232	118928	69533	1380432

Table 1.1: Total number of unique phishing reports received, according to APWG

MarkMonitor Inc. found that companies in the Retail are the high value targets that constitute around 41.85% of the total phishing attacks, on the second is the financial services sectors with 19.60% and ISP is third on the chart with 12.458% and Payment Services are at 4th with 11.33%.

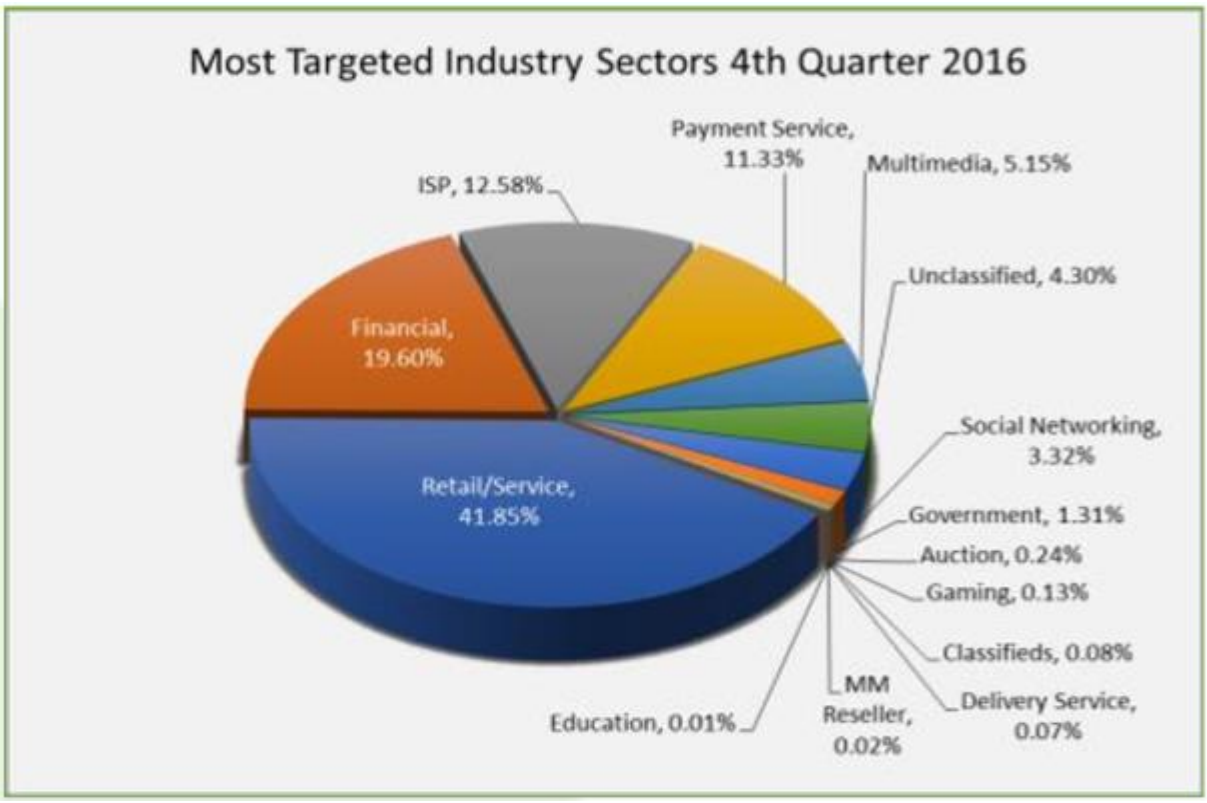

Figure1. 3: Most Targeted Industry Area

1.5 Phishing Detection

1.5.1 Phishing Tactics

Phishers use various tricks to carry out a successful deception. These tricks include the following:

1. Link manipulation (the contents of<A>tag content are made to display a web link going to an authentic URL, where as in the background it actually goes to a phished or malicious URL).
2. Evading phishing detection filters [19] (with the use of images instead of text that can remain undetected by many phishing filters [20]).
3. Malicious use of web scripting languages (using Java script to hide browser address bar and create a custom address bar displaying a hard coded authentic URL to the user).
4. Using pop-up windows to ask user names and passwords.
5. Utilizing browser vulnerabilities (e.g., Tabnabbing).

Figure1. 4: Phishing Tactics

1.5.2 Phishing Prevention Schemes

Phishing prevention schemes try to prevent phishing attacks by providing an extra layer of security to the authentication schemes and user interaction platforms (via two factor authentication and two-way authentication). This reduces the probability of a user being deceived by an attacker's phishing website. Some of the advanced phishing prevention practices include watermarking, RFID-based, external authentication devices based, picture password based, dynamic security skin based, smart card based, and QR Code based techniques, and so on. [40] These techniques can prevent most phishing attacks, but they require changes and support on the website's side and cooperation and understanding on the user's side for their success. Furthermore, these solutions may lead to complex user interfaces, may incur extra cost for the computation of each authentication, and may also require users to keep extra authentication devices, making it cumbersome to implement and use. Figure 1.5 shows the aforementioned phishing prevention schemes.

Figure1. 55: Phishing Prevention Schemes

1.5.3 Phishing Detection Schemes

Recent researches have provided many state of the art schemes in the area of phishing detection, which can help industries, researchers, and academia to review latest schemes with their pros and cons and find the most suitable scheme for phishing detection at their end. A broad classification of phishing detection schemes based on the underlying technique utilized for phished website identification is shown in Figure 6. The techniques are majorly classified as:

1. Search engine based (SEB)
2. Heuristics and machine learning based (HMLB)
3. Phishing blacklist and whitelist based (PBWB)
4. Visual similarity based (VSB)
5. DNS based (DNSB)
6. Proactive phishing URL detection-based (PPUDB) schemes.

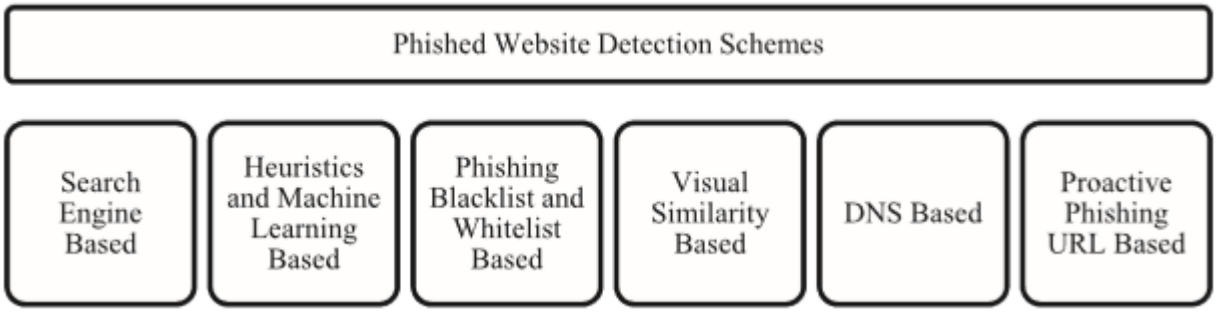

Figure1. 6: Classification of Phishing Detection Schemes

1.5.3.1 Search engine based

In search engine-based phishing detection technique, features are extracted using multiple or single search engines and later on findings are analyzed. Though, normal websites have higher index then phishing pages, this assumption, of remaining active for very short period of time, is used to differentiate between the normal and suspicious web pages.

1.5.3.2 Heuristics and machine learning based

These techniques extract a set of features of text, image, or URL-specific information from normal or abnormal websites. A set of heuristics is utilized, and the thresholds or rules obtained from the learning algorithms are used for anomaly detection.

1.5.3.3 Phishing blacklist and whitelist based

The methods in this category utilize the whitelist of normal websites and the blacklist containing anomalous websites to detect phishing. The blacklist is obtained either by user feedback or via reporting by the third parties who perform phishing URL detection using one of the other phishing detection schemes.

1.5.3.4 Visual similarity based

The technique utilizes the visual similarity between webpages to detect phishing. When phishing web sites are matched in terms of their visual characteristics with the authentic websites, it checks whether the URL is on the authentic domain URL list. If not, the website is marked as a phishing website.

1.5.3.5 DNS based

DNS is used to validate the IP address of a phishing website. For example, DNS will identify whether the IP address over which the phishing website is running is on the list of authentic website IPs. If it is not, the website is marked as phishing. DNS can also be utilized by these techniques in other ways, based on the needs of the user.

1.5.3.6 Proactive phishing URL detection based

This scheme detects probable phishing URLs by generating different combinatorial URLs from existing authentic URLs and determining whether they exist and are involved in phishing-related activities on the web.

1.6 Phishing Problem Type and Proposed Solution

1.6.1 Problem

In this research “Link Manipulation” based phishing problem is considered. Link based phishing problem is basically about the malicious URLs that are obscured to appear if they are linked to valid organizations resulting in difficultly to detect making it difficult to be detected. In Figure 1.7; an example is given showing the website that seems to be PayPal. Figure 1.8 shows the original website of PayPal. One can identify the differences between both images by deep observation only as the differences are not obvious to point out. Mostly, differences are observed in logos, SSLs and favicon. Phishers uses uncommon ways to redirect users to their malicious web pages and easily become victims.

Hackers use the desired confidential user information in multiple ways of forgeries resulting in losses to the people. These hackers also blackmail different political and media celebrities after stealing their personal credentials.

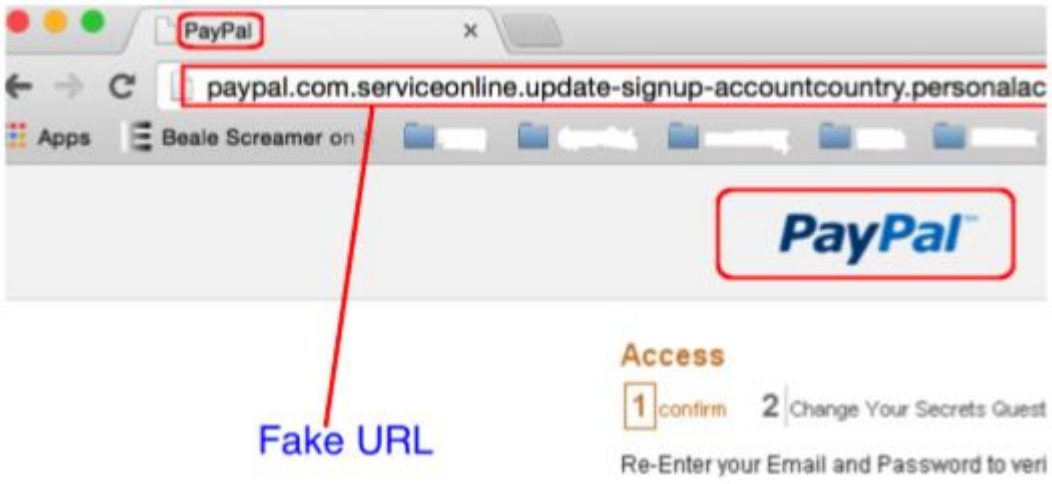

Figure1. 7: Phished URL of PayPal

Figure1. 8: Legitimate URL of PayPal

1.6.2 Proposed Solution

The detection approach that is proposed in this research is a hybrid approach of the “Blacklist Based” approach and the “Heuristic Based” approach. Here we have tried to cater the drawback of blacklist based approach by utilizing the heuristic based approach. The proposed solution is divided into two parts. In first part the Heuristic approach is used to generate rules while in the second part the blacklist based method along with the rules from the rules from the first part are used to detect phishing. The Abstract level diagram of the proposed solution is shown below in figure 1.9.

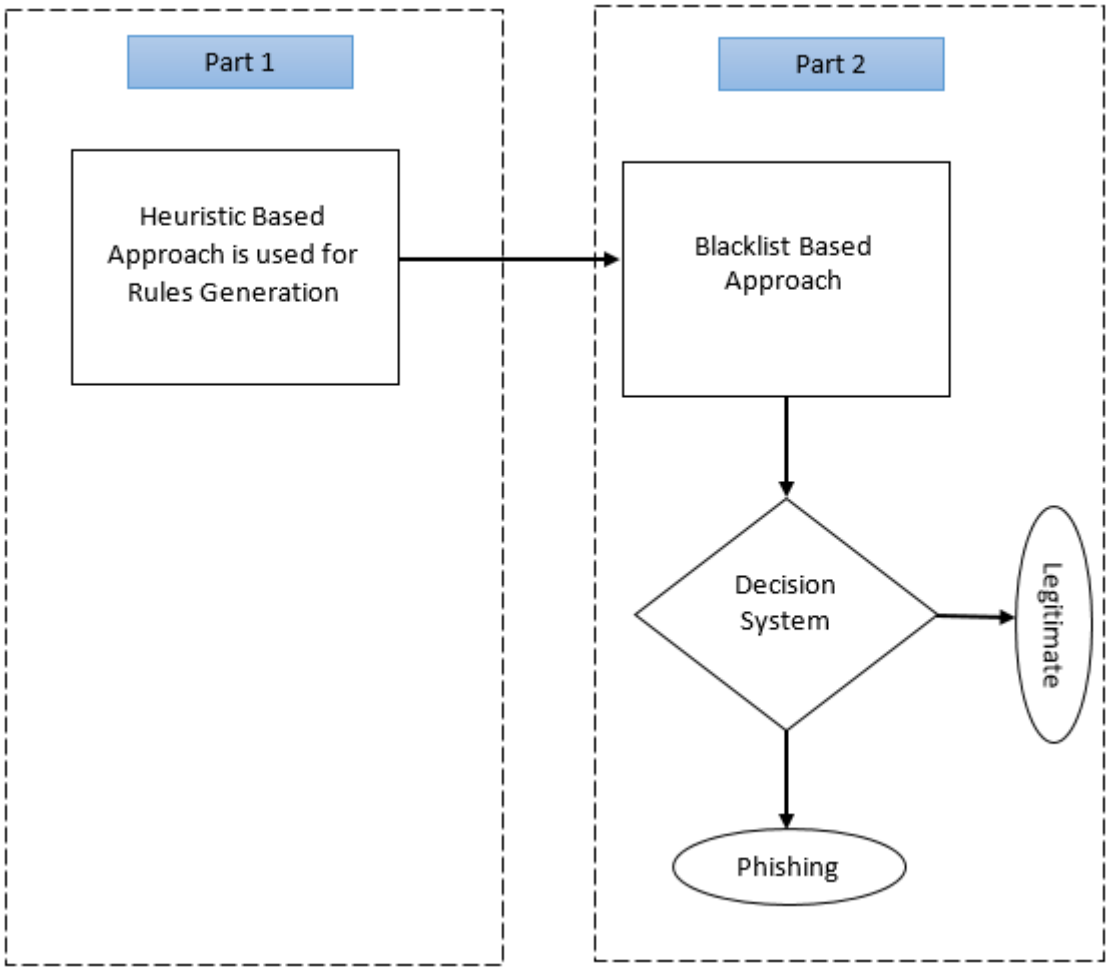

Figure1. 9: Abstract model of the proposed solution

1.7 Structure of Thesis

The research thesis has been structured in following sections:

Chapter 1 is about the introduction of the research topic and basic terminologies related the research.

In Chapter 2 Literature reviews has been discussed. Previous work relating to phishing detection and different methodologies with their advantages and limitations has been discussed.

Chapter 3 focuses on the proposed methodology based on the Hybrid Approach for phishing classification and detection.

Chapter 4- Experiment and Results: Describes the full experimental execution of the hybrid model and the results obtained as well as the comparison of the results.

Chapter 5- The Conclusion: To present a summary of the whole system, from the problem to model working to results.

CHAPTER 2: LITERATURE REVIEW

Since the last decade many phishing detection approaches have been proposed, here in this chapter we have explained some of the latest literatures that have contributed well to the detection of phishing solutions.

2.1 Effective CSS based feature extraction of web pages

In this paper [41] *Visual Similarity* technique of web pages is used for detecting phishing. The researcher has used CSS based features for measuring the similarity of different suspicious pages and selected the effective feature set for similarity rating. In visual similarity technique elements and CSS rules of web pages are used for determining the visual appearance of web pages. In CSS, a *selector* and a series of *declarations* are used to define the CSS rules. *Selector* is a pattern to specify HTML elements. *Declaration* comprises two components, property followed by its value. Subsequently, *CSS rules* are organized in the following formats.

- Selector1 {Property1-1: Value1-1; Property1-2: Value1-2; ...};
- Selector2 {Property2-1: Value2-1; Property2-2: Value2-2; ...}; ……

Page appearance will not be affected by the element of matching the page. We call the set of CSS rules that are actually affecting the web page appearance *effective CSS rules*. The proposed solution has three main steps with the following respective objectives:

2.1.1 Extracting and representing effective CSS features

For extracting and representing effective CSS feature, CSS structure *CSS (Ps)* and elements *Ele (Ps)* of a suspicious page *Ps* is determined. Afterwards, from interaction between these *CSS (Ps)* and *Ele (Ps)* identification of the set of effective CSS features *ECSS (Ps*) is done

2.1.2 Measuring Similarity between the Suspicious Page and the Target Pages (Computing similarity scores)

On the bases of effective CSS features of the suspicious page and prospective target pages, we design metrics to measure their *complexity scores* and *similarity score* accordingly.

2.1.3 Detecting phishing pages

It is checked whether the pages *similarity score* is over a preset threshold *E* or not. By the *similarity score* and a list of *target web sites*, the approach adopts if the apprehensive page is a phishing one or not.

2.1.4 Advantages:

Efficiently detects the phishing on top web sites with high accuracy.

2.1.5 Limitations:

The domains that have a low ranking will not be in the Target List & hence will not be identified.

2.2 New Rule-based Phishing Detection Method

In this paper [42] the author introduced a new set of features Proposed Features along with the Relevant Features that are extracted from the previous work experience and created a new feature vector for the phishing classification. SVM approach was used for classification based on the new feature vector. For Rules generation Decision Tree algorithm C4.5 is used.

The relevant 17 features which are used in this paper are IP address (F1), Web address length (F2), SSL certificate, we define (F3) for length of the “host”, (F4) for the length of the “path” Number of dots in URL (F5) and (F6) for the length of “file” and "query” parts. Blacklist keywords (F7) feature for the rate of keywords appears in <host> part, (F8) as the rate of keywords appears in <path> part and (F9) as the rate of related keywords in <query> part of a URL.

The proposed features are evaluated from Levenshtein distance for Approximate string matching Algorithm (F10) for LD_Links, (F11) for LD_JS, (F12) for LD_CSS, (F13) for LD_Images and from Page resource access protocol as (F14) to show the rate of secure access links, (F15) to show the rate of secure access JavaScript files, (F16) to show the rate of secure access style sheet files, (F17) to show the rate of secure access to images.

Dataset comprises 3066 phishing web pages and 686 legitimate webpages collected from PhishTank (*http://www.phishtank.com*) and Yahoo directory service (*http://dir.yahoo.com*) respectively.

For rule generation N subsets of the dataset with equal size and class distribution were used. N subsets are tested for N times test runs each with an altered subset, then save the result of extrapolation for each subset. Then correctly predicted cases by SVM are selected into new data set. Now create the artificial datasets from the initial test blocks plus the new truly classified and train on the DT. Finally, common rules in each category were removed or the rules with confidence less than 50% were removed and again formed DT to combine the overlapped rules & in the end got 10 rules.

2.2.1 Advantages:

A new feature set was used proposed based on the content of the page the improved the accuracy of detection.

2.2.2 Limitations

Proposed approach is entirely depending on the webpage content. Means if images or flash is used to display content then with this approach we would be unable to detect phishing.

2.3 Supervised Learning Based Model for Phishing Sites Detection

Authors in this paper has proposed a hybrid model for detection of phishing. [43] Initially the author took 7 classifiers [Random Forest (RF), Decision Tree (J48), Sequential Minimal Optimization (SMO), Bayesian net (BN), Naive Bayes (NB), Fuzzy Unordered Rule Induction (FURIA) and Instance based learning (IBk)] and trained them all individually on the dataset obtained from UCI repository having 30 attributes & 11055 instances. The results showed that IBk and RF accuracies greater than 95%. Then these two approaches were combined with other low performance classifiers to improve their accuracy/performance.

From the above different combinations we got hybrid classifiers that almost outperformed all the individual classifiers.

The hybrid approach of J48+IBk and BN+IBk resulted in the highest accuracy among other hybrid classifier.

2.3.1 Advantages:

Performance of weak classifiers can be improved by combining with one or more strong classifiers.

2.3.2 Limitation:

The proposed approach did not take into consideration the curse of dimensionality. The number of feature needs to be reduced.

2.4 Online credibility and performance data using Machine Learning

This paper [44]used two datasets with malware and phishing domains. 16 data features and 09 machine learning models (5 distinct classifiers and 4 ensemble classifiers) were considered for results.

Furthermore, for improving the performance of single classifier, a binary particle swarm's optimization (BPSO) feature selection method was used. R programming environment was used to run all reported experiments along with BPSO-based feature selection technique.

2.4.1 Advantages:

One great advantage is that we don't require the feature extraction from the content or url. The other advantage could be conclusion that ensemble models outperform single models as well as the feature selection based single models.

2.4.2 Limitation:

The sites that are newly built will not have great values for the attributes selected by the author, hence will be miss-classified.

2.5 New Fast Associative Classification Algorithm

In this paper AC mining classifier is employed which is also known as the Fast Associative Classification Algorithm (FACA) which is based on the vertical mining based Diffsets for discerning all recurrent item sets.

Fast associative Classification Association Algorithm (FACA):

For detection of single items, FACA scanning is done on all instances of the training sets, then it combines single item set rule to get rules with two items in body & so on and removes the rule with support less than the minimum support threshold. Once all recurrent rules have been revealed, confidence value of all rules is calculated, rule with confidence value equal to

or greater than minimum confidence will be added to classification association rules (CARs) otherwise will be removed.

FACA algorithm sorts rules on CARs in order to give the more helpful rules a higher priority for being selected as part of the model.

- Rule with least number of feature values in its body is given a higher rank.
- For features with same feature values, feature with higher confidence value is given a higher rank.
- If same confidence values in two or more rules occurs then the rules with higher support is given a higher rank.
- For a situation with prevailing similar criteria occurs then the rule that was produced first is given a higher rank.

FACA begins with high rank rules & examines it on all training occurrences; if a rule matches at least one instance then it is added to FACA model. All the occurrences that equal the rule body & head are added to FACA otherwise deleted.

The process is repeated until no more training instance or rules in the CARs are remaining. This help to get only the useful rules.

2.5.1 Advantages:

FACA outperforms in terms of accuracy, and with the lowest error rate than that of other related classifiers like CBA, CMAR, MCAR, and ECAR by 4%, 3.7%, 3.2%, and 2.9% respectively.

2.5.2 Limitations:

Class assignment to an instance is done by the max count in a cluster class while all the rules are ranked so it should also take into account the rank of the rules along with the count.

2.6 Effect of Feature Selection on Phishing Website Classification Problem

This paper proposed an effective model based on 4 Feature selection (dimensionality reduction) and 5 classification algorithms to classify a websites as legitimate or phishy on a data mining tool WEKA. [45]

Feature selection algorithms- Correlation Feature Selection (CFS), Information gain (lG) and Consistency-Based Subset and for extensity reduction Principal component analysis (PCA).

Classification algorithms - We have used J48, Naive Bayes, SVM, Random Forest and AdaBoost in order to find the most reliable technique by comparing their accuracy and AUC for classification of phishing websites.

Dataset: 30 Attributes whose details are given in the . Total Instances 2456, phishy instances are 1094 & legitimate website instances are1362.

2.6.1 Advantages:

The feature selection step save us time as well as space. Consistency subset produces high accuracy rates of 97.4756 % by using 15 features of the 30 features, thus saves experimental time.

2.6.2 Limitations:

If we change the features in the dataset it might give us different results, should consider weighted features.

2.7 An Efficient Approach Using Single-layer Neural Network

The proposed model is built to detect phishing sites by using single-layer neural network and six heuristics/features. [46]

Dataset: Training dataset containing 11,660 sites and 2 testing datasets that each dataset contains 5,000 phishing sites and 5,000 legitimate sites. The best results show that 98.43% phishing websites are detected.

In the proposed method the value of features are calculated from url & valid internet resources and not depend on training dataset and also the weights of heuristics are more optimize

because the weights are trained by neural network. The model is classified into two classes so the site is phishing if the value of the output node is less than 0.5 and legitimate if the value is greater than or equal to 0.5. ANN algorithm performs two phases as:

- Propagation: This phase calculates two values, the input value of the output node, the output value of the output node.
- Weight Update: This phase calculates the error of the output node and updates the weights

2.7.1 Advantages:

Using neural network the feature weights are better optimized automatically.

2.7.2 Limitation:

Three features of the features will not give good value for the new site & will eventually give wrong results.

2.8 Hybrid Model Using Clustering and Bayesian Approach

In this paper the author selected features of URL along with the features of Web Page. [47] K-Means Clustering is applied on initial URL features and Validity is checked if still we are not able to determine the Validity of Web Site then Naive Bayes Classifier is applied onto URL as well as HTML tag features of Site. System Architecture steps:

- Step 1: Given the web site X.
- Step 2: Extract the URL features from X.
- Step 3: Apply K-Means Clustering on dataset of X and predict the cluster in which the X is nearer to centroid (-1, 0, +1). //-1: Legit Site, 0: Suspicious, +1: Phishing Site
- Step 4: If outputis -1 or +1, predict the result. If output is 0 then go to step 5.
- Step 5: Download the source code of webpage and extract the HTML tag features and enter into X.
- Step 6: Classify X using Naive Bayes Classifier and predict the output -1 or +1.

The whole model can be split into two main modules, one URL Features & K-Means Clustering and second HTML Feature & NB Classifiers.

URL Features and K-Means Clustering:

- URL Features: IP Address in URL, Dots in URL, Suspicious Characters, Slashes in URL
- K-Means Clustering: We can create database with the application of clustering which is divided into three categories as Valid Phish (feature set containing higher values), Suspicious Phish (some feature with high value & some with low values so can be legit or phishing), Invalid Phish (Having very low values for the feature set).

HTML Features and NB Classifier:

- HTML Features: NULL Anchors, Foreign Anchors, SSL Certificate
- Naive Bayes Classifier: Here in this module the author have used NB classifier as it classify the unknown or null attribute values by omitting from the probability computation which is not handled by other classifiers like decision tree.

2.8.1 Advantage:

The proposed mechanism uses KMeans Clustering which is effective to produce output at higher throughput but with deficiency of competence and this deficiency of efficiency is improved with the Naive Bayes Classifier.

2.8.2 Limitation:

The second part of the method is dependent on the web page content, it page is built using image/flash it won't be detected.

2.9 Feature Selection for Improved Phishing Detection

In this paper the author evaluates two feature selection techniques correlation-based and wrapper-based with two feature space searching techniques genetic algorithm and greedy forward selection and machine learning algorithms used for classification are Naïve Bayes, Logistic Regression and Random Forests. [48] Using real-world phishing data sets with more than 16,000 phishing and 32,000 non-phishing webpages and 177 initial features.

Experiments: We used 10 times 10-fold cross-validation to estimate test accuracy and compare feature selection & search techniques by applying classification algorithms NB, LR & RF using a datamining framework WEKA. We also tried evaluating C4.5 and Multilayer Perceptron but the wrapper feature selection technique was slower taking months for this classifier.

Dataset: In First dataset (DS1) 11,240 phishing webpages form PhishTank are considered. The criteria of selection of these pages are done on the bases of submission before October 31, 2010 and while 21946 legitimate webpages are collected from Yahoo! and seed URLs. Second dataset (DS2) has 5,454 phishing webpages obtained from PhishTank which are acquiesced between January 1 and May 3 of 2011 and while number of legitimate webpages are 9,635 from DMOZ. Total of 177 features of which 38 are content-based and the rest are URL based.

Correlation Based Feature selection and Wrapper based Feature Selection is carried on DS1 using Genetic & Greedy Forward search on NB, LR & RF. Here NB with genetic search has select 42 features with improved accuracy & lower error. Smaller subset is selected with improvement in accuracies & degradation in errors. RF performs best with Greedy Forward search.

2.9.1 Advantages:

Using efficient search method improves the feature selection.

2.9.2 Limitation:

Main drawback about this technique is that by changing dataset the performance changes. Evaluate other feature selection techniques such as PCA, chi-squared attribute evaluation, latent semantic analysis etc. and other search approaches as best first and greedy backward eradication.

2.10 Automated Technique for Feature Assessment

Previously different researchers had used their experience for feature extraction but in recent studies new tool have been used for automatic feature extraction form different web links

instead of relying on manual or human experiences. [49] And for automatic feature extractions, new rules are developed for each feature which helps in developing applications for phishing detection. Using automatic feature extraction techniques help in dramatic increase in the dataset and it allows analyzing large number of phishing pages.

Among the different features frequencies of phishing datasets, "Request URL" has the highest significant in identifying phishing web links followed by "Age of Domain" which is presented by almost 2392 datasets. Subsequently, HTTPS and SSL ranks next with 92.8%. Whereas, “Disabling Right Click” is lowest significant feature for phishing which appears for only 40 times, followed by "URL having @ symbol" with only 3.6% appearance in datasets.

For identification of phishing web links following two approaches are deployed.

- Blacklisting(Compare of requested URL with those in the blacklist)
- Heuristic-based (feature selection from different web links to categorize it as phishing or legit based on some predefined criteria.)

Following two methods can be performed for Feature Extraction:

- Manual: User derive features & judge site legitimacy
- Automatic: Web page properties are usually derived from the HTML tags, URL Address & JavaScript code.

Almost 2500 phishing and legitimate web links are collected from the PhishTank archives.

- Features based on Address bar: For extraction of features based on address bar, a JavaScript program was built.
- Feature based on Abnormalities: for extraction of features based on some abnormal services, a PHP scripts was developed.
- Features based on HTML and JavaScript: for extracting such features a JavaScript program was built.
- Features based on Domain: As these features are to be extracted from the Alexa.com and WHOIS database, a PHP script was developed for feature extractions.

Prediction: Prediction is done on the bases of the rules manually generated.

2.10.1 Advantages:

The weight features can help classify more accurately.

2.10.2 Limitations:

Based on these rules it would be difficult to manually classify.

CHAPTER 3: PROPOSED METHODOLOGY

Our proposed hybrid model for URL based phishing detection is divided into two parts, first part is used for rules generation using heuristic approach and these rules are then forwarded to the second part where based on these rules & blacklist based approach phish is detected.

3.1. Part-1: Rules Generation Using Heuristic Approach

This part has further two sections, in the first dataset is generated based on the URL feature, that are selected on the basis our previously discussed literature review, and in the second section this dataset is then fed to three different rule-based machine learning models, rules generated by that model are selected that have high accuracy.

3.1.1 Dataset Generation

3.1.1.1.Phished & legitimate URLs

We collected the URLs from the two different resources to generate our dataset. The PhishTank provide the service of the phished URLs in four different formats i.e. XML, JSON, CSV, and Serialized PHP. We downloaded the CSV format file that contained 25121 verified phished URLs. For the Legitimate URLs we downloaded the CSV format file from Majestic that contained 1 million site verified and listed by rank order. From both the files we selected 1000 URLs each for our dataset creation.

3.1.1.2.Feature Generation & Extraction

We assigned the phishing URLs class label “1” and legitimate URLs as “-1”. We combined both type of URLs into a single file. We extracted seven features from the file using PHP/MySQL that is as:

- Have IP
- URL Length
- Have @ Symbol
- Double Slashes (//)
- Prefix-Suffix (-)
- Having Sub Domain (No. of Dots)

- https_token
- SSL_state

After creating dataset we further applied different feature selection techniques to select the feature set that are most relevant and remove the ones having less impact on the classification.

3.1.2. Applying Rule Based Model on Dataset

In the second section we applied rule based machine learning models on the dataset created in the first section. Rules from that model are selected that performs with high accuracy on our dataset.

3.1.2.1. C 4.5

For extracting rules while implementing C 4.5 with Indirect Method, unpruned decision trees are entry point for extraction of rules and for each rule,

$$r: RHS \rightarrow c$$

deliberating to pruning rule.

For class ordering:

- On the bases of simplest set of rules, classes tend to appear first
- Collection of each subset of rules belongs to the same rule class

3.1.2.2. RIPPER

To deal with 2-class problem with RIPPER method, Direct Method and Indirect Method is used,. Firstly Direct Method is discussed.

In Direct Method for 2-class problem, one class is chosen as a positive class whereas other one to be negative one. Rules Learning is done using the positive class and the negative class will be default one.

For considering multi-class problem in Direct Method, classes are ordered in the ascending class occurrence. Rules learning start from the smallest class first and the rest of the classes

are preserved as the negative classes. This procedure is reiterated for the subsequent smallest class to be treated as the positive class and so on.

Learn one rule:

- The process starts from empty rule
- Conjuncts are added as long as they improve FOIL's information gain
- The process is stopped when rule no longer covers negative instances
- Build rules with accuracy = 1 (if possible)
- Snip the rule instantly using reduced error pruning
- Pruning measurement: W(R) = (p-n)/ (p+n)

Where;

p→ number of positive examples covered by the rule in the validation set
n→number of negative examples covered by the rule in the validation set

- Optimization process is started after adding the last test is added to the rule. Some rules may also be created for covering few of the negative instances (accuracy < 1). A global optimization (pruning) strategy is also applied

3.1.2.3. PART

Divide-and-conquer strategy and separate-and-conquer are combines in Indirect Method for rule learning:

1. On current set of instances, a partial decision tree is built
2. A rule is created from the decision tree i.e. rule is made with the largest leaf coverage
3. Decision tree is discarded
4. Those instances which are covered by the rules are removed
5. Jump back to step one

3.1.3. Evaluation Methods

The performance of prediction models is shown by the confusion matrix as shown in Table 3.1.

	YES (predicted)	NO (predicted)
YES (actual)	True Positive	False Positive
NO (actual)	False Negative	True Negative

Table 2. 1: Confusion Matrix

TP = true positives: number of examples predicted positive that are actually positive

FP = false positives: number of examples predicted positive that are actually negative

TN = true negatives: number of examples predicted negative that are actually negative

FN = false negatives: number of examples predicted negative that are actually positive

To evaluate which model performs the best we utilized average model accuracy, recall, f-measure and precision using 10-fold cross validation tests as follows.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{.....(3.1)}$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad \text{.....(3.2)}$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad \text{.....(3.3)}$$

$$\text{F-Measure} = 2 \cdot \frac{precision \cdot recall}{precision + recall} \quad \text{.....(3.4)}$$

Now we select the rules generated by that approach that have higher values for the above performance measure and then these rules are converted into simple if-then structure and they fed as input to the second section/part of the model as given in the Figure 3.1 .

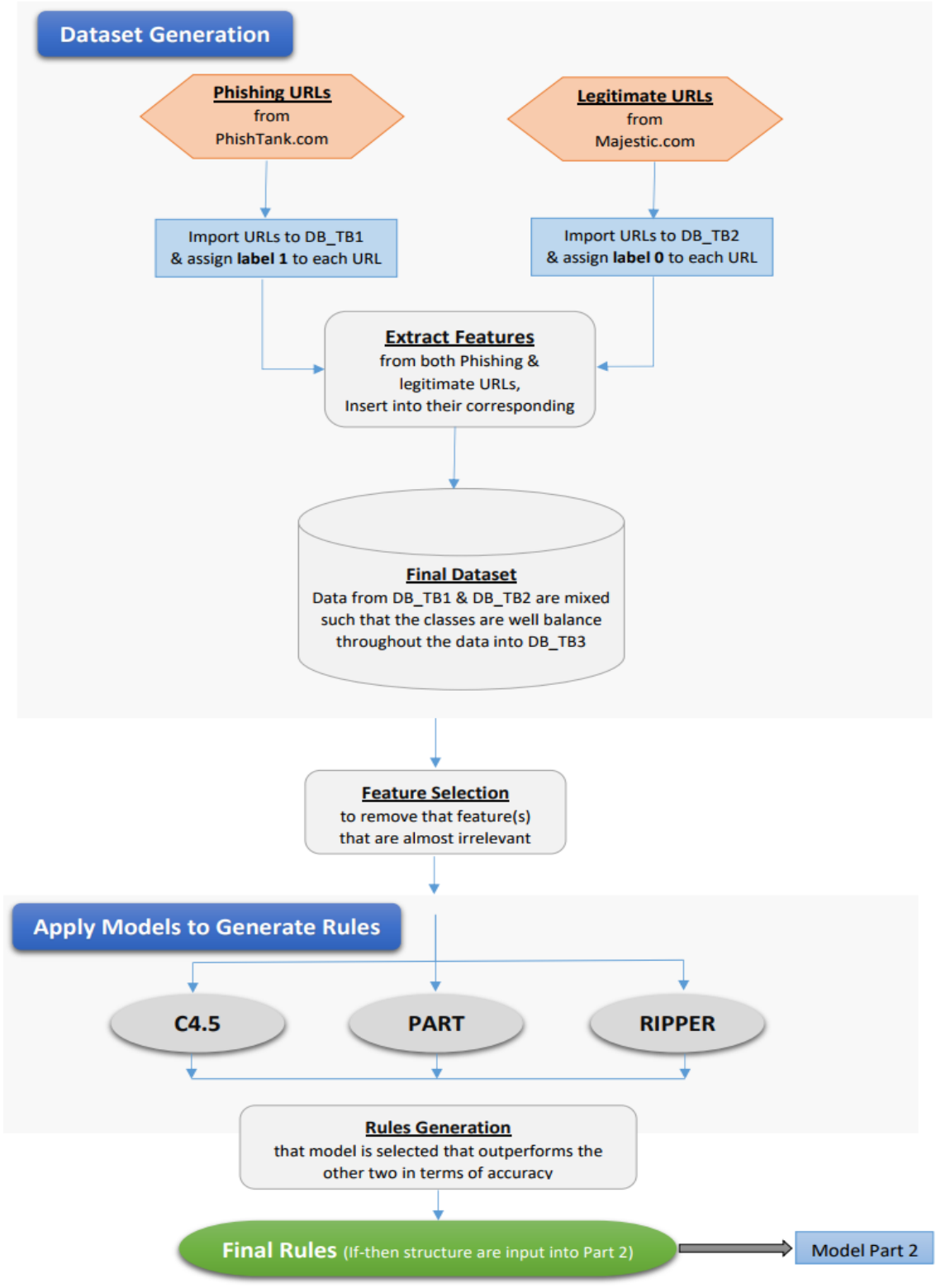

Figure 3. 1: Rules Generation Using Heuristic Models

3.2. Part-2: Proposed Hybrid Model

In part-2 of the model the rules from the part-1 act as input to this second part. Here in this part of the model the first we import all the 25,121 verified phished URLs from the PhishTank csv files into MySQL DB table, we then extract the host of the URL and also store it in the DB table alongside the URLs, further we generate hash of both the URL and the host and these two hashes are also store in the same DB table.

Whenever we have a new URL to check if its phished or not, we generate the hash of the URL and check it with the previously stored blacklist (verified phishes) in the DB if match is found the URL is declared as phish if not then we extract the host from this new URL and generate its hash which is checked against the stored hash of the hosts, if match is found URL is declared as phish if not then we extract the same features from the URL that we extracted in part-1.

These feature values are then fed into a decision system, it is here in the decision system that the rules generated in the first part are fed into the decision system. The URLs that are not detected by the blacklist, their feature values are extracted and sent to decision system. The decision system based on the rules from part-1 and the feature values make decision whether the URL is phished or not. If no rule is matched the URL is declared as legitimate where as if it matches even a single rule it is declared as phished and hash of this URL is generated along with the hash of its host and stored in our DB system so that if this comes again our model will detect it in the initial phase at the blacklist stage and won't have to go through the feature extraction and decision system which takes processing time. This URL reported to the PhishTank as well. The figure 3.2 shows the whole process of who the part-2.

We used hash for matching URL strings as its of fixed length, storing us the space as well as making process thing fast as if we try to match URL strings that have length more than 100 characters it would take more time than a simple hash matching, here we have used MD5 hash which returns a 32-character hexadecimal number.

Our model automatically updates itself with the new phished URLs resolving the issue that exits in the traditional blacklist based approaches.

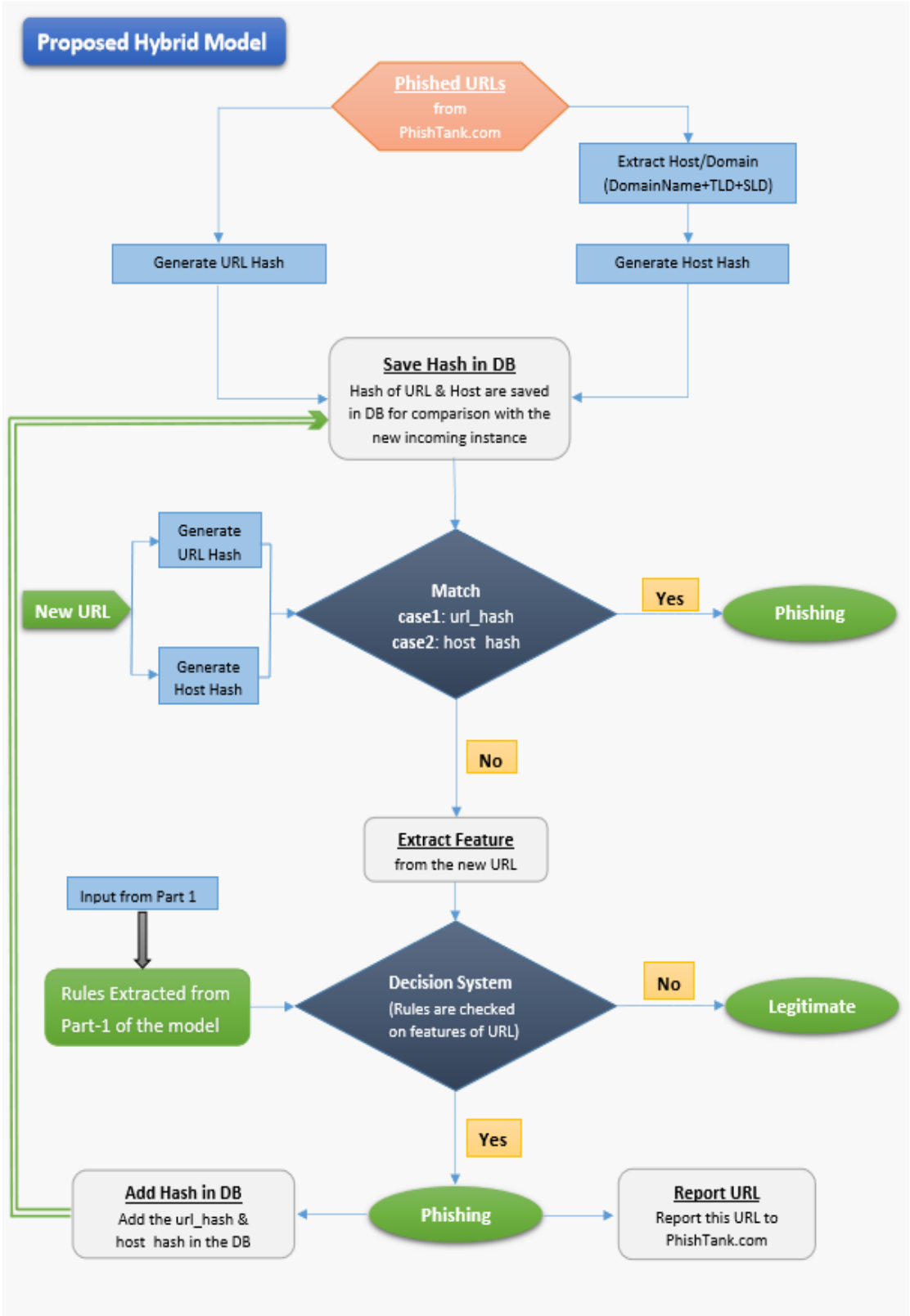

Figure 3. 2: Proposed Hybrid Model

CHAPTER 4: EXPERIMENT AND RESULTS

4.1 PART-1:

4.1.1. Dataset Generation

First we start by loading both the phished & legitimate URL files in the .csv format into MySQL database; both the files are loaded into separate tables. Then we specified eight (8) features that are to be extracted from both types of URLs based on our literature study. We have followed the same approach for feature extraction by rules as proposed by Rami at el [49]. Here a feature can have 2(Phishing, Legitimate) or 3 values (Phishing, Suspicious, Legitimate) as devised by the rules. The attributes that we extracted are as follows.

1. Have IP

Instead of using domain name, if IP address is used as an alternative in the URL, such as "http://132.58.5.56/fake.html", then it is obvious that Phishers are trying to steal their personal information. In some of the cases Hexadecimal converted IP address is also used as shown in the following link "http://0x4C.0xC5.0xBE.0x13/1/paypal.cn/index.html".

$$\underline{Rule}: \text{IF}\begin{cases}\text{If The Domain Part has an IP Address} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate}\end{cases}$$

2. URL Length

Mostly lengthy URLs can be used by Phishers in order to hide the doubt about the URLs. For example:

http://extrodmacdefloec.com.my/jh/fde/selected_item_id=100014998572120/?cmd=_home&display=764110g43f2s09li32c8r7q2lc3d5877e9551004t5kf5b7dw6523vv1m3sd5e8@phishing.fbook.html

To overcome this issue and to ensure the accurateness of URL length, average length of URLs used in the dataset are less than 54 characters. Researches have shown that if URLs length ranges greater than 54 characters then they must be phishing URLs. In the used dataset about 1220 URLs have lengths equals to 54 characters or more, coming out to be 48.8% of the total dataset size. So, a rule can be defined as:

$$Rule: IF \begin{cases} URL\ length < 54 \rightarrow feature = \text{Legitimate} \\ else\ if\ URL\ length \geq 54\ and\ \leq 75 \rightarrow feature = Suspicious \\ otherwise \rightarrow feature = \text{Phishing} \end{cases}$$

3. Having At(@) Symbol

Whenever "@" symbol is used in the URLs, everything coming ahead of it is ignored by the browsers and the tangible address frequently follows the "@" symbol. A rule for this situation comes to be:

$$\text{Rule: IF} \begin{cases} \text{Url Having @ Symbol} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$$

4. Double Back Slash (//)

Whenever there is double back slash"//" in the URLs, it is ultimate that the user will be redirected to another link or website. For example: URLs like: "http://www.legitimate.com//http://www.phishing.com". To distinguish legitimate links to phishing links it is suggested to examine the location of "//" where it appears. "//" should appear in the sixth position if the URL is starting with "HTTP". Similarly for "HTTPS" double back slash "//" should appear in seventh position.

$$\text{Rule: IF} \begin{cases} \text{ThePosition of the Last Occurrence of "//" in the URL} > 7 \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$$

5. Prefix Suffix (-)

It is rarely in practice to use dash symbol in legitimate URLs. Whenever it is observed prefix or suffixes separated by (-), it should be dealt as phishing link as Phishers tend to add (-) in the domain name to give a feel of legitimate webpage. For example: http://www.logine-paypal.com/.

$$\text{Rule: IF} \begin{cases} \text{Domain Name Part Includes } (-) \text{ Symbol} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate} \end{cases}$$

6. Having Sub Domain (.)

Let's assume a following link: http://www.puc.edu.cn/students/. A domain name might include the country-code top-level domains (ccTLD), which in our example is "cn". The "edu" part is shorthand for "education", the combined "edu.cn" is called a second-level domain (SLD) and "puc" is the actual name of the domain. For such type of links, the rules are generated by extracting features by firstly omitting the (www.) from the URL which represents a sub domain. Now, (ccTLD) is removed if it exists. Finally, counting the remaining dots will decide about the legitimacy of the URLs. If dots are greater than one, then ultimately URL is classified as "Suspicious". Though, if the dots are greater than two, it is classified as "Phishing" as it shows that it will have multiple sub domains. Finally if there is no subdomain then it will be assigned "Legitimate" to the feature.

$$\text{Rule: IF} \begin{cases} \text{Dots In Domain Part} = 1 \rightarrow \text{Legitimate} \\ \text{Dots In Domain Part} = 2 \rightarrow \text{Suspicious} \\ \text{Otherwise} \rightarrow \text{Phishing} \end{cases}$$

7. SSL State (HTTP/HTTPS)

For a website to be legitimate HTTPS existence plays a vital role in it but it is not enough to give impression of legitimacy. The authors in [50] [51] have given suggestions of checking certificates assigned with HTTPS containing the level of the conviction license issuer, and the certificate age [52]. Certificate Authorities that are consistently listed among the top trustworthy names include: "GeoTrust, GoDaddy, Network Solutions, Thawte, Comodo, Doster and VeriSign". Moreover, in the used datasets, minimum age of a reputable certificate is two years.

$$\text{Rule: IF} \begin{cases} \text{Use https and Issuer Is Trusted and Age of Certificate} \geq 1 \text{ Years} \rightarrow \text{Legitimate} \\ \text{Using https and Issuer Is Not Trusted} \rightarrow \text{Suspicious} \\ \text{Otherwise} \rightarrow \text{Phishing} \end{cases}$$

8. "HTTPS Token" in the Domain Part of the URL

To dodge the users, Phishers may use HTTPS tokens in the domain parts of the URL which doesn't give the feeling of Phishing links. For example; http://https.www.epaypal.mpp.home.payhair.com.

$$\text{Rule: IF}\begin{cases}\text{Using HTTP Token in Domain Part of The URL} \rightarrow \text{Phishing} \\ \text{Otherwise} \rightarrow \text{Legitimate}\end{cases}$$

9. Class

This is the target attribute that shows if a particular instance is phishing or legitimate. For Phished its value is -1 and for non-phished means legitimate its value is +1.

After extracting the features from both phished & legitimate URLs, we selected 1000 records of each type and combined them into a single table to create a full dataset of 2000 records. We then shuffled the rows as to balance the data to be equally distributed.

From the dataset of 2000 records we divided it into two parts, 1800(1004 phished and 796 legitimate) records for training-testing with 10-fold cross validation and 200(111 phished and 89 legitimate) records as evaluation set. We applied several feature extraction techniques but we found that removing any of the features reduces the accuracy. We applied Correlation Based Feature Selection, Chi-Square and Gain Ratio Based Feature selection approaches using WEKA tool.

4.1.2. Rules Generation

To generate rules we applied three different rule generation machine learning approaches. C4.5, PART and RIPPER were applied on the dataset that we created in the previous section. Weka tool was used to implement these rule generation methods, C4,5 in implemented in WEKA by the name J48 and RIPPER by the name JRip. Results obtained by C4.5, RIPPER and PART with default settings are as follows:

4.1.2.1 C4.5

Size of the tree: 6

Number of Leaves: 4

- If (ssl_state = -1) Then -1 (phished)
- If (ssl_state = 0) Then -1 (phished)
- If (ssl_state = 1)
    - | prefix_suffix = -1 Then -1 (phished)
    - | prefix_suffix = 1 Then 1 (legit)

Correctly Classified Instances = 1685

Incorrectly Classified Instances = 115

Accuracy = 93.6111 %

Error = 6.3889 %

Class	TP Rate	FP Rate	Precision	Recall	F-Measure
-1	0.957	0.090	0.930	0.957	0.944
1	0.910	0.043	0.944	0.910	0.926

Table 4. 1: Evaluations measure of C4.5

	P(-1)	N(1)
T(-1)	961	43
F(1)	72	724

Table 4. 2: Confusion Matrix of C4.5

4.1.2.2 JRip

Number of Rules: 4

JRIP rules:

- (ssl_state = 1) and (prefix_suffix = 1) and (having_sub_domain = 1) => class = 1 (legit)
- (ssl_state = 1) and (prefix_suffix = 1) and (having_sub_domain = 0) => class = 1 (legit)
- (ssl_state = 1) and (prefix_suffix = 1) and (url_length = -1) => class = 1 (legit)
- else => class = -1 (phishing)

Correctly Classified Instances = 1686

Incorrectly Classified Instances = 114

Accuracy = 93.6667 %

Error = 6.3333 %

Class	TP Rate	FP Rate	Precision	Recall	F-Measure
-1	0.958	0.090	0.930	0.958	0.944
1	0.910	0.042	0.945	0.910	0.927

Table 4. 3: Evaluations measure of RIPPER

	P(-1)	N(1)
T(-1)	962	42
F(1)	72	724

Table 4. 4: Confusion Matrix of RIPPER

4.1.2.3 PART

Number of Rules: 7

- ssl_state = -1: -1 (phishing)
- prefix_suffix = 1 AND ssl_state = 1 AND having_sub_domain = 1: 1 (legit)
- prefix_suffix = 1 AND ssl_state = 1 AND having_sub_domain = 0: 1 (legit)
- prefix_suffix = -1: -1 (phishing)
- ssl_state = 1 AND url_length = -1: 1 (legit)
- https_token = 1: -1 (phishing)
- else : 1 (legit)

Correctly Classified Instances = 1679

Incorrectly Classified Instances = 121

Accuracy = 92.3333 %

Error = 7.6667 %

Class	TP Rate	FP Rate	Precision	Recall	F-Measure
-1	0.956	0.095	0.927	0.956	0.941
1	0.905	0.044	0.942	0.905	0.923

Table 4. 5: Evaluations measure of PART

	P(-1)	N(1)
T(-1)	960	44
F(1)	76	720

Table 4. 6: Confusion Matrix of PART

4.1.3. Comparison of C4.5, RIPPER and PART

A comparison based on different evaluation measures are provided in the table 4.4. According to the results (accuracy, error, F-measure, precision, recall) shown, RIPPER outperformed the other two approaches in all evaluation measures.

Class	Number of Rules	Accuracy	Error	Precision	Recall	F-Measure
C4.5	4	93.61 %	6.39 %	0.930	0.957	0.944
Ripper	4	93.67 %	6.33 %	0.930	0.958	0.944
PART	7	92.33 %	7.67 %	0.927	0.956	0.941

Table 4. 7: Comparison of C4.5, RIPPER and PART

So rules of RIPPER are converted to if-then structure & passed to the part-2 of the model.

Final RIPPER rules are as follows:

1. **IF** ((ssl_state = 1) and (prefix_suffix = 1) and (having_sub_domain = 1))
    - **THEN** class = 1 (legit)
2. **IF** ((ssl_state = 1) and (prefix_suffix = 1) and (having_sub_domain = 0))
    - **THEN** class = 1 (legit)
3. **IF** ((ssl_state = 1) and (prefix_suffix = 1) and (url_length = -1))
    - **THEN** class = 1 (legit)
4. **ELSE** class = -1 (phishing)

4.2. PART-2:

We start by upload the list of the valid phish list, in the csv format on to the database table only containing the URLs, which is obtained from the PhishTank. These URLs are used to

compare with the new URLs, as if the new URL is a phished one or not, which is the normal process of blacklist-based approach.

4.2.1. URL and Host Checking

Our approach works ahead of the simple blacklist based approach, what mostly happens is that the attacker just changes the path, query part or both without changing the host which is mostly not catered in the blacklist based approach, and keeping in view this gap we extract the host of each URL in the blacklist and save it in the database alongside the URL. Then we compare the new URL not only with the URLs in the blacklist but also the host of the new URL with the hosts saved in the database, if the match is found it's declared as phishing else we go to next step of feature extraction.

url	host_domain	url_hash	host_domain_hash
http://signin.amazon.co.uk-prime.form-unsuscribe.i...	signin.amazon.co.uk-prime.form-unsuscribe.id-5564....	c10701899af3c1a26895fa6a8a99bd31	32c2ad1ec47b197a07327513c96e701e
http://allnotifyarea20.ml	allnotifyarea20.ml	ff57d9afe14ad3e8bef0b745fd73eddf	3ae080059652362727837 8c4b87a8c3a
https://inexcusable-storms.000webhostapp.com/verif...	inexcusable-storms.000webhostapp.com	bd36c55c08dd74ba9479e256367561c1	e2255c3e94e8685f8b4b88adcb2d010a
http://jasatradingsa.com/remoteeee/GD/	jasatradingsa.com	5acc4ba9763d1af908f2d81bfa34d192	d62a7ba433fa0b179e0a88f33886e488
http://us.battle.net.login.login.xml.account.suppo...	us.battle.net.login.login.xml.account.support.html...	05d4ff309023067ee9bb72f6dee4e9c6	e0f5f1e159fbd17e7b0306605e78286a
http://integratedpreservations.us/wp-admin/goo/bid...	integratedpreservations.us	8aaf0e15aa469f81b872fb8cf30dca19	4a426f2a48de352b251bfd883d0df83a
http://office-docsign.eyebani.com/doccszxwesr/	office-docsign.eyebani.com	9100eaaf392bcd5ec3ffa169b8666e5d	2e528cc45c14467207d08c6d8874318b
https://offidocx.com/3dfcb45c68ff3457029d34af9dc03...	offidocx.com	f470b93a824493785621facc9a107f67	340aee0cef1b5a587be4b2aad91027c4

Figure 4. 1: URL, Host of URL, Hash or URL and its host

4.2.2. Hashing for fast processing

The phishing URLs are in string format and are very long than the ordinary normal legitimate URLs, so comparing new URL with thousands of URLs or their hosts can take a lot of time. To save the processing time we have used MD5 function for hashing the URL and its host as shown in the figure 4.1 above.

4.2.3. URL Feature Extraction and Decision System

The URLs whose match is not found in the blacklist, now we extract the same features from the URL that we extracted in part first and pass these feature values to the decision system containing the rules generated from the first part as can be seen in figure 4.2. Here in the decision system based on the feature values and the rules it is declared if the URL is legitimate or phishing.

id	url	have_ip	url_length	having_at_symbol	double_slash	prefix_suffix	having_sub_domain	https_token	ssl_state	class
2	https://www.etsy.com/c/craft-supplies-and-tools/ho...	1	-1	1	1	1	-1	1	1	1
3	http://acessoclientestilo.com.br/brasil.php?lider/...	1	-1	1	1	1	0	1	-1	-1
4	http://www.recant-odoamanhecer.com.br/sym/excel/in...	1	-1	1	1	-1	1	1	-1	-1
5	http://www.network-2017.esy.es/recovery.php	1	1	1	1	-1	-1	1	1	-1
6	http://ietf.org/about/standards-process.html	1	-1	-1	1	1	1	1	1	1
7	http://mashable.com/entertainment/?utm_cid=mash-pr...	1	-1	1	1	1	1	1	1	1
8	http://video.foxnews.com/v/5552604087001/?#sp=show...	1	-1	1	1	1	0	1	-1	1
9	http://www.recantodoamanhecer.com.br/sym/excel/ind...	1	-1	1	1	1	1	1	0	-1
10	http://https-securedirsserver.verifyrecipientemail...	1	-1	1	1	-1	0	1	-1	-1
11	https://www.buzzfeed.com/?utm_term=.gtk40EDEvM#.my...	1	-1	1	1	1	1	1	1	1
12	http://news.baidu.com/advanced_news.html	1	-1	1	1	1	1	1	1	1
13	https://secures1-paypalsignin.com.enscrowlookup.co...	1	-1	1	1	-1	1	1	-1	-1
14	http://www.noaa.gov/stories/hurricane-andrew-what-...	1	1	1	1	1	1	1	1	1

Figure 4. 2: Feature Extracted from the new URLs.

4.2.4. Updating Database

If the decision system declares the new URL as legitimate then the program execution exits without any further processing, else if it is marked as phished then the hash of the new URL and its host are saved in DB, so that if the same URL or its host repeats itself it can be checked at the first stage thus saving the processing time of feature extraction and checking it through the decision system

	P(-1)	N(1)
T(-1)	110	01
F(1)	06	83

Table 4. 8: Evaluation dataset Confusion Matrix

Accuracy = 96.5%

Error = 3.5%

Precision = 99.09%

Recall = 94.82 %

F-Measure = 2. (0.939571)/(1.9391) = 0.969079 = 96.91%

2.11 Comparison of different techniques

In the following table comparison of different techniques have been given.

Methods	Precision	Recall	F1-Measure
2012 Belabed et al	**96.6%**	**98.0%**	**0.973**
2014 Corbetta et al	**95.3%**	**73.08%**	**0.827**
2016 Z. Hu et al	**91.50%**	**0.9440**	**0.9300**
2017 Proposed Hybrid Approach	**99.09%**	**94.82 %**	**0.9691**

Table 4. 6: Comparison with Proposed & Previous Algorithms

CHAPTER 5: CONCLUSION

Phishing is one of most common online scam used to lure the user to provide his/her personal information that can be social security identification number, bank account information or online account used for purchasing and selling of goods. In the second chapter we have discussed several recent approaches introduced to stop the attacks due to phishing. Various methods have extracted features from URL and checked for the anomalies to declare the page as phished or not. Where some have extracted feature from the web page content and relevant features are collected that use similarity index measure to identify the phishing. Online domain features like rank of website in Google, Alexa and similar others have also been implemented in different approaches in the detection of phish scam. Combination of different weak & strong supervised classifiers have been implemented in the form of hybrid approaches that yields better results like in one case J48+IBk & BN+IBk outperforms other pairings. Many algorithms of association classification have also been exercised; one of the recent ones is discovering frequent item set using *Diffset* which is a vertical mining approach. All these techniques performs good in their respective environments but each one is lacking or in other words restricting the perfect accuracy one way or the other i.e some approaches focuses on the URL features while others on visual similarity. We tried to introduce an approach that results in computationally optimized performance that is neither dependent on the similarity of the visual attributes nor rely on the online popularity of the site domain.

The solution proposed in this thesis has two separate sections, in the first section/part we have created our own dataset from the URLs collected from PhishTank.com and Majestic.com. We extracted 8 features from the ULRs in the database and store these attributes along with the URLs in DB, then we employed 3 rule generation based algorithms and finally picked the rules generated by the model with high accuracy which concludes our first part of the anticipated model and these rules are input to the second part of the model.

Second part of the model starts with the hash generation of the complete URL as well as the host part in the URL separately that are too stored in the DB, so as we receive a new URL that is to be checked its hash is initiated and matched to the ones in database, if match not found then the host hash is generated and check to find if it's a phish or not. If hash doesn't define then we extract the same features that we extracted in the first part and here comes the rules from part first to evaluate the features of URL based on these rules to judge the scam URL from legitimate one. If phishing the website is reported to the phishtank and hash of the site along with its host is saved in DB, in this way our DB get more and more populated

reducing the extra computational cost that is mostly part of all models. Our proposed model out performs in terms accuracy of 99.09% where recall and F1-measure near to the previous algorithms.

Future Work: We can increase the performance more if we also cater the attributes of online popularity and creating a new better rule based approach to generate rules.

References

[1] Junaid Ahsenali Chaudhry, Shafique Ahmad Chaudhry and Robert G. Rittenhouse, "Phishing Attacks and Defenses," *International Journal of Security and Its Applications,* vol. 10, no. 1, pp. 247-256, 2016.

[2] http://www.phishing.org/.

[3] Qian Cui, Guy-Vincent Jourdan, Gregor V. Bochmann, Russell Couturier and Iosif-Viorel Onut, "Tracking Phishing Attacks Over Time," in *International World Wide Web Conference Committee (IW3C2)*, Perth, 2017.

[4] M. J. a. S. Myers, "Phishing and Countermeasures: Understanding the Increasing Problem of Electronic Identity Theft, Wiley-Interscience, 2006".

[5] "Internet Technologies Workshop: Tel-Aviv University, "Current Anti Phishing Methods","http://tau-itw.wikidot.com/deleted:saphe-current-anti-phishing-methods", 2009".

[6] Gartner, "Number of Phishing Attacks on U.S. Consumers Increased 40 Percent in 2008, http://www.gartner.com/it/page.jsp?id=936913"," 2008.

[7] L. P. E.-S. planet, "Top ten phishing facts, "http://www.esecurityplanet.com/views/article.php/3875866/Top-Ten-Phishing-Facts.htm"," 2010.

[8] Z. Ramzan, "Phishing attacks and countermeasures," in *In Stamp, Mark & Stavroulakis, Peter. Handbook of Information and Communication Security. Springer. ISBN 9783642041174.*, 2010.

[9] A. J. L. M. D. M. Van der Merwe, "Characteristics and Responsibilities involved in a Phishing Attack," in *Winter International Symposium on Information and Communication Technologies*, Cape Town, 2005.

[10] 2014, "20% Indians are victims of Online phishing attacks," in *Microsoft IANS.news.biharprabha*, 2014.

[11] A. Jøsang and e. al., "Security Usability Principles for Vulnerability Analysis and Risk Assessment," in *Proceedings of the Annual Computer Security Applications Conference 2007 (ACSAC'07)*, 2007.

[12] J. T. L. L. H. Huang, "Countermeasure Techniques for Deceptive Phishing Attack","

International Conference on New Trends in Information and Service Science, pp. 636-641, 2009.

[13] 2. "What is spear phishing?". Microsoft Security At Home. Retrieved June 11.

[14] D. ". P. W. G. C. F. R. J. 2. 2. Stephenson.

[15] 1. c. f. c. T. R. A. f. t. o. o. J. 3. 2. R. A. 1. 2. Fake subpoenas harpoon 2.

[16] 2. R. M. 2. 2. What Is 'Whaling'? Is Whaling Like 'Spear Phishing'?. About Tech. Archived from the original on October 18.

[17] 2. R. D. 1. 2. Get smart on Phishing! Learn to read links!. Archived from the original on December 11.

[18] C. Cimpanu, "Hidden JavaScript Redirect Makes Phishing Pages Harder to Detect," in *Softpedia News Center. Softpedia*, 2016.

[19] P. ". s. t. m. p. s. u. b. c. f. N. A. f. t. o. o. J. 3. 2. Mutton.

[20] "The use of Optical Character Recognition OCR software in spam filtering powerpoint ppt presentation".

[21] P. Mutton, "Phishing Web Site Methods," in *FraudWatch International. Archived from the original on January 31, 2011*.

[22] "Phishing con hijacks browser bar," in *BBC News. April 8, 2004.*.

[23] B. Krebs, "Flaws in Financial Sites Aid Scammers, Security Fix. Archived from the original on January 31, 2011. Retrieved June 28, 2006".

[24] P. Mutton, "PayPal Security Flaw allows Identity Theft, Netcraft. Archived from the original on January 31, 2011. Retrieved June 19, 2006.".

[25] R. Miller, "Phishing Attacks Continue to Grow in Sophistication, Netcraft. Archived from the original on January 31, 2011. Retrieved December 19, 2007".

[26] O. d. C. M. 2. 2. R. N. 1. 2. Serious security flaw in OAuth.

[27] 2. R. N. 1. 2. Covert Redirect Vulnerability Related to OAuth 2.0 and OpenID". Tetraph. May 1.

[28] G. u. t. b. n. s. f. F. N. M. 5. 2. R. N. 1. 2. Facebook.

[29] "Nasty Covert Redirect Vulnerability found in OAuth and OpenID. The Hacker News. May 3, 2014," 2014.

[30] "Facebook, Google Users Threatened by New Security Flaw. Yahoo. May 2, 2014.,"

2014.

[31] "Covert Redirect' vulnerability impacts OAuth 2.0, OpenID," in *SC Magazine. May 2, 2014*, 2014.

[32] "Covert Redirect Flaw in OAuth is Not the Next Heartbleed," in *Symantec. May 3, 2014*, 2014.

[33] "Graham, Meg (19 January 2017). This Gmail phishing attack is tricking experts. Here's how to avoid it," 2017.

[34] K. Tomlinson, "Fake news can poison your computer as well as your mind," 2017.

[35] A. Gonsalves, "Phishers Snare Victims With VoIP," in *Techweb*, 2006.

[36] "Identity thieves take advantage of VoIP," in *Silicon.com. March 21, 2005,* 2005.

[37] "Phishing, Smishing, and Vishing: What's the Difference?," 2008.

[38] "Internet Banking Targeted Phishing Attack," in *Metropolitan Police Service. June 3, 2005*, 2009.

[39] "https://www.antiphishing.org/resources/apwg-reports/," 2016.

[40] M. M. a. P. K. A. Gaurav Varshney, "A survey and classification of web phishing detection schemes," in *Security Comm. Networks 2016, Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/sec.1674*, 2016.

[41] T. W. L. P. W. T. L. Z. Mao J., "Phishing Website Detection Based On Effective CSS Features Of Web Pages," in *International Conference On Wireless Algorithms, Systems, And Application, pp 804-815, Springer, May 2017*, 2017.

[42] A. Y. V. Mahmood Moghimi, "New Rule-based Phishing Detection Method," *Expert Systems With Applications: An International Journal*, vol. 53, no. July 2016, pp. 231-242, 2016.

[43] M. A. U. H. Tahir , . Asghar , . Zafar and . Gillani, "A Hybrid Model To Detect Phishing-sites Using Supervised Learning Algorithms," in *International Conference On Computational Science And Computational Intelligence (CSCI), IEEE Conference*, 2016.

[44] Z. Hu, . Chiong, . Pranata, . Susilo and . Bao, "Identifying Malicious Web Domains Using Machine Learning Techniques With Online Credibility And Performance Data," *IEEE Congress On Evolutionary Computation (CEC),* pp. 5186 - 5194, 2016.

[45] N. J. A. M. Pradeep Singh, "Investigating The Effect Of Feature Selection And

Dimensionality Reduction On Phishing Website Classification Problem," in *Pages: 388 - 393, IEEE 2015 1st International Conference on Next Generation Computing Technologies (NGCT)*, 2015 .

[46] B. L. T. H. K. N. M. H. N. Luong Anh Tuan Nguyen, "An Efficient Approach For Phishing Detection Using Single-layer Neural Network," in *Pages: 435 - 440, IEEE 2014 International Conference on Advanced Technologies for Communications (ATC)*, 2014 .

[47] B. D. D. K. S. D. R. G. C. S. B. M. Rahul Patil, "A Hybrid Model To Detect Phishing-sites Using Clustering And Bayesian Approach," in *IEEE , Pages: 1 - 5, International Conference for Convergence of Technology (I2CT)*, 2014.

[48] A. Q. Ram B. Basnet, "Feature Selection For Improved Phishing Detection," in *pp 252-261, Springer, Conference on Advanced Research In Applied Artificial Intelligence*, 2012.

[49] R. M. Mohammad, F. Thabtah and L. Mccluske, "An Assessment Of Features Related To Phishing Websites Using An Automated Technique," in *pages: 492 - 497, IEEE International Conference For Internet Technology And Secured Transactions, ,* 2012.

[50] R. M. Mohammad, F. Thabtah and L. McCluskey, "An Assessment of Features Related to Phishing Websites using an Automated Technique," in *The 7th International Conference for Internet Technology and Secured Transactions (ICITST-2012)*, London, 2012.

[51] R. M. Mohammad, F. Thabtah and L. McCluskey, "Intelligent Rule based Phishing Websites Classification," *IET Information Security*, vol. 7, no. 3, July 2013.

[52] "https://www.sslshopper.com/ssl-checker.html".

[53] https://majestic.com/support/glossary#AlexaRank.

[54] a. shoukat, "abc," 2017.

[55] Khuong Nguyen, Minh Hoang Nguyen, "An Efficient Approach For Phishing Detection Using Single-layer Neural Network," in *ieee*, 2014.

[56] p. a. c. c. d. (. i. m. o. f. m. r. b. d. a. a. t. e. i. a. e. c. T. w. i. a. Phishing is the attempt to obtain sensitive information such as usernames.

[57] A. J. L. M. D. M. (. C. a. R. i. i. a. P. A. W. I. S. o. I. a. C. T. C. T. J. 2. Van der Merwe.

[58] F. A. S. A. Wa'el Hadi, "A new fast associative classification algorithm for detecting

phishing websites," *Elsevier 2016 Journal of Applied Soft Computing*, 2016 .