

DETECTING PHISHING WEBSITES USING MACHINE LEARNING

Submitted in partial fulfillment of the requirements
for the award of
Bachelor of Engineering degree in Computer Science and Engineering

by

Patibanda Pranav Naga Raja Prakash (38110419)
Akula Abinay (38110009)



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SCHOOL OF COMPUTING**

SATHYABAMA

**INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)**

Accredited with Grade "A" by NAAC

**JEPPIAAR NAGAR, RAJIV GANDHI SALAI,
CHENNAI - 600 119**

MARCH – 2022



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with "A" grade by NAAC

Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai - 600119

www.sathyabama.ac.in



DEPARTMENT OF COMPUTER SCIENCE ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **Patibanda Pranav Naga Raja Prakash (38110419)** and **Akula Abinay (38110009)** who carried out the project entitled "**DETECTING PHISHING WEBSITES USING MACHINE LEARNING**" under my supervision from November 2021 to March 2022.

Internal Guide

Ms.Dr.T. JUDGI, M.E., Ph. D.,

Head of the Department

Submitted for Viva voce Examination held on _____

Internal Examiner

External Examiner

DECLARATION

I **Patibanda Pranav Naga Raja Prakash (38110419)** and **Akula Abinay (38110009)** hereby declare that the Project Report entitled "**DETECTING PHISHING WEBSITES USING MACHINE LEARNING**" done by us under the guidance of **Ms.Dr.T. JUDGI, M.E., Ph. D** is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering degree in 2018-2022.

DATE:

PLACE: CHENNAI

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management of SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T.Sasikala M.E.,Ph.D., Dean**, School of Computing **Dr.S.Vigneshwari M.E., Ph.D.** and **Dr.L.Lakshmanan M.E., Ph.D.** , Heads of the Department of Computer Science and Engineering for providing us necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and a deep sense of gratitude to my Project Guide **Ms.Dr.T. JUDGI, M.E., Ph. D.**, for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

ABSTRACT

Phishing sites which expects to take the victims confidential data by diverting them to surf a fake website page that resembles a honest to goodness one is another type of criminal acts through the internet and its one of the especially concerns toward numerous areas including e-managing an account and retailing. Phishing site detection is truly an unpredictable and element issue including numerous components and criteria that are not stable. On account of the last and in addition ambiguities in arranging sites because of the intelligent procedures programmers are utilizing, some keen proactive strategies can be helpful and powerful tools can be utilized, for example, fuzzy, neural system and data mining methods can be a successful mechanism in distinguishing phishing sites. We applied Random Forest (RF), one of the different types of machine learning based algorithms used for detection of Phishing websites. Finally we measured and compared the performance of the classifier in terms of accuracy.

TABLE OF CONTENTS

Chapter No.	TITLE	Page No.
	ABSTRACT	v
	LIST OF FIGURES	vii
1	INTRODUCTION	1
	1.1. OUTLINE OF THE PROJECT	1
	1.2 MACHINE LEARNING	2
	1.3 SCOPE AND OBJECTIVE	
	1.2	2
2	LITERATURE SURVEY	3
	2.1. RELATED WORK	3
3	METHODOLOGY	4
	3.1. EXISTING SYSTEM	4
	3.2. PROPOSED SYSTEM	8
		5
	3.3 ALGORITHMS USED	
		5

3.3.1 NAIVE BAYES ALGORITHM

3.3.1. DECISION TREE CLASSIFIER

5

3.3.2 . RANDOM FOREST

6

3.4 SYSTEM ARCHITECTURE

7

3.5 SYSTEM REQUIREMENTS

7

3.6 MODULES

8

3.7 UML DIAGRAMS

18

3.8 LANGUAGES USED

20

3.4.

3.9 REQUIREMENT ANALYSIS

24

4	RESULTS AND DISCUSSION	27
	4.1. WORKING	27
5	CONCLUSION	32
	5.1. CONCLUSION	32
		32
	5.2. REFERENCES	32
	5.3. APPENDIX	33
	A.SOURCE CODE	33
	B.SCREENSHOTS	36
	C.PLAGIARISM REPORT	39
	D.JOURNAL PAPER	40

LIST OF FIGURES

Figure No.	Figure Name	Page No.
	System Architecture	
3.4		7
3.7.1	Data Flow Diagram	20
B.1	Input Page	36
B.2	Output Page	37
B.3	Output Page	37
B.4	Accuracy Graph	38

CHAPTER 1

INTRODUCTION

1.1 OUTLINE OF THE PROJECT

Phishing is a type of extensive fraud that happens when a malicious website act like a real one keeping in mind that the end goal to obtain touchy data, for example, passwords, account points of interest, or MasterCard numbers.

In spite of the fact that there are a few contrary to phishing programming and methods for distinguishing potential phishing endeavours in messages and identifying phishing substance on sites, phishes think of new and half breed strategies to go around the accessible programming and systems.

Phishing is a trickery system that uses a blend of social designing what's more, innovation to assemble delicate and individual data, for example, passwords and charge card subtle elements by taking on the appearance of a dependable individual or business in an electronic correspondence. Phishing makes utilization of spoof messages that are made to look valid and implied to be originating from honest to goodness sources like money related foundations, ecommerce destinations and so forth, to draw clients to visit fake sites through joins gave in the phishing email. The misleading sites are intended to emulate the look of a genuine organization site page.

The employing so as to phishing invader's trap clients diverse social building strategies, for example, debilitating to suspend client accounts on the off chance that they don't finish the account upgrade process, give other data to approve their records or a few different motivations to get the clients to visit their satirize page.

Supervised learning (Classification Technique) accommodates a vastly improved precision while unsupervised learning accommodates a quick and dependable way to deal with infer information from a dataset. That's why we used supervised learning in our work.

1.2 MACHINE LEARNING

Machine learning could be a subfield of computer science (AI). The goal of machine learning typically is to know the structure information of knowledge of information and match that data into models which will be understood and used by folks. Although machine learning could be a field inside technology, it differs from ancient process approaches.

In ancient computing, algorithms are sets of expressly programmed directions employed by computers to calculate or downside solve. Machine learning algorithms instead give computers to coach on knowledge inputs and use applied math analysis so as to output values that fall inside a particular vary. thanks to this, machine learning facilitates computers in building models from sample knowledge so as to modify decision-making processes supported knowledge inputs.

1.3 SCOPE AND OBJECTIVE

The system should be useful in many e-commercial websites for maintaining the security and reliability of customers and people online

The system should be useful in preventing online frauds leading to leakage of important and private user data

The scope of using Machine Language over other Traditional Detecting Methods

Objectives:

- Understanding phishing domain (or Fraudulent Domain) characteristics, its distinguishing features from legitimate domains
- Why it is so important to detect this domain and how they can be detected using machine learning and natural language processing techniques
- Reviewing the state-of-the-art machine learning techniques for malicious URL detection in literature
- Understanding the newly emerging concept of Malicious URL Detection as a service and the principles to be used while designing such a system.

To distinguish the phishing websites from the legitimate websites and ensure secure transactions to users

CHAPTER 2

LITERATURE SURVEY

2.1 RELATED WORK

(Mao, J., Bian, J., Tian, W., Zhu, S., Wei, T., Li, A., & Liang, Z. 2018) [1] In this article, we intend to further develop understanding abilities through AI. Specifically, it is suggested that the exploration technique be founded on the page design choice strategy utilized for page search. The consequences of the review show that our techniques are exact and helpful in deciding the phishing sheet.

(Atharva Deshpande , Omkar Pedamkar , Nachiket Chaudhary , Dr. Swapna Borde/ 2021) [2] This page analyzes the apparatuses used to learn and get machines. Phishing is known for its gatecrashers since duping somebody is more straightforward to hit on a terrible line than beating a safeguard framework. The negative connections in the principle body of the message are expected to show that these corporate images and other real items are utilized to arrive at degenerate associations.

(Ishant Tyagi; Jatin Shad; Shubham Sharma; Siddharth Gaur; Gagandeep Kaur/ 2018) [3] This page centers around different AI calculations pointed toward foreseeing whether a site is misled or real. Machine preparing is famous on the grounds that it can distinguish party time assaults and is great at beating new kinds of phishing assaults. In our work, we had the option to precisely decide 98.4% by anticipating phishing or lawful area.

(Vahid Shahrivari, Mohammad Mahdi Darabi, Mohammad Izadi/ 2020) [4] Understanding Phishing Using Machine Learning Techniques Wahid Shahrivari, Mohammad Mahdi Darabi, Mohammad Izadi/2020. The best method for recognizing these awful encounters is through AI. This is on the grounds that numerous phishing assaults are the most widely recognized types of AI. In this article, we think about the aftereffects of many AI strategies to foresee phishing.

(Mohith Gowda HR, Adithya MV, Gunesh Prasad S & Vinay S/ 2020) [5] In this article, we need specialized ability to effortlessly recognize a phishing site on the client side that requirements to assemble a web crawler. In this framework, we utilize the erase rule to eliminate content or site highlights utilizing just the URL. The

rundown comprises of 30 unique URLs and will then, at that point, be utilized to discover reality with regards to the site by irregular words arrangement.

(Banu, R., Anand, M., Kamath, A., Ashika, S., Ujwala, H. S., & Harshitha, S. N. 2019) [6] The approach will also use Deep Learning frameworks with hierarchical long-short term memory networks (H-LSTMs) and attention mechanisms to model the emails simultaneously at the word and sentence level. Phishing attacks categorizes the emails based on certain properties which give more details about the source of phishing. Generally, most of the existing systems focus on email classification depending upon header part or body part.

(Karabatak, M., & Mustafa, T. 2018) [7] This article inspects information assortment on the UCI site. Diminishing its size and looking at the presentation of positioning calculations is being contemplated in the news site of the phishing site. The portrayal of the phishing site is taken from the UCI information base of AI. The data set comprises of 11055 passages and 31 exercises. The presentation of the arranging calculation is currently contrasted with other data on the grouping calculations. At long last, contrasting the requesting elements of the informational indexes utilizing the overall calculations gave.

(Shima, K., Miyamoto, D., Abe, H., Ishihara, T., Okada, K., Sekiya, Y., ... & Doi, Y. 2018) [8] In this review, we tested with realistic URL access history data taken from a research organization and data from the famous archive site of phishing site information, PhishTank.com. Our approach achieved 2~3% better accuracy compared to the existing DL- based approach.

CHAPTER 3

METHODOLOGY

3.1 EXISTING SYSTEM

There are innumerable spaces that can be defrauded, like web-based installments, webmail, monetary foundations, document stockpiling or distributed storage. Web and online installments are remembered for the rundown of best practices. Since phishing should be possible by email or lance phishing , the client ought to know about the effect and not be 100% certain about the general security activity. Machine preparing

is probably the most ideal way to master phishing procedures as it dispenses with the risks of existing strategies.

3.2 PROPOSED SYSTEM

Endeavors to gather individual data deceitfully are turning out to be more normal today. To assist clients with knowing how to access such a site, a framework has been executed that tells clients by means of email and a spring up window when they attempt to get to the site. This page gives a boycott identification framework known as a phishing site with the goal that the site can be told when looking or signing in. Along these lines, it tends to be utilized as a genuine apparatus to distinguish, convince, and forestall misrepresentation.

3.3 ALGORITHMS USED

3.3.1 DECISION TREE CLASSIFIER

Quite possibly the most impressive and famous calculation. The choice tree calculation is beneath the control level of the calculation. It works in two ways, similar to yield. Input:

- At first, we believe all arranged activities to be root.
- Highlights In the data recovery class, capacities are thought of as nonstop.
- Presented on numerous occasions relying upon the idea of the archive.
- We utilize factual strategies to import capacities like root or interior.

A choice tree that makes a class or retreat structure as a tree. Little information is erased as the related tree develops. The choice takes at least two branches, and the leaves show the arrangement or choice. The most noteworthy exactness in the tree grouping compares to the anticipated root. The authentication tree can be utilized both by classification and number. A choice tree that makes a class or retreat structure as a tree. It is designed together and utilizations the on the off chance that standard, which is characterized exhaustively by class. Methods are concentrated successively utilizing concurrent preparation data. However long you gain proficiency

with the law, you can kill the twists of the law. This cycle will proceed until the preparation bundle meets the prerequisites. It depends on the "offer and win" circle through and through. All properties should be characterized. In any case, they need to think that it is first. The idea of the highest points of the trees enormously impacts the arrangement and is known by the possibility of obtaining data. The choice tree is not difficult to plant and can deliver many branches, and can show commotion or strange outer causes.

3.3.2 RANDOM FOREST ALGORITHM

Random forest algorithm can use both for classification and the regression kind of problems. In this article, you are going to learn, how the random forest algorithm works in machine learning for the classification task. In the next coming another article, you can learn about how the random forest algorithm can use for regression.

STEP 1: BEGIN

STEP 2: Randomly select "k" features from total "m" features.

STEP 3: Among the "k" features, calculate the node "d" using the best split point.

STEP 4: Split the node into daughter nodes using the best split.

STEP 5: Repeat 1 to 3 steps until "l" number of nodes has been reached.

STEP 6: Build forest by repeating steps 1 to 4 for "n" number times to create "n" number of trees.

STEP 7: END

3.4 SYSTEM ARCHITECTURE

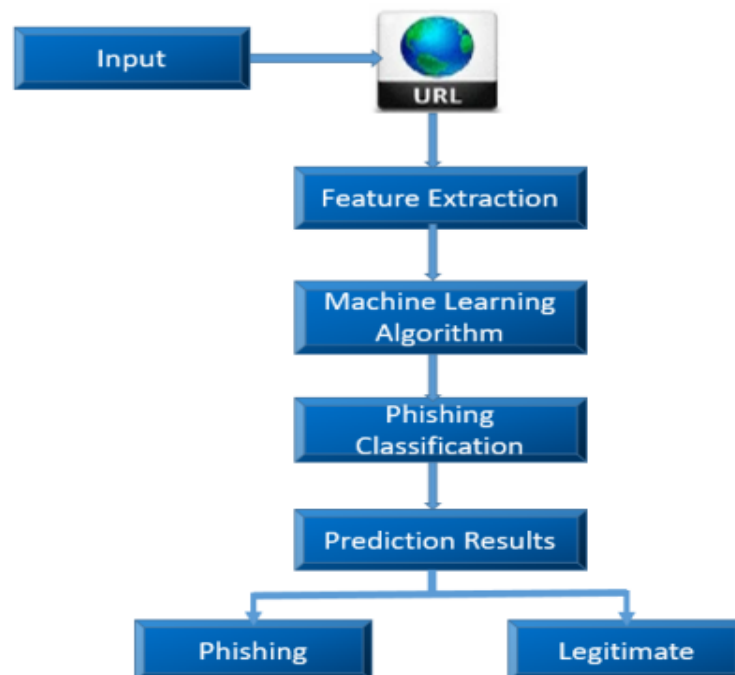


Fig:3.4 :System Architecture

3.5 SYSTEM REQUIREMENTS

Hardware Requirements:

- System : Intel Pentium IV 2.80 GHz.
- Monitor : LED.
- Mouse : Logitech.
- Ram : 4.00 GB or above 4.00 GB
- Hard Disk : 250 GB

Software Requirements:

- Operating system : Windows 7, Ubuntu
- Language : Python 3

3.6 MODULES

- Detection Technique
- Phishing Websites Features
- Data Set

DETECTION TECHNIQUE

- Detection of phishing websites has received a lot of attention recently due to their impact on users' security.
- Therefore, many techniques have been developed to detect phishing websites varying from communication-oriented techniques, such as authentication protocols, blacklisting, and white-listing, to content-based filtering techniques.
- The blacklisting and white-listing techniques have not proven though to be sufficiently efficient when used in different domains, and thus they are not commonly used.
- Meanwhile, the content-based phishing filters have been widely used and have proven to be of high efficiency. In light of this, researches have focused on content-based mechanism and on developing machine learning and data mining techniques based on the header and body of emails.

Address Bar based Features

Using the IP Address

If an IP address is used as an alternative of the domain name in the URL, such as "http://125.98.3.123/fake.html", users can be sure that someone is trying to steal their personal information. Sometimes, the IP address is even transformed into hexadecimal code as shown in the following link "http://0x58.0xCC.0xCA.0x62/2/paypal.ca/index.html".

Rule: IF {If The Domain Part has an IP Address → Phishing Otherwise → Legitimate

Long URL to Hide the Suspicious Part

Phishers can use long URL to hide the doubtful part in the address bar. For example:

http://feder Macedo adv.com.br/3f/aze/ab51e2e319e51502f416dbe46b773a5e/?cmd=_home&dispatch=11004d58f5b74f8dc1e7c2e8dd4105e811004d58f5b74f8dc1e7c2e8dd4105e8@phishing.website.html

To ensure accuracy of our study, we calculated the length of URLs in the dataset and produced an average URL length. The results showed that if the length of the URL is greater than or equal 54 characters then the URL classified as phishing. By reviewing our dataset we were able to find 1220 URLs lengths equals to 54 or more which constitute 48.8% of the total dataset size.

Rule:

IF

{URLlength < 54 → feature = Legitimate else if URLlength ≥ 54 and ≤ 75 → feature = Suspicious}

We have been able to update this feature rule by using a method based on frequency and thus improving upon its accuracy.

Using URL Shortening Services “TinyURL”

URL shortening is a method on the “World Wide Web” in which a URL may be made considerably smaller in length and still lead to the required webpage. This is accomplished by means of an “HTTP Redirect” on a domain name that is short, which links to the webpage that has a long URL. For example, the URL “http://portal.hud.ac.uk/” can be shortened to “bit.ly/19DXSk4”.

Rule: *IF {TinyURL → Phishing Otherwise → Legitimate}*

URL’s having “@” Symbol

Using “@” symbol in the URL leads the browser to ignore everything preceding the “@” symbol and the real address often follows the “@” symbol.

Rule: *IF {Url Having @ Symbol → Phishing Otherwise → Legitimate}*

Redirecting using “//”

The existence of “//” within the URL path means that the user will be redirected to another website. An example of such URL’s is: “http://www.legitimate.com//http://www.phishing.com”. We examine the location where the “//” appears. We find that if the URL starts with “HTTP”, that means the “//”

should appear in the sixth position. However, if the URL employs “HTTPS” then the “/” should appear in seventh position.

Rule:

IF

{ThePosition of the Last Occurrence of "/" in the URL > 7 → Phishing Otherwise → Legitimate

Adding Prefix or Suffix Separated by (-) to the Domain

The dash symbol is rarely used in legitimate URLs. Phishers tend to add prefixes or suffixes separated by (-) to the domain name so that users feel that they are dealing with a legitimate webpage. For example <http://www.Confirme-paypal.com/>.

Rule:IF

{Domain Name Part Includes (–) Symbol → Phishing Otherwise → Legitimate

Sub Domain and Multi Sub Domains

Let us assume we have the following link: <http://www.hud.ac.uk/students/>. A domain name might include the country-code top-level domains (ccTLD), which in our example is “uk”. The “ac” part is shorthand for “academic”, the combined “ac.uk” is called a second-level domain (SLD) and “hud” is the actual name of the domain. To produce a rule for extracting this feature, we firstly have to omit the (www.) from the URL which is in fact a sub domain in itself. Then, we have to remove the (ccTLD) if it exists. Finally, we count the remaining dots. If the number of dots is greater than one, then the URL is classified as “Suspicious” since it has one sub domain. However, if the dots are greater than two, it is classified as “Phishing” since it will have multiple sub domains. Otherwise, if the URL has no sub domains, we will assign “Legitimate” to the feature.

Rule:IF

{Dots In Domain Part = 1 → Legitimate Dots In Domain Part = 2 → Suspicious Otherwise → Phishing

HTTPS (Hyper Text Transfer Protocol with Secure Sockets Layer)

The existence of HTTPS is very important in giving the impression of website legitimacy, but this is clearly not enough. The authors in (Mohammad, Thabtah and McCluskey 2012)(Mohammad, Thabtah and McCluskey 2013) suggest checking the certificate assigned with HTTPS including the extent of the trust certificate issuer, and the certificate age. Certificate Authorities that are consistently listed among the top trustworthy names include: “GeoTrust, GoDaddy, Network Solutions, Thawte,

Comodo, Doster and VeriSign". Furthermore, by testing out our datasets, we find that the minimum age of a reputable certificate is two years.

Rule:IF

{Use https and Issuer Is Trusted and Age of Certificate ≥ 1 Years $\rightarrow$ Legitimate Using https and I

Domain Registration Length

Based on the fact that a phishing website lives for a short period of time, we believe that trustworthy domains are regularly paid for several years in advance. In our dataset, we find that the longest fraudulent domains have been used for one year only.

Rule:IF{Domains Expires on ≤ 1 years $\rightarrow$ Phishing Otherwise $\rightarrow$ Legitimate

Favicon

A favicon is a graphic image (icon) associated with a specific webpage. Many existing user agents such as graphical browsers and newsreaders show favicon as a visual reminder of the website identity in the address bar. If the favicon is loaded from a domain other than that shown in the address bar, then the webpage is likely to be considered a Phishing attempt.

Rule:IF{Favicon Loaded From External Domain $\rightarrow$ Phishing Otherwise $\rightarrow$ Legitimate

Using Non-Standard Port

This feature is useful in validating if a particular service (e.g. HTTP) is up or down on a specific server. In the aim of controlling intrusions, it is much better to merely open ports that you need. Several firewalls, Proxy and Network Address Translation (NAT) servers will, by default, block all or most of the ports and only open the ones selected. If all ports are open, phishers can run almost any service they want and as a result, user information is threatened. The most important ports and their preferred status are shown in Table 2.

Rule:IF{Port # is of the Preferred Status $\rightarrow$ Phishing Otherwise $\rightarrow$ Legitimate

Table 1 Common ports to be checked

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

1.1.1. The Existence of “HTTPS” Token in the Domain Part of the URL

The phishers may add the “HTTPS” token to the domain part of a URL in order to trick users. For example,

<http://https-www-paypal-it-webapps-mpp-home.soft-hair.com/>.

Rule:IF

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

Abnormal Based Features

Request URL

Request URL examines whether the external objects contained within a webpage such as images, videos and sounds are loaded from another domain. In legitimate webpages, the webpage address and most of objects embedded within the webpage are sharing the same domain.

Rule:IF

$\{\% \text{ of Request URL} < 22\% \rightarrow \text{Legitimate } \% \text{ of Request URL} \geq 22\% \text{ and } 61\% \rightarrow \text{Suspicious Otherwise}$

URL of Anchor

An anchor is an element defined by the `<a>` tag. This feature is treated exactly as “Request URL”. However, for this feature we examine:

1. If the `<a>` tags and the website have different domain names. This is similar to request URL feature.
2. If the anchor does not link to any webpage, e.g.:
 - A. `<a href="#">`
 - B. `<a href="#content">`
 - C. `<a href="#skip">`
 - D. `<a href="JavaScript ::void(0)">`

Rule:

IF

$\{\% \text{ of URL Of Anchor} < 31\% \rightarrow \text{Legitimate } \% \text{ of URL Of Anchor} \geq 31\% \text{ And } \leq 67\% \rightarrow \text{Suspicious } \}$

Links in <Meta>, <Script> and <Link> tags

Given that our investigation covers all angles likely to be used in the webpage source code, we find that it is common for legitimate websites to use <Meta> tags to offer metadata about the HTML document; <Script> tags to create a client side script; and <Link> tags to retrieve other web resources. It is expected that these tags are linked to the same domain of the webpage.

Rule:

IF

{% of Links in " < Meta > ", " < Script > " and " < Link > " < 17% → Legitimate % of Links i

Server Form Handler (SFH)

SFHs that contain an empty string or “about:blank” are considered doubtful because an action should be taken upon the submitted information. In addition, if the domain name in SFHs is different from the domain name of the webpage, this reveals that the webpage is suspicious because the submitted information is rarely handled by external domains.

Rule:IF

{SFH is "about: blank" Or Is Empty → Phishing SFH Refers To A Different Domain→ Suspicious O

Submitting Information to Email

Web form allows a user to submit his personal information that is directed to a server for processing. A phisher might redirect the user's information to his personal email. To that end, a server-side script language might be used such as “mail()” function in PHP. One more client-side function that might be used for this purpose is the “mailto:” function.

Rule:IF

{Using "mail()" or "mailto:" Function to Submit User Information→ Phishing Otherwise → Legiti

Abnormal URL

This feature can be extracted from WHOIS database. For a legitimate website, identity is typically part of its URL.

Rule:

IF

{The Host Name Is Not Included In URL → Phishing Otherwise→ Legitimate

HTML and JavaScript based Features

Website Forwarding

The fine line that distinguishes phishing websites from legitimate ones is how many times a website has been redirected. In our dataset, we find that legitimate websites have been redirected one time max. On the other hand, phishing websites containing this feature have been redirected at least 4 times.

Rule:IF

{#ofRedirect Page≤1 → Legitimate #of Redirect Page≥2 And < 4→ Suspicious Otherwise → Phis

Status Bar Customization

Phishers may use JavaScript to show a fake URL in the status bar to users. To extract this feature, we must dig-out the webpage source code, particularly the “onMouseOver” event, and check if it makes any changes on the status bar.

Rule:IF

{onMouseOver Changes Status Bar→ Phishing It Does't Change Status Bar→Legitimate

Disabling Right Click

Phishers use JavaScript to disable the right-click function, so that users cannot view and save the webpage source code. This feature is treated exactly as “Using onMouseOver to hide the Link”. Nonetheless, for this feature, we will search for event “event.button==2” in the webpage source code and check if the right click is disabled.

Rule:IF{*Right Click Disabled → Phishing Otherwise→Legitimate*

Using Pop-up Window

It is unusual to find a legitimate website asking users to submit their personal information through a pop-up window. On the other hand, this feature has been used in some legitimate websites and its main goal is to warn users about fraudulent activities or broadcast a welcome announcement, though no personal information was asked to be filled in through these pop-up windows.

Rule: IF {*Popoup Window Contains Text Fields→ Phishing Otherwise → Legitimate*

IFrame Redirection

IFrame is an HTML tag used to display an additional webpage into one that is currently shown. Phishers can make use of the “iframe” tag and make it invisible i.e. without frame borders. In this regard, phishers make use of the “frameBorder” attribute which causes the browser to render a visual delineation.

Rule: IF {*Using iframe→ Phishing Otherwise → Legitimate*

Domain based Features

Age of Domain

This feature can be extracted from WHOIS database (Whois 2005). Most phishing websites live for a short period of time. By reviewing our dataset, we find that the minimum age of the legitimate domain is 6 months.

Rule: IF $\{Age\ of\ Domain \geq 6\ months \rightarrow Legitimate\ Otherwise \rightarrow Phishing\}$

DNS Record

For phishing websites, either the claimed identity is not recognized by the WHOIS database (Whois 2005) or no records founded for the hostname (Pan and Ding 2006). If the DNS record is empty or not found then the website is classified as “Phishing”, otherwise it is classified as “Legitimate”.

Rule: IF $\{no\ DNS\ Record\ For\ The\ Domain \rightarrow Phishing\ Otherwise \rightarrow Legitimate\}$

Website Traffic

This feature measures the popularity of the website by determining the number of visitors and the number of pages they visit. However, since phishing websites live for a short period of time, they may not be recognized by the Alexa database (Alexa the Web Information Company., 1996). By reviewing our dataset, we find that in worst scenarios, legitimate websites ranked among the top 100,000. Furthermore, if the domain has no traffic or is not recognized by the Alexa database, it is classified as “Phishing”. Otherwise, it is classified as “Suspicious”.

Rule: IF

$\{Website\ Rank < 100,000 \rightarrow Legitimate\ Website\ Rank > 100,000 \rightarrow Suspicious\ Otherwise \rightarrow Phishing\}$

PageRank

PageRank is a value ranging from “0” to “1”. PageRank aims to measure how important a webpage is on the Internet. The greater the PageRank value the more important the webpage. In our datasets, we find that about 95% of phishing webpages have no PageRank. Moreover, we find that the remaining 5% of phishing webpages may reach a PageRank value up to “0.2”.

Rule: IF $\{PageRank < 0.2 \rightarrow Phishing\ Otherwise \rightarrow Legitimate\}$

Google Index

This feature examines whether a website is in Google's index or not. When a site is indexed by Google, it is displayed on search results (Webmaster resources, 2014). Usually, phishing webpages are merely accessible for a short period and as a result, many phishing webpages may not be found on the Google index.

Rule: $\text{IF}\{\text{Webpage Indexed by Google} \rightarrow \text{Legitimate Otherwise} \rightarrow \text{Phishing}\}$

Number of Links Pointing to Page

The number of links pointing to the webpage indicates its legitimacy level, even if some links are of the same domain (Dean, 2014). In our datasets and due to its short life span, we find that 98% of phishing dataset items have no links pointing to them. On the other hand, legitimate websites have at least 2 external links pointing to them.

Rule:

IF

$\{\# \text{Of Link Pointing to The Webpage} = 0 \rightarrow \text{Phishing} \# \text{Of Link Pointing to The Webpage} > 0 \text{ and } \leq 1 \rightarrow \text{Legitimate}\}$

Statistical-Reports Based Feature

Several parties such as PhishTank (PhishTank Stats, 2010-2012), and StopBadware (StopBadware, 2010-2012) formulate numerous statistical reports on phishing websites at every given period of time; some are monthly and others are quarterly. In our research, we used 2 forms of the top ten statistics from PhishTank: "Top 10 Domains" and "Top 10 IPs" according to statistical-reports published in the last three years, starting in January 2010 to November 2012. Whereas for "StopBadware", we used "Top 50" IP addresses.

Rule:

IF

$\{\text{Host Belongs to Top Phishing IPs or Top Phishing Domains} \rightarrow \text{Phishing Otherwise} \rightarrow \text{Legitimate}\}$

PHISHING WEBSITES FEATURES

- One of the challenges faced by our research was the unavailability of reliable training datasets. In fact, this challenge faces any researcher in the field. However, although plenty of articles about predicting phishing websites using

data mining techniques have been disseminated these days, no reliable training dataset has been published publically, maybe because there is no agreement in literature on the definitive features that characterize phishing websites, hence it is difficult to shape a dataset that covers all possible features.

- In this article, we shed light on the important features that have proved to be sound and effective in predicting phishing websites. In addition, we proposed some new features, experimentally assign new rules to some well-known features and update some other features.

DATA SET

One of the challenges faced by our research was the unavailability of reliable training datasets. In fact, this challenge faces any researcher in the field. However, although plenty of articles about predicting phishing websites using data mining techniques have been disseminated these days, no reliable training dataset has been published publically, maybe because there is no agreement in literature on the definitive features that characterize phishing websites, hence it is difficult to shape a dataset that covers all possible features.

In this article, we shed light on the important features that have proved to be sound and effective in predicting phishing websites. In addition, we proposed some new features, experimentally assign new rules to some well-known features and update some other features.

3.7 UML DIAGRAMS

UML is simply another graphical representation of a common semantic model. UML provides a comprehensive notation for the full lifecycle of object-oriented development.

ADVANTAGES

- To represent complete systems (instead of only the software portion) using object oriented concepts
- To establish an explicit coupling between concepts and executable code

- To take into account the scaling factors that are inherent to complex and critical systems
- To creating a modeling language usable by both humans and machines
- UML defines several models for representing systems
- The class model captures the static structure
- The state model expresses the dynamic behavior of objects
- The use case model describes the requirements of the user
- The interaction model represents the scenarios and messages flows
- The implementation model shows the work units
- The deployment model provides details that pertain to process allocation

DATA FLOW DIAGRAM

1. The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system.

2. The data flow diagram (DFD) is one of the most important modeling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.

3. DFD shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data moves from input to output.

4. DFD is also known as bubble chart. A DFD may be used to represent a system at any level of abstraction. DFD may be partitioned into levels that represent increasing information flow and functional detail.

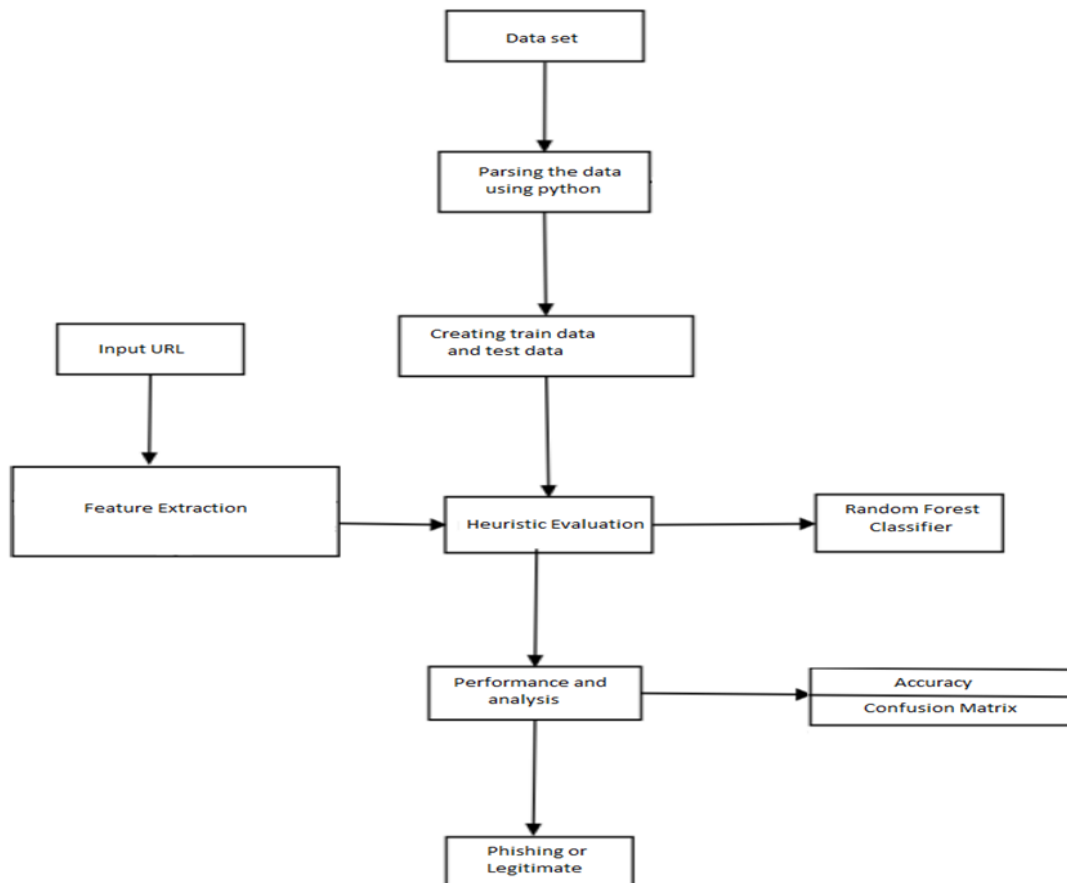


Fig:3.7.1 :Data Flow diagram

3.8 LANGUAGES USED

Python:

Python is a high-level, interpreted, interactive and object-oriented scripting language. Python is designed to be highly readable. It uses English keywords frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages.

- **Python is Interpreted** – Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is similar to PERL and PHP.
- **Python is Interactive** – You can actually sit at a Python prompt and interact with the interpreter directly to write your programs.

- **Python is Object-Oriented** – Python supports Object-Oriented style or technique of programming that encapsulates code within objects.
- **Python is a Beginner's Language** – Python is a great language for the beginner-level programmers and supports the development of a wide range of applications from simple text processing to WWW browsers to games.

History of Python

Python was developed by Guido van Rossum in the late eighties and early nineties at the National Research Institute for Mathematics and Computer Science in the Netherlands.

Python is derived from many other languages, including ABC, Modula-3, C, C++, Algol-68, SmallTalk, and Unix shell and other scripting languages.

Python is copyrighted. Like Perl, Python source code is now available under the GNU General Public License (GPL).

Python is now maintained by a core development team at the institute, although Guido van Rossum still holds a vital role in directing its progress.

Python Features

Python's features include –

- **Easy-to-learn** – Python has few keywords, simple structure, and a clearly defined syntax. This allows the student to pick up the language quickly.
- **Easy-to-read** – Python code is more clearly defined and visible to the eyes.
- **Easy-to-maintain** – Python's source code is fairly easy-to-maintain.
- **A broad standard library** – Python's bulk of the library is very portable and cross-platform compatible on UNIX, Windows, and Macintosh.
- **Interactive Mode** – Python has support for an interactive mode which allows interactive testing and debugging of snippets of code.
- **Portable** – Python can run on a wide variety of hardware platforms and has the same interface on all platforms.

- **Extendable** – You can add low-level modules to the Python interpreter. These modules enable programmers to add to or customize their tools to be more efficient.
- **Databases** – Python provides interfaces to all major commercial databases.
- **GUI Programming** – Python supports GUI applications that can be created and ported to many system calls, libraries and windows systems, such as Windows MFC, Macintosh, and the X Window system of Unix.
- **Scalable** – Python provides a better structure and support for large programs than shell scripting.

Apart from the above-mentioned features, Python has a big list of good features, few are listed below –

- It supports functional and structured programming methods as well as OOP.
- It can be used as a scripting language or can be compiled to byte-code for building large applications.
- It provides very high-level dynamic data types and supports dynamic type checking.
- It supports automatic garbage collection.
- It can be easily integrated with C, C++, COM, ActiveX, CORBA, and Java.

Python is available on a wide variety of platforms including Linux and Mac OS X. Let's understand how to set up our Python environment.

Getting Python

The most up-to-date and current source code, binaries, documentation, news, etc., is available on the official website of Python <https://www.python.org>.

Windows Installation

Here are the steps to install Python on Windows machine.

- Open a Web browser and go to <https://www.python.org/downloads/>.

- Follow the link for the Windows installer python-XYZ.msifile where XYZ is the version you need to install.
- To use this installer python-XYZ.msi, the Windows system must support Microsoft Installer 2.0. Save the installer file to your local machine and then run it to find out if your machine supports MSI.
- Run the downloaded file. This brings up the Python install wizard, which is really easy to use. Just accept the default settings, wait until the install is finished, and you are done.

The Python language has many similarities to Perl, C, and Java. However, there are some definite differences between the languages.

Flask Framework:

Flask is a web application framework written in Python. Armin Ronacher, who leads an international group of Python enthusiasts named Pocco, develops it. Flask is based on Werkzeug WSGI toolkit and Jinja2 template engine. Both are Pocco projects.

Http protocol is the foundation of data communication in world wide web. Different methods of data retrieval from specified URL are defined in this protocol.

The following table summarizes different http methods –

Sr.No	Methods & Description
1	GET Sends data in unencrypted form to the server. Most common method.
2	HEAD Same as GET, but without response body

3	POST Used to send HTML form data to server. Data received by POST method is not cached by server.
4	PUT Replaces all current representations of the target resource with the uploaded content.
5	DELETE Removes all current representations of the target resource given by a URL

By default, the Flask route responds to the **GET** requests. However, this preference can be altered by providing methods argument to **route()** decorator.

In order to demonstrate the use of **POST** method in URL routing, first let us create an HTML form and use the **POST** method to send form data to a URL.

3.9 REQUIREMENT ANALYSIS

Requirement analysis, also called requirement engineering, is the process of determining user expectations for a new modified product. It encompasses the tasks that determine the need for analyzing, documenting, validating and managing software or system requirements. The requirements should be documentable, actionable, measurable, testable and traceable related to identified business needs or opportunities and define to a level of detail, sufficient for system design.

FUNCTIONAL REQUIREMENTS

It is a technical specification requirement for the software products. It is the first step in the requirement analysis process which lists the requirements of particular software systems including functional, performance and security requirements. The function of the system depends mainly on the quality hardware used to run the software with given functionality.

Usability

It specifies how easy the system must be use. It is easy to ask queries in any format which is short or long, porter stemming algorithm stimulates the desired response for user.

Robustness

It refers to a program that performs well not only under ordinary conditions but also under unusual conditions. It is the ability of the user to cope with errors for irrelevant queries during execution.

Security

The state of providing protected access to resource is security. The system provides good security and unauthorized users cannot access the system there by providing high security.

Reliability

It is the probability of how often the software fails. The measurement is often expressed in MTBF (Mean Time Between Failures). The requirement is needed in order to ensure that the processes work correctly and completely without being aborted. It can handle any load and survive and survive and even capable of working around any failure.

Compatibility

It is supported by version above all web browsers. Using any web servers like localhost makes the system real-time experience.

Flexibility

The flexibility of the project is provided in such a way that is has the ability to run on different environments being executed by different users.

Safety

Safety is a measure taken to prevent trouble. Every query is processed in a secured manner without letting others to know one's personal information.

NON- FUNCTIONAL REQUIREMENTS

Portability

It is the usability of the same software in different environments. The project can be run in any operating system.

Performance

These requirements determine the resources required, time interval, throughput and everything that deals with the performance of the system.

Accuracy

The result of the requesting query is very accurate and high speed of retrieving information. The degree of security provided by the system is high and effective.

Maintainability

Project is simple as further updates can be easily done without affecting its stability. Maintainability basically defines that how easy it is to maintain the system. It means that how easy it is to maintain the system, analyse, change and test the application. Maintainability of this project is simple as further updates can be easily done without affecting its stability.

FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

The feasibility study investigates the problem and the information needs of the stakeholders. It seeks to determine the resources required to provide an information systems solution, the cost and benefits of such a solution, and the feasibility of such a solution.

The goal of the feasibility study is to consider alternative information systems solutions, evaluate their feasibility, and propose the alternative most suitable to the organization. The feasibility of a proposed solution is evaluated in terms of its components.

ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 WORKING

INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling

the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

- What data should be given as input?

- How the data should be arranged or coded?

- The dialog to guide the operating personnel in providing input.

- Methods for preparing input validations and steps to follow when error occur.

OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

The output form of an information system should accomplish one or more of the following objectives.

- Convey information about past activities, current status or projections of the Future.

- Signal important events, opportunities, problems, or warnings.

- Trigger an action.

- Confirm an action.

Test plan

Software testing is the process of evaluation a software item to detect differences between given input and expected output. Also to assess the feature of a software item. Testing assesses the quality of the product. Software testing is a process that should be done during the development process. In other words software testing is a verification and validation process.

Verification

Verification is the process to make sure the product satisfies the conditions imposed at the start of the development phase. In other words, to make sure the product behaves the way we want it to.

Validation

Validation is the process to make sure the product satisfies the specified requirements at the end of the development phase. In other words, to make sure the product is built as per customer requirements.

Basics of software testing

There are two basics of software testing: black box testing and white box testing.

Black box Testing

Black box testing is a testing technique that ignores the internal mechanism of the system and focuses on the output generated against any input and execution of the system. It is also called functional testing.

White box Testing

White box testing is a testing technique that takes into account the internal mechanism of a system. It is also called structural testing and glass box testing. Black box testing is often used for validation and white box testing is often used for verification.

Types of testing

There are many types of testing like

- Unit Testing
- Integration Testing
- Functional Testing
- System Testing
- Stress Testing

- Performance Testing
- Usability Testing
- Acceptance Testing
- Regression Testing
- Beta Testing

Unit Testing

Unit testing is the testing of an individual unit or group of related units. It falls under the class of white box testing. It is often done by the programmer to test that the unit he/she has implemented is producing expected output against given input.

Integration Testing

Integration testing is testing in which a group of components are combined to produce output. Also, the interaction between software and hardware is tested in integration testing if software and hardware components have any relation. It may fall under both white box testing and black box testing.

Functional Testing

Functional testing is the testing to ensure that the specified functionality required in the system requirements works. It falls under the class of black box testing.

System Testing

System testing is the testing to ensure that by putting the software in different environments (e.g., Operating Systems) it still works. System testing is done with full system implementation and environment. It falls under the class of black box testing.

Stress Testing

Stress testing is the testing to evaluate how system behaves under unfavorable conditions. Testing is conducted at beyond limits of the specifications. It falls under the class of black box testing.

Performance Testing

Performance testing is the testing to assess the speed and effectiveness of the system and to make sure it is generating results within a specified time as in performance requirements. It falls under the class of black box testing.

Usability Testing

Usability testing is performed to the perspective of the client, to evaluate how the GUI is user-friendly? How easily can the client learn? After learning how to use, how proficiently can the client perform? How pleasing is it to use its design? This falls under the class of black box testing.

Acceptance Testing

Acceptance testing is often done by the customer to ensure that the delivered product meets the requirements and works as the customer expected. It falls under the class of black box testing.

Regression Testing

Regression testing is the testing after modification of a system, component, or a group of related units to ensure that the modification is working correctly and is not damaging or imposing other modules to produce unexpected results. It falls under the class of black box testing.

CHAPTER 5

CONCLUSION

5.1 CONCLUSION

Phishing is a cyber crime procedure utilizing both social building and specialized deception to take individual sensitive data. Besides, Phishing is considered as another extensive type of fraud. Experimentations against recent dependable phishing data sets utilizing different classification algorithm have been performed which received different learning methods. The base of the experiments is accuracy measure.

The aim of this research work is to predict whether a given URL is phishing website or not. It turns out in the given experiment that Random forest based classifiers are the best classifier with great classification accuracy of 82.644% for the given dataset of phishing site. As a future work we might use this model to other Phishing dataset with larger size then now and then testing the performance of those classification algorithm's in terms of classification accuracy.

5.2 REFERENCES

- [1] Mao, J., Bian, J., Tian, W., Zhu, S., Wei, T., Li, A., & Liang, Z. (2018). Detecting phishing websites via aggregation analysis of page layouts. *Procedia Computer Science*, 129, 224-230..
- [2]Atharva Deshpande, Omkar Pedamkar, Nachiket Chaudhary, Swapna Borde "Detection of Phishing Websites using Machine Learning", *International Journal of Engineering Research & Technology*. 10(5), 2527-2531, 2021
- [3] Tyagi, I., Shad, J., Sharma, S., Gaur, S., & Kaur, G. (2018, February). A novel machine learning approach to detect phishing websites. In *2018 5th International conference on signal processing and integrated networks (SPIN)* (pp. 425-430). IEEE..
- [4] Shahrivari, V., Darabi, M. M., & Izadi, M. (2020). Phishing Detection Using Machine Learning Techniques. *arXiv preprint arXiv:2009.11116*.
- [5] HR, M. G., Adithya, M. V., & Vinay, S. (2020). Development of anti-phishing browser based on random forest and rule of extraction framework. *Cybersecurity*, 3(1), 1-14.
- [6] Banu, R., Anand, M., Kamath, A., Ashika, S., Ujwala, H. S., & Harshitha, S. N. (2019, May). Detecting phishing attacks using natural language processing and

machine learning. In 2019 International Conference on Intelligent Computing and Control Systems (ICCS) (pp. 1210-1214). IEEE.

[7].Karabatak, M., & Mustafa, T. (2018, March). Performance comparison of classifiers on reduced phishing website dataset. In 2018 6th International Symposium on Digital Forensic and Security (ISDFS) (pp. 1-5). IEEE.

[8] Shima, K., Miyamoto, D., Abe, H., Ishihara, T., Okada, K., Sekiya, Y., ... & Doi, Y. (2018, February). Classification of URL bitstreams using bag of bytes. In 2018 21st Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN) (pp. 1-5). IEEE.

5.3 APPENDIX

A. SOURCE CODE

Decision Tree Classifier

```
import pandas as pd

legitimate_urls = pd.read_csv("legitimate-urls.csv")
phishing_urls = pd.read_csv("phishing-urls.csv")
legitimate_urls.head(10)
phishing_urls.head(10)
urls = legitimate_urls.append(phishing_urls)
urls.head(5)
urls.columns

urls = urls.drop(urls.columns[[0,3,5]],axis=1)
urls = urls.sample(frac=1).reset_index(drop=True)
urls_without_labels = urls.drop('label',axis=1)
urls_without_labels.columns

labels = urls['label']

from sklearn.model_selection import train_test_split

data_train, data_test, labels_train, labels_test =
train_test_split(urls_without_labels, labels, test_size=0.20, random_state=100)

from sklearn.tree import DecisionTreeClassifier

model = DecisionTreeClassifier()
```

```

model.fit(data_train,labels_train)
pred_label = model.predict(data_test)
from sklearn.metrics import confusion_matrix,accuracy_score
cm = confusion_matrix(labels_test,pred_label)
print(cm)
accuracy_score(labels_test,pred_label)

```

Random Forest Classifier

```

import pandas as pd
legitimate_urls = pd.read_csv("C:extracted_csv_files/legitimate-urls.csv")
phishing_urls = pd.read_csv("C:extracted_csv_files/phishing-urls.csv")
legitimate_urls.head(10)
phishing_urls.head(10)
urls = legitimate_urls.append(phishing_urls)
urls.head(5)
urls = urls.drop(urls.columns[[0,3,5]],axis=1)
urls = urls.sample(frac=1).reset_index(drop=True)
urls_without_labels = urls.drop('label',axis=1)
urls_without_labels.columns
labels = urls['label']
from sklearn.model_selection import train_test_split
data_train, data_test, labels_train, labels_test =
train_test_split(urls_without_labels, labels, test_size=0.30, random_state=110)
from sklearn.ensemble import RandomForestClassifier
random_forest_classifier = RandomForestClassifier()
random_forest_classifier.fit(data_train,labels_train)
prediction_label = random_forest_classifier.predict(data_test)
from sklearn.metrics import confusion_matrix,accuracy_score
cpnfusionMatrix = confusion_matrix(labels_test,prediction_label)
print(cpnfusionMatrix)
accuracy_score(labels_test,prediction_label)

```

```

custom_random_forest_classifier = RandomForestClassifier(n_estimators=500,
max_depth=20, max_leaf_nodes=10000)

custom_random_forest_classifier.fit(data_train,labels_train)

custom_classifier_prediction_label =
custom_random_forest_classifier.predict(data_test)

confusionMatrix2 =
confusion_matrix(labels_test,custom_classifier_prediction_label)

print(confusionMatrix2)

accuracy_score(labels_test,custom_classifier_prediction_label)

import matplotlib.pyplot as plt

import numpy as np

#feature_importances_ : array of shape = [n_features] ----- The feature
importances (the higher, the more important the feature).

#feature_importances_ -- This method returns the quantified relative importance
in the order the features were fed to the algorithm

importances = custom_random_forest_classifier.feature_importances_

#std = np.std([tree.feature_importances_ for tree in
custom_random_forest_classifier.estimators_],axis=0) #[[[estimators_ :explanation
--- list of DecisionTreeClassifier ----- (The collection of fitted sub-estimators.)]]]

#To make the plot pretty, we'll instead sort the features from most to least
important.

indices = np.argsort(importances)[::-1]

print(f"indices of columns : {indices}")

# Print the feature ranking

print("\n ***Feature ranking: *** \n")

print("Feature name : Importance")

for f in range(data_train.shape[1]):

    print(f"{f+1} {data_train.columns[indices[f]]} : {importances[indices[f]]} \n")

print("**** The blue bars are the feature importances of the randomforest classifier,
along with their inter-trees variability****")

# Plot the feature importances of the forest

plt.figure()

plt.title("Feature importances")

plt.bar(range(data_train.shape[1]), importances[indices],

```

```
color="b", align="center")
```

#yerr=std[indices] -- this is another parameter that can be included if std is calculated above

#and also it gives error bar that's the reason we calculate std above. but here we are not making it plot.

```
plt.xticks(range(data_train.shape[1]), data_train.columns[indices])
```

```
plt.xlim([-1, data_train.shape[1]])
```

```
plt.rcParams['figure.figsize'] = (35,15) #this will increase the size of the plot
```

```
plt.show()
```

B. SCREENSHOTS

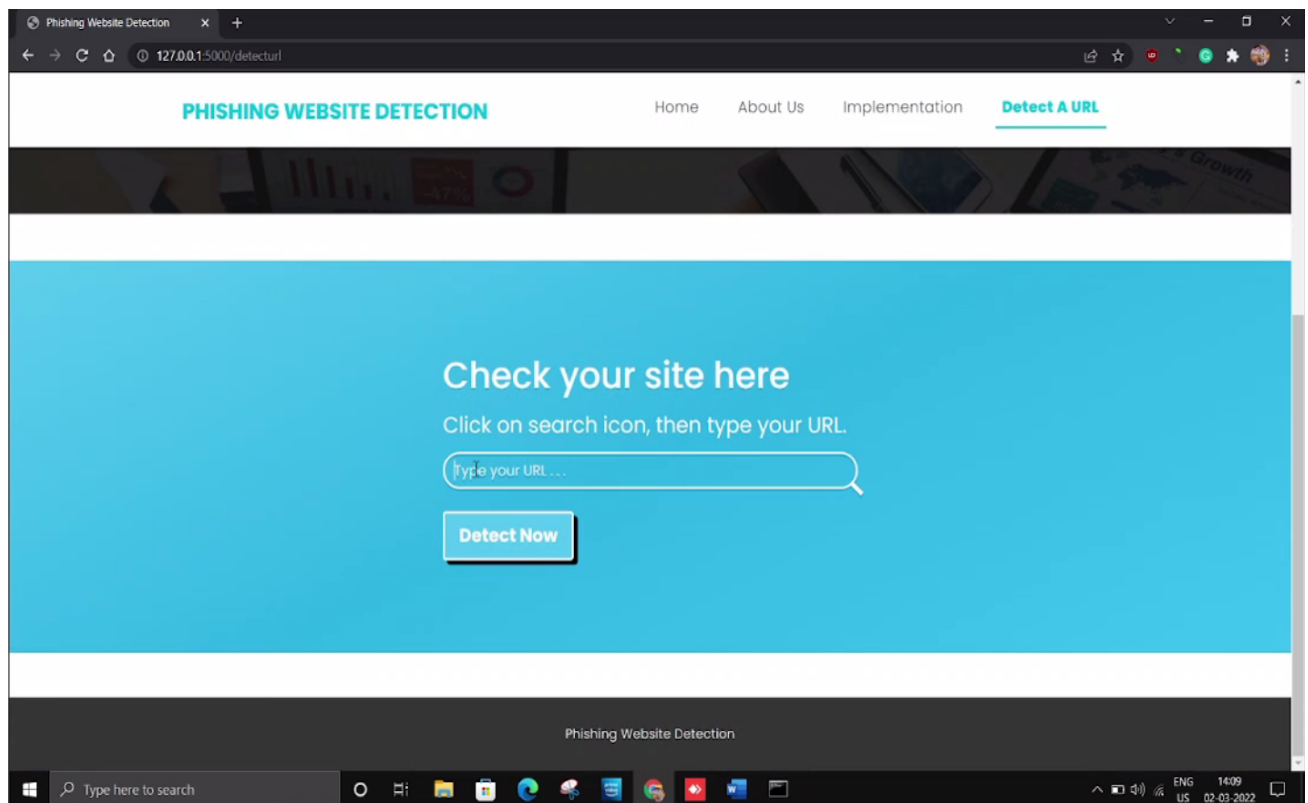


Fig:B.1 : Input page

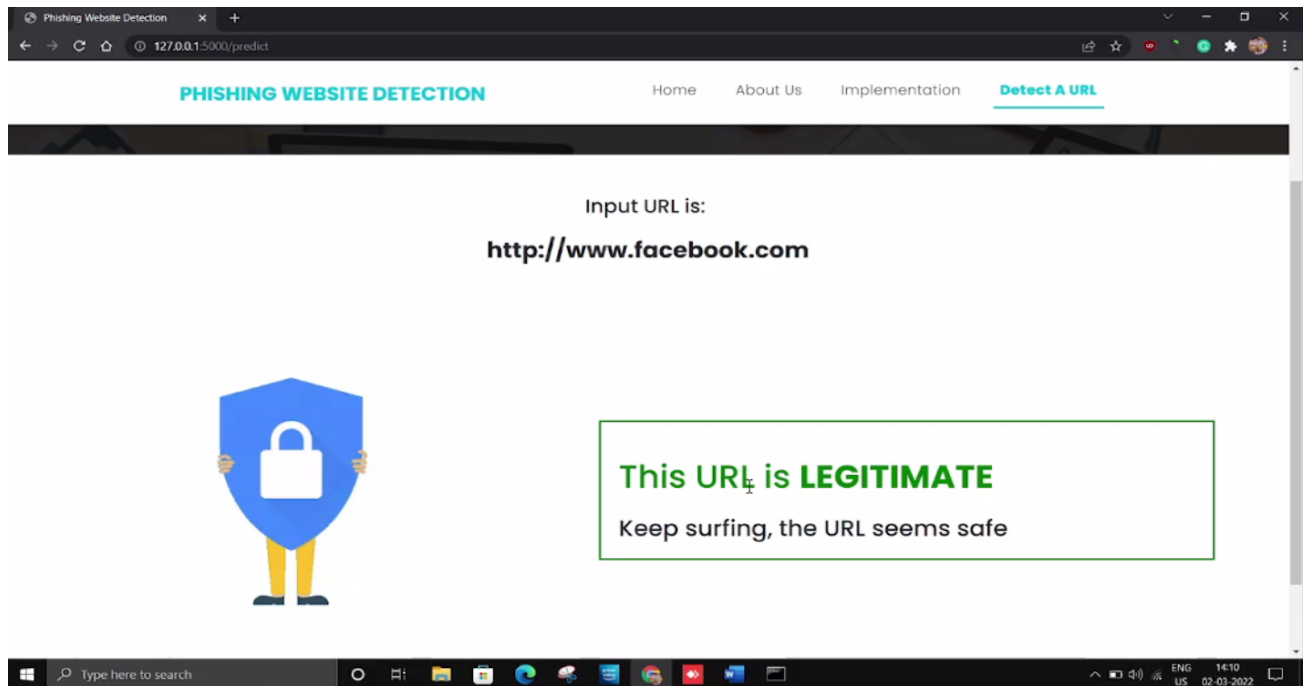


Fig:B.2 : Output page

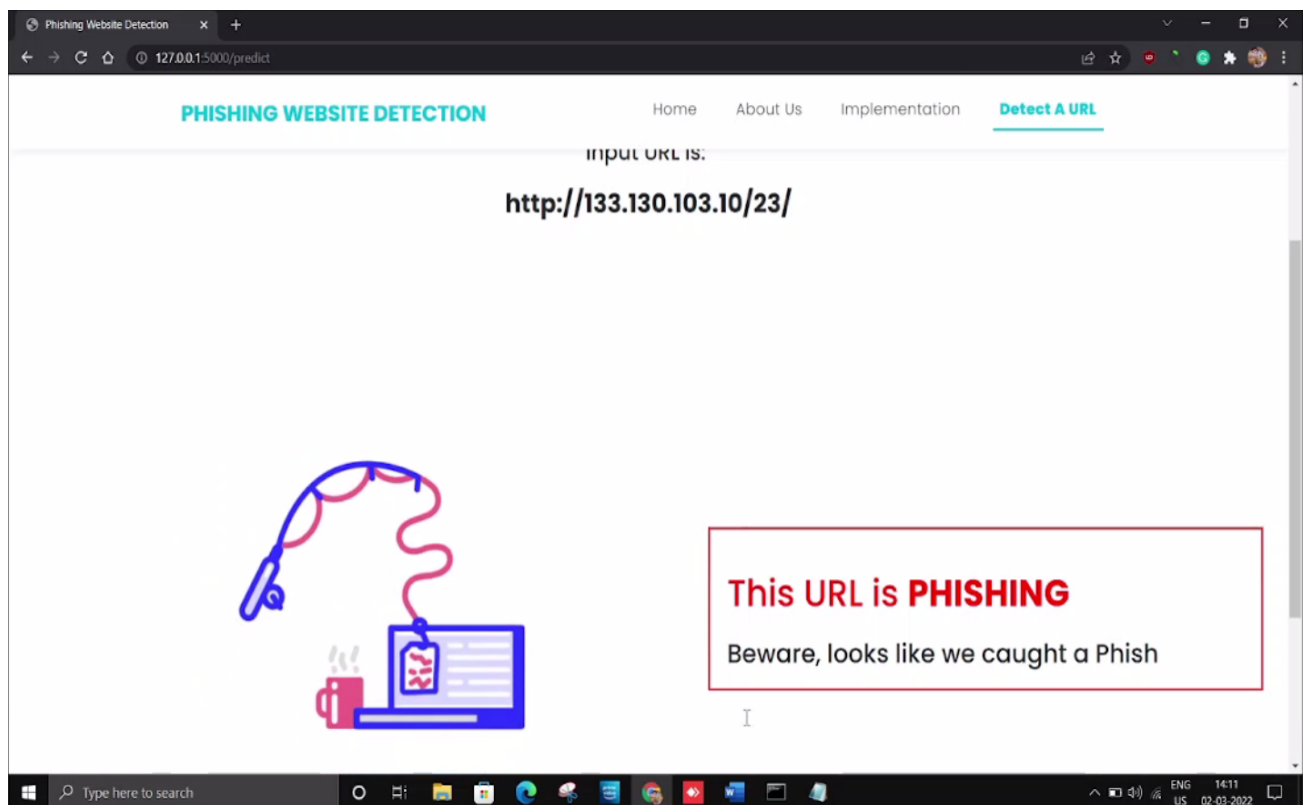


Fig:B.3 : Output page

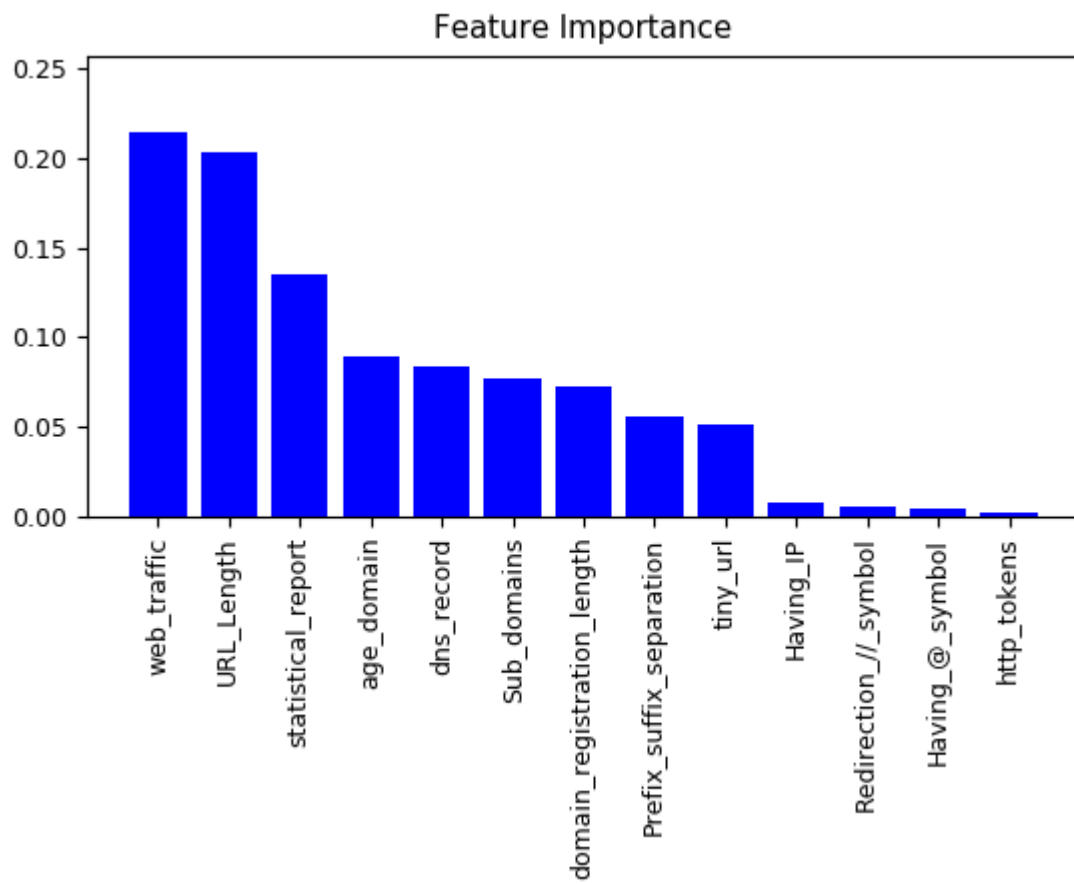
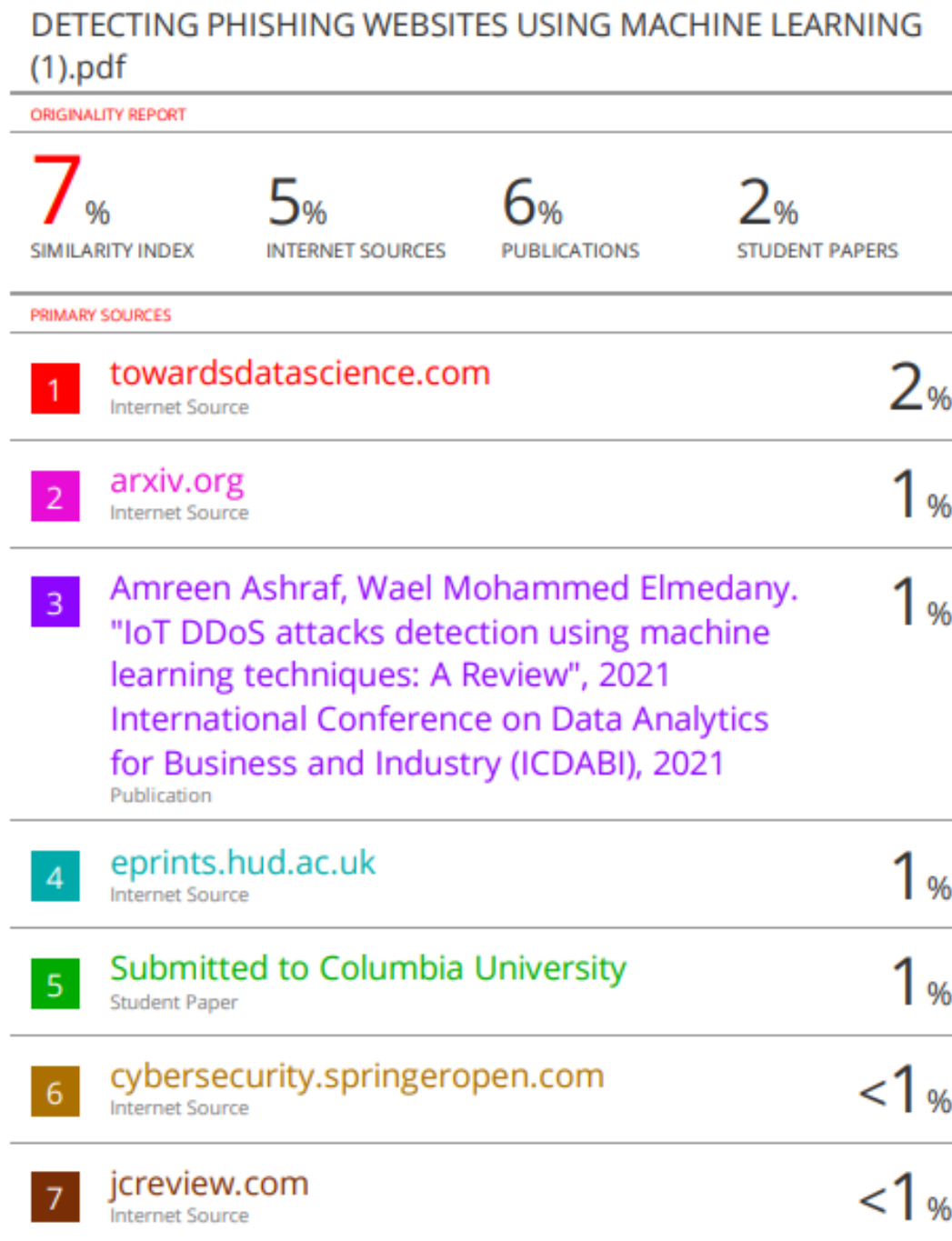


Fig:B.4 : Accuracy Graph

C. PLAGIARISM REPORT



Detecting Phishing Websites Using Machine Learning

Patibanda Pranav Naga Raja Prakash¹
Student
pranavpatibanda@gmail.com

Akula Abinay²
Student
akulaabinay@gmail.com

Dr. Judgi³
Professor
judgi.cse@sathyabama.ac.in

¹²³Sathyabama Institute of Science and Technology, India

Abstract: Phishing attack is a simplest way to obtain sensitive information from innocent users. The presentation of the phishing site isn't actually arranged and is a significant issue with numerous and unaltered parts. Contingent upon the technique utilized by the product designers to make the site, the foundation and different ambiguities can be useful, as certain procedures can be useful and incredible, like the equivocalness, the strong framework, and the data mining innovation, can be utilized successfully. the most effective method to isolate destinations. We utilized the Random Forest (RF), one of the distinctive AI strategies we use to recognize phishing locales. At last, we estimated and thought about the presentation of the classifier completely.

Keywords: *Decision Tree, Random Forest, KNN, Naive Bayes*

I. Introduction

Recollect that extortion is a definitive objective of getting functional data, for example, passwords, benefit records, and MasterCard numbers, and that noxious sites are a type of misrepresentation that happens when such moves are made. While there is little distinction between the phishing program and the manners by which the phishing exercises in the message are distinguished and the on-location phishing factors are recognized, fish are considering new techniques and types to go around the program and framework. Phishing is a technique that utilizes a blend of more point-by-point plan and development to gather basic and remarkable data, for example, passwords, individual and web-based business cards, and basic business cards. Phishing

causes your courier to seem noticeably more appealing, and notwithstanding cash related activities, for example, items coming from organizations, web-based business can lead clients to visit counterfeit destinations. Sites that mirror the picture of a genuine site.

Managers can trap different givers during the time spent making an interpersonal organization, for example, not shutting a client's record because of an inadequate record recharging process, or giving other data to clients to affirm various reports and objectives. Visit their criticism page. Strategy Technique is the best method for utilizing quick and solid strategies to interface with information from an informational collection for solo preparing. That is the reason we utilized the illustration in our work.

II. Scope and objectives

A. Objectives:

- Comprehend the qualities of a model (or model) and how it varies from a genuine system
- Why it is essential to know this space, and how to recognize it utilizing AI and language handling abilities
- Survey current strategies for AI to distinguish awful URLs in books
- Comprehend the URL that looks terrible and the standards used to plan the framework as an assistance.

Separate the phishing site from the authority site and make the exercises more straightforward for the clients

III. METHODOLOGY

There is a great deal of exploration on calculations and various kinds of data in science course books and business items. Phishing URLs and related pages have many separating them from terrible URLs.

For example; an attacker can register long and confusing domain to hide the actual domain name (Cyber-squatting, Typo squatted)

Features collected from academic studies on the detection of identity theft by sensitive information about machine learning techniques are included in the category as provided below.

- URL-Based Features
- Domain-Based Features
- Page-Based Features
- Content-Based Features

Moreover, much use more technical features and process them using machine learning algorithms has been imposed.

IV. LITERATURE SURVEY

(Mao, J., Bian, J., Tian, W., Zhu, S., Wei, T., Li, A., & Liang, Z. 2018) [1] In this article, we intend to further develop understanding abilities through AI. Specifically, it is suggested that the exploration technique be founded on the page design choice strategy utilized for page search. The consequences of the review show that our techniques are exact and helpful in deciding the phishing sheet.

(Atharva Deshpande , Omkar Pdamkar , Nachiket Chaudhary , Dr. Swapna Borde/ 2021) [2] This page analyzes the apparatuses used to learn and get machines. Phishing is known for its gatecrashers since duping somebody is more straightforward to hit on a terrible line than beating a safeguard framework. The negative connections in the principle body of the message are expected to show that these corporate images and other real items are utilized to arrive at degenerate associations.

(Ishant Tyagi; Jatin Shad; Shubham Sharma; Siddharth Gaur; Gagandeep Kaur/ 2018) [3] This page centers around different AI calculations pointed toward foreseeing whether a site is misled or real. Machine preparing is famous on the grounds that it can distinguish party time assaults and is great at beating new kinds of phishing assaults. In our work, we had the option to precisely decide 98.4% by anticipating phishing or lawful area.

(Vahid Shahrivari, Mohammad Mahdi Darabi, Mohammad Izadi/ 2020) [4] Understanding Phishing Using Machine Learning Techniques Wahid Shahrivari, Mohammad Mahdi Darabi, Mohammad Izadi/2020. The best method for recognizing these awful encounters is through AI. This is on the grounds that numerous phishing assaults are the most widely recognized types of AI. In this article, we think about the aftereffects of many AI strategies to foresee phishing.

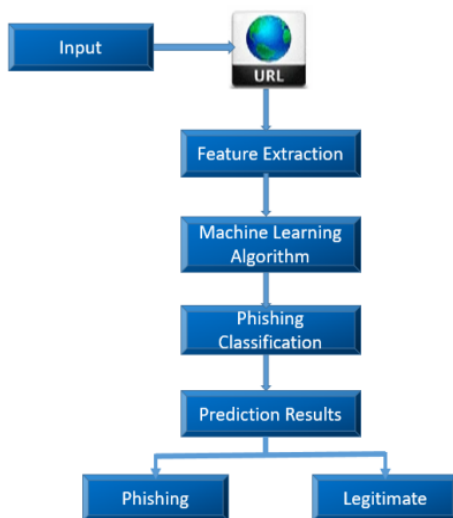
(Mohith Gowda HR, Adithya MV, Gunesh Prasad S & Vinay S/ 2020) [5] In this article, we need specialized ability to effortlessly recognize a phishing site on the client side that requirements to assemble a web crawler. In this framework, we utilize the erase rule to eliminate content or site highlights utilizing just the URL. The rundown comprises of 30 unique URLs and will then, at that point, be utilized to discover reality with regards to the site by irregular words arrangement.

(Banu, R., Anand, M., Kamath, A., Ashika, S., Ujwala, H. S., & Harshitha, S. N. 2019) [6] The approach will also use Deep Learning frameworks with hierarchical long-short term memory networks (H-LSTMs) and attention mechanisms to model the emails simultaneously at the word and sentence level. Phishing attacks categorizes the emails based on certain properties which give more details about the source of phishing. Generally, most of the existing systems focus on email classification depending upon header part or body part.

(Karabatak, M., & Mustafa, T. 2018) [7] This article inspects information assortment on the UCI site. Diminishing its size and looking at the presentation of positioning calculations is being contemplated in the news site of the phishing site. The portrayal of the phishing site is taken from the UCI information base of AI. The data set comprises of 11055 passages and 31 exercises. The presentation of the arranging calculation is currently contrasted with other data on the grouping calculations. At long last, contrasting the requesting elements of the informational indexes utilizing the overall calculations gave.

(Shima, K., Miyamoto, D., Abe, H., Ishihara, T., Okada, K., Sekiya, Y., ... & Doi, Y. 2018) [8] In this review, we tested with realistic URL access history data taken from a research organization and data from the famous archive site of phishing site information, PhishTank.com. Our approach achieved 2~3% better accuracy compared to the existing DL- based approach.

V. SYSTEM ARCHITECTURE



VI. EXISTING SYSTEM

There are innumerable spaces that can be defrauded, like web-based installments, webmail, monetary foundations, document stockpiling or distributed storage. Web and online installments are remembered for the rundown of best practices. Since phishing should be possible by email or lance phishing , the client ought to know about the effect and not be 100% certain about the general security activity. Machine preparing is probably the most ideal way to master phishing procedures as it dispenses with the risks of existing strategies.

VII. PROPOSED SYSTEM

Endeavors to gather individual data deceitfully are turning out to be more normal today. To assist clients with knowing how to access such a site, a framework has been executed that tells clients by means of email and a spring up window when they attempt to get to the site. This page gives a boycott identification framework known as a phishing site with the goal that the site can be told when looking or signing in. Along these lines, it tends to be utilized as a genuine

apparatus to distinguish, convince, and forestall misrepresentation.

VIII. MODULES

- A. Detection Technique
- B. Phishing Websites Features
- C. Data Set

DETECTION TECHNIQUE

Promoting on the phishing site is exceptionally well known with regards to the security of your clients. Thusly, numerous procedures have been created to distinguish phishing site, from correspondence innovations like confirmation conventions, boycotting, and whitelists, to separating. The boycott and the arranging list have not been demonstrated to be sufficient to be utilized in various areas, so they are not utilized. Simultaneously, the fishery content has been broadly utilized and shown to be compelling. In such manner, the exploration centers around the improvement of content-based techniques, email and AI, just as data innovation.

PHISHING WEBSITES FEATURES

One of the obstructions to our learning is the absence of dependable preparing data. Indeed, this issue is normal to any specialist in the field. In any case, the greater part of the issues identified with the forecast of the phishing site utilizing data mining innovation are boundless nowadays, however solid data on the preparation has not been disclosed, maybe because of misconceptions in the depiction of the phishing site. Building an information pack incorporates all that could be within reach. In this article, we will feature the absolute most significant elements that have demonstrated to be dependable and powerful in anticipating phishing destinations. Also, we have presented new things, given new guidelines a shot a portion of the natural capacities, and refreshed different capacities.

DATA SET

One of the challenges faced by researchers in predicting phishing websites is the lack of reliable

training datasets. Due to the lack of such data, it is difficult to develop a comprehensive data set that covers all the possible features that can be used to predict phishing websites. This article aims to provide a list of the most effective features that can be used to predict phishing websites.

ALGORITHMS

- Decision Trees
- Random forest

RANDOM FOREST:

As the name suggests, a characteristic woodland is comprised of an enormous number of trees that cooperate. Each tree in the woods makes a phase, and afterward countless uproarious gatherings become our model.

Decision Trees for Classification: A Machine Learning Algorithm

The choice tree is a sort of control machine where the data is continually dispersed by specific measures (this implies the information and result of the preparation). A tree can be characterized by two variables: direction and fiendishness. Leaves are an official conclusion or arrangement. Choices are made based on data sharing. An illustration of a tree arrangement can be clarified utilizing a twofold tree above. Assume you need to foresee an individual's life dependent on data like age, dietary patterns, and active work. The subjects covered are "The manner by which old are you?", "Do you work out?", "Do you work out, etc. eat a ton of pizza? Also leaves are the "right" or "wrong" arrangement. For this situation, it was a two-section issue (type no issue). There are two fundamental kinds of choice trees: Sorting trees (Yes/No Type) What we saw above is an illustration of a tree, and the outcomes were variable as 'fitting' or 'unseemly'. Here the choices shift by classification. Compromise Tree (Media Type Continued) Here the choice or reaction changes Continued, for example Numbers like 123, etc. Movement Now that we have distinguished the choice tree, we will take a gander at how it functions inside. There are numerous calculations that make an

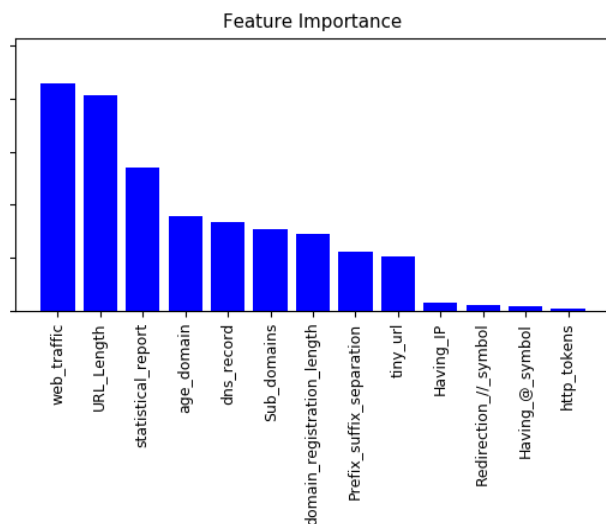
answer, yet the best is known as the ID3 calculation. ID3 implies dichotomator causes 3. Before we talk about the ID3 calculation, we will take a gander at a couple of subtleties.

IX. SCREENSHOTS

Fig.1.Table detecting the malicious websites based on different domains

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Domain	Having_@	Having_IP	Path	Prefix_suf	Protocol	Redirectic	Sub_dom	URL_Leng	age_dom	dns_recor	domain_r	http_tole	label	statistical	tiny_url	web_traffic
2	www.liqu	0	0	/	0	http	0	0	0	0	0	1	0	0	0	0	2
3	www.onli	0	0	/	0	http	0	0	0	0	0	1	0	0	1	0	1
4	www.cere	0	0	/mekol/s	0	http	0	1	0	1	0	1	0	0	0	0	0
5	www.gale	0	0	/kmh/	0	http	0	0	0	0	0	0	0	0	0	0	0
6	www.famv	0	0	/	0	http	0	0	0	1	1	1	0	0	1	0	1
7	www.anir	0	0	/	0	http	0	0	0	0	0	1	0	0	1	0	1
8	www2.111	0	0	/mb1996	0	http	0	1	0	1	0	1	0	0	0	0	2
9	archive.rh	0	0	/fritters/y	0	http	0	2	0	0	0	1	0	0	0	0	2
10	www.free	0	0	/	0	http	0	0	0	0	0	1	0	0	0	1	2
11	www.cute	0	0	/	0	http	0	0	0	2	0	0	0	0	0	0	2
12	www.tare	0	0	/	0	http	0	0	0	2	0	2	0	0	0	0	2
13	www.inte	0	0	/users/po	0	http	0	2	2	0	0	1	0	0	0	0	1
14	darkkamir	0	0	/	0	http	0	0	0	1	1	1	0	0	1	0	1
15	www.lei.r	0	0	/bkosl/v	0	http	0	2	0	0	0	1	0	0	1	0	1
16	www9.kir	0	0	/fetish/he	0	http	0	0	0	2	0	2	0	0	0	1	0
17	www.jaso	0	0	/	0	http	0	0	0	0	0	1	0	0	0	0	1
18	www.geo	0	0	/kaseycha	0	http	0	2	0	2	0	2	0	0	0	0	2
19	www.ang	0	0	/journal/c	0	http	0	2	2	0	0	0	0	0	0	0	0
20	e.webring	0	0	/hub	0	http	0	0	2	0	0	0	0	0	0	0	2
21	www.nem	0	0	/	0	http	0	0	0	1	1	1	0	0	1	0	1
22	j-heaven	0	0	/library.ht	1	http	0	2	0	0	0	0	0	0	0	0	1
23	www.ang	0	0	/poetry/n	0	http	0	0	0	0	0	0	0	0	0	0	0

Fig. 2.Graph depicting the accuracy of different classifications of url.



X. CONCLUSIONS

This paper aims to enhance detection method to detect phishing websites using machine learning technology. Dependable enemy of phishing sites have been tried utilizing various calculations, utilizing diverse preparing techniques. The premise of the test

is the real rate. The motivation behind this review is to decide whether a URL is a phishing site. It turns out in the given experiment that Random forest based classifiers are the best classifier with great classification accuracy of 82.644% for phishing. For future work, we would now be able to utilize this model in other enormous phishing sites and test the presentation of these calculation stages to discover more.

XI. REFERENCES

- [1] Mao, J., Bian, J., Tian, W., Zhu, S., Wei, T., Li, A., & Liang, Z. (2018). Detecting phishing websites via aggregation analysis of page layouts. *Procedia Computer Science*, 129, 224-230..
- [2] Atharva Deshpande, Omkar Pedamkar, Nachiket Chaudhary, Swapna Borde "Detection of Phishing Websites using Machine Learning", *International Journal of Engineering Research & Technology*. 10(5), 2527-2531, 2021
- [3] Tyagi, I., Shad, J., Sharma, S., Gaur, S., & Kaur, G. (2018, February). A novel machine learning approach to detect phishing websites. In 2018 5th International conference on signal processing and integrated networks (SPIN) (pp. 425-430). IEEE..
- [4] Shahrivari, V., Darabi, M. M., & Izadi, M. (2020). Phishing Detection Using Machine Learning Techniques. *arXiv preprint arXiv:2009.11116*.
- [5] HR, M. G., Adithya, M. V., & Vinay, S. (2020). Development of anti-phishing browser based on random forest and rule of extraction framework. *Cybersecurity*, 3(1), 1-14.
- [6] Banu, R., Anand, M., Kamath, A., Ashika, S., Ujwala, H. S., & Harshitha, S. N. (2019, May). Detecting phishing attacks using natural language processing and machine learning. In 2019 International Conference on Intelligent Computing and Control Systems (ICCS) (pp. 1210-1214). IEEE.
- [7] Karabatak, M., & Mustafa, T. (2018, March). Performance comparison of classifiers on reduced phishing website dataset. In 2018 6th International Symposium on Digital Forensic and Security (ISDFS) (pp. 1-5). IEEE.
- [8] Shima, K., Miyamoto, D., Abe, H., Ishihara, T., Okada, K., Sekiya, Y., ... & Doi, Y. (2018, February). Classification of URL bitstreams using bag of bytes. In 2018 21st Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN) (pp. 1-5). IEEE.

