JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Phishing Detection: Using Machine Learning

[1]Khalid Shaikh, [2]Supriya Shelar, [3]Abhishek Prasad

[1]Student,[1]Information Technology Department, [1]Shree L R Tiwari College of Engineering(Mumbai University)

[2]Student,[2]Information Technology Department, [2]Shree L R Tiwari College of Engineering(Mumbai University)

[3]Student,[3]Information Technology Department, [3]Shree L R Tiwari College of Engineering(Mumbai University)

Abstract: Phishing is a massive threat to all Internet users, and it's difficult to track down or protect against because it doesn't seem to be malicious. In today's world, everything is uploaded to the internet, putting personal information at risk. Phishing affects people all over the world and is carried out on a large scale, making it impossible to track down and prosecute the perpetrators. Just one common technique that phishers have used is a tactic known as "quick flux," in which they use a wide pool of proxies and URLs to hide the true location of the phishing site. This makes it more difficult to blacklist the site, and it takes longer to locate the server. Through blacklisting or banning phishing pages, or filtering out phishing emails, phishing can be stopped before it hits the user. The first approach involves manually inspecting URLs and the places that they claim to be, or using machine learning to automate the process. Owing to the more nuanced nature of phishing, there are few effective spam filters used by email servers. Machine learning techniques are now being used to create phishing filters. ‘‘Modeling and Preventing Phishing Attacks': The author of this paper has included a set of visual aids to help readers understand how phishing attacks function. The authors also clarify how each of these variables is interpreted. In their paper, "Classification of Phishing Email Using Random Forest Machine Learning Technique," the authors describe the characteristics used to identify phishing emails. Using URLs with an IP address, non-matching 'href' attributes and connection text, the number of dots in a domain name, and testing domain names against the email sender are a few examples. Their experiment yielded an accuracy of 99.7 percent with a very low false positive rate of about 0.06 percent. There is a comparison study of anti-phishing detection, prevention, and security mechanisms from the previous decade. On the basis of email structure, the vulnerable area is divided into three groups. The number of vulnerabilities protected by current anti-phishing mechanisms is identified in order to determine if the vulnerability is centered or unfocused. This research paper can be thought of as a tutorial for a decade's worth of anti-phishing research. The current study investigates the efficacy of methods and strategies used to combat email phishing.

Index Terms **- Phishing detection system, Malicious URL, Email spam, Anti-phishing, Cyber Threat.**

I. INTRODUCTION

In today’s world technology is advancing very fast and there is a huge volume of data transfer. Advancement in technologies leads to risk of data Confidentiality, Integrity and Availability which leads to the security concern of the data. Computer security or Information Technology security is the safeguard of computer systems and networks from the theft of or harm to their hardware, software, or electronic data. Owing to its complexity, both in terms of policies and technologies, cybersecurity is also one of the major challenges in the modern-day world. To safeguard digital devices, it is significant to know the attacks that can be made against it. One of the Threats is Protecting System or Data against Phishing Attacks. Phishing is characterized as the deceitful procurement of secret information by the proposed beneficiaries and the abuse of such information. The phishing assault is frequently done by websites. In a typical case of Phishing, a legitimate website is seen by all accounts, to be from realized sites, from a client's bank, MasterCard organization, website, or Internet specialist co-op. For the most part, individual data, for example, MasterCard number or secret word is approached to refresh accounts. These messages contain a URL connection that guides clients to another site. This webpage is really a phony or adjusted site. At the point when clients go to this site, they are approached to enter individual data to be sent to the phishing attacker. The Phishing Application, which is intended to forestall genuine dangers like this, gets pernicious messages showing up at websites, and tends to be incorporated into the framework. The hacker can register any domain name that has not been listed before. This part of Uniform Resource Locator can be set only once. The phisher can modify the Free URL at any time to create a new URL. The reason why security professionals find it difficult to detect phishing domains is because of the unique and irreplaceable part of the website domain. When a domain is identified as fraudulent, it is easy to avert this domain before a user accesses it.

II. PROBLEM DEFINITION

In today's cyber world, there are several phishing attacks. Phishing attacks occur in a variety of ways, including website phishing, phishing applications, and phishing URLs, which are clones of legitimate websites. There are several application tools and prevention techniques available on the market to protect against all types of attacks. Fake Email, Fake Website, and Dummy Device Detection techniques are also included in the application. In today's world, these three aspects of the Cyber domain are primarily targeted by these types of attacks in order to gain access to data or steal sensitive information. This Application based on Machine Learning Algorithms and Techniques can help to detect Phishing Application, URLs, and also Emails respectively.

III. LITERATURE SURVEY

[1] Jyoti Chhikara Ritu Dahiya Neha Garg Monika Rani in paper "Phishing & Anti-Phishing Techniques" in this paper Phishing differs from traditional scams primarily in the scale of the fraud that can be committed. In order to combat phishing, business and consumers need to adopt best practices and practice awareness, educate themselves about phishing techniques, use current security protection and protocols, and report suspicious activities. And its methodology is branded Monitoring, Behavior Detection, and Security Event Monitoring. And the final technical solution to phishing does not involve significant infrastructure changes in the Internet However there are few techniques that can be taken now to reduce the consumer's vulnerability to phishing attacks.
[2] Dr.Radha Damodarm in paper "Study on phishing attacks and anti-phishing tools" in this paper Phishing is the attempt to acquire sensitive information such as usernames, passwords, and credit card details (and sometimes, indirectly, money), often for malicious reasons, by masquerading as a trustworthy entity in an electronic communication. A content-Based Approach to Detecting Phishing Web sites, Ant Colony Optimization, Particle Swarm Optimization and Periodical updating of anti-phishing tools or software's in their own systems may helpful to secure their confidential information and credentials. This study may give the awareness about the phishing problems and solutions.
[3] Muhammet Baykara Zahit Ziya Gürel in paper "Detection of phishing attacks" in this paper Anti Phishing Simulator collects phishing and spam message at a common point. In addition to getting spam box, it allows you to control the "spam box" whenever you want. Those who are technically qualified by the "URL Control" and Bayesian spam filtering,WEKA,Naive Bayes techniques,UCI Machine Learning storage.It only aims to the spam contents and not the interconnected URLs that is artificial neuron networks.
[4]Kalpana,Naveen Kuma,Parul Saharavat in paper "Email Phishing- An open threat to everyone" in this paper An end-host based anti-phishing algorithm which we call LinkGuard, based on the characteristics of the phishing hyperlink. Since LinkGuard is character based, it can detect and prevent not only known phishing attacks but also unknown ones, This Link Guard algorithm plus includes one step forward with the detection of phishing measures. Phishing websites is captured using the text reading to find out the malicious. It only Detect known attack which means it can detect up to 75% of accuracy as they have not used link guard attack.

3.1 COMPARATIVE STUDY

Parameters	Decision Tree	Random Forest	Support vector machine
Accuracy	93.1	97.14	90.5
False Negative Rate	4.7	3.14	5.26
False Positive Rate	2.98	2.61	3.45
Method used	To create a tree structure by using the attribute information of the training data can be expressed as asking questions to the data and reaching the results as soon as possible according to the obtained answers.	using decision trees as the basic classifier and creating a collective learning model by combining multiple decision trees	a nonparametric classifier method suggested to solve classification problems in data sets where patterns between variables are unknown
Gini index	used	used	not used
Speed of classification	Fast	Fast	slow

IV. PROPOSED SYSTEM

The proposed system is a Website. The website is built with the use of HTML, CSS and JavaScript. This Project aims illuminating the different components of URL phishing. Through this comprehensive examination and investigation, it is recognized that the specialists were centered for the most part on page content vulnerabilities which was for the most part used to deceive credulous clients. Area vulnerabilities and code scripting vulnerabilities are inclined to greater tricks and uses trend setting innovation to deceive innocent just as capable clients. It has been seen that part of work finished with respect to lessening phishing, significant work is being done to give a sheltered and secure URL and network access, the equivalent can likewise be said to be valid for the other the fence yet at the same time is required to prevent from root.

V. METHODOLOGY

The main goal of our paper is to control the security of information and to avert Violations, to check and identify whether spam is available from the current database, to enable the user to create his own spam list, and to check whether the incoming or receiving mail has dangerous and risky content. Adding the mail account to the system will secure the mail account and the user will also communicate and control it without having to open the mail. With this module, it is possible to determine the classification results of keywords and passages in the database by Link Guard algorithm. There is a panel where the URLs in the mail are thoroughly inspected and the fake URL is detected after that user is informed that this URL is malicious. Phishing is a social engineering attack that uses electronic media such as websites to deliver content with languages to persuade and trick victims to perform certain actions. Another part is Malicious URL Detector. The main purpose of this is to detect malicious URLs by using Machine Learning Algorithm to Classify the URL as Legitimate and Bad URL i.e. Phishing URL. Last Part of the Project is the Fake Application Detector. It aims to detect Legitimate or Fake Application Dynamically Checking Ratings, Reviews and Ranking of the app through Machine Learning Techniques.

The Anti-Phishing application aims to control the security of information and to prevent Infringements, to check whether spam is available from the current database, and to check whether the incoming mail has dangerous content. The inclusion of the mail account to be protected in the system with this module, the user will also control and communicate without having to open the mail. With this module, it is possible to determine the classification results of keywords and passages in the database by Random Forest algorithm. The panel where the URLs in the mail are checked and the fake URL is detected. A comparative analysis report of anti-phishing detection, prevention and protection mechanisms from last decade is listed. This comparative analysis reports the anti-phishing mechanism run on server side or client side and which the vulnerable area is divided into three categories on the basis of website structure. The amount of vulnerabilities covered by existing anti-phishing mechanisms is listed to identify the focus on vulnerability by researchers. This paper could be said as a tutorial of an existing anti-phishing research work from decades ago. It gives a glimpse and a deep analysis on anti-phishing which could lead recent researchers to fill gaps of security breach causing phishing attacks and innovate. When phishing assaults are made, they are frequently utilized for ridiculous news, for example, those made around significant occasions, occasions and commemorations. Normally a casualty gets a message that seems to have been sent by a known individual or association. The assault is completed by means of a malevolent record infusion that incorporates phishing programming or through connections to noxious sites. In either case, the objective is to guide the client to a malicious site to introduce malignant programming on the gadget or to fool them into uncovering individual and money related data, for example, casualties, passwords, account IDs, or charge card subtleties. An effective phishing message normally appeared from a notable organization, it is hard to tell from the first message: in phishing messages, organization logos and other clear designs and information gathered from the organization. Similarly, as with other connection control methods, the utilization of subdomains and incorrectly spelled URLS (frequently spelling botches) is normal. Phishing assailants use JavaScript to put an authentic URL of the URL onto the program's location bar. The URL produced by exploring through an installed connection can likewise be altered utilizing JavaScript. Safeguard against phishing assaults should start with preparing and illuminating clients to distinguish phishing messages; however, there are different systems that can lessen effective assaults. For instance: a system portal website channel can catch many focused-on phishing messages and diminish the quantity of phishing messages arriving at clients' inboxes. In the implementation phase of the project, we will include the bug fixes and changes that we find in the testing phase of the project. To conduct beta testing for identifying any further errors, bugs and improvements that can be performed. Once the improvements are included, the system will be implemented as a live application to the end user. IT projects require resources in terms of money, time, human resources, infrastructure and technology, both hardware and software. Resources are not just a mean, but also an approximation of constraints.

VI. PROJECT PLANNING

Project planning is essential to managing the scope, schedule and budget of the project. For this, we used tools such as MS Excel, MS PowerPoint, and online MS project as well as various modeling tools, such asdraw.io. Stakeholder perspective is crucial to the success of this project. Part of the reason is that phishing detection is a highly sensitive field, and even the slightest of errors, which are evidently unavoidable in even the most sophisticated software, can lead to the patient's condition worsening. Thus, we made it a necessity to search for user consensus before we planned for features to be built in our project. This was done by researching search interest on search engines, visiting forums pertaining to machine learning, artificial intelligence, healthcare as well as wearable technology. The involvement of stakeholders is critical to the project's progress. Part of the explanation is that medical diagnosis is a delicate area, and even the tiniest of errors, which are inevitable even in the most advanced software,

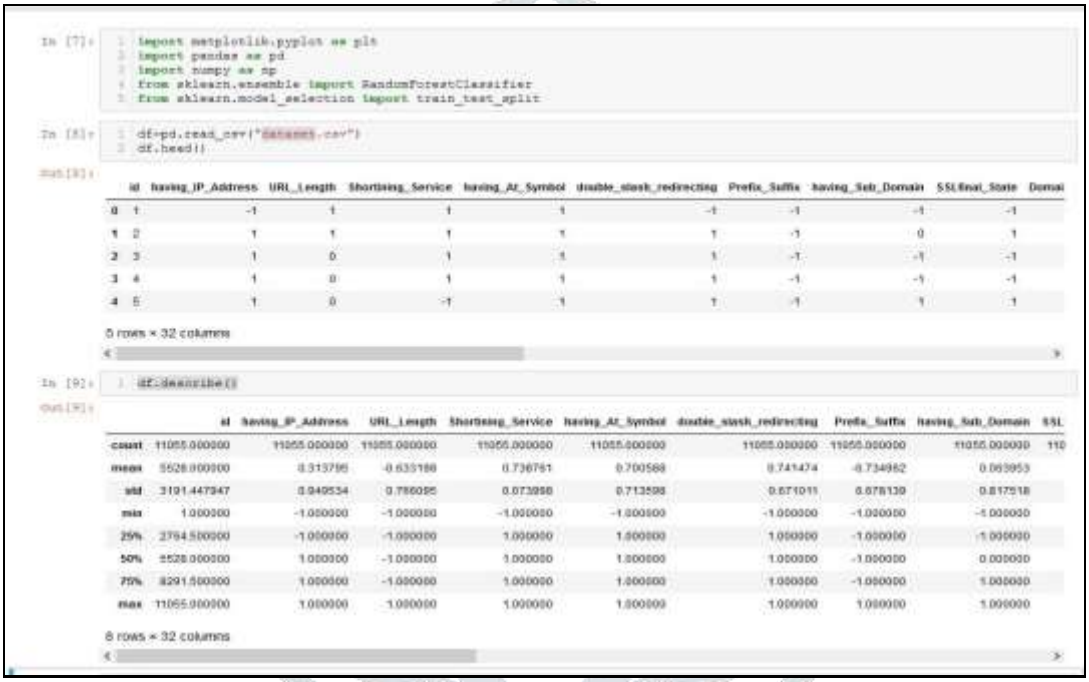
will result in the patient's condition worsening. As a result, we made it a requirement to seek user consent before planning features for our project. This was accomplished by conducting keyword analysis on search engines, as well as visiting forums devoted to machine learning, artificial intelligence, healthcare, and wearable technology. We divided our project into phases and sub phases, with each sub-phase receiving a date range of one to three weeks. The timeline map function of online MS Project was used to accomplish this. The schedule dependencies were then modelled and fine-tuned using a Gantt map. The Gantt chart was developed in MS Excel using the Gantt chart function.

VII. IMPLEMENTATION

7.1 DATASET USED

The dataset for Smart Phishing System known as “phisp coop” was collected from Kaggle. The dataset contains many features which includes Ip addresses, Url Length, Prefix, SSL, double slash etc. The data was having irregularities and hence were removed in the data preprocessing step. The dataset for smart phishing was taken from the open source GitHub repository. Following is a snapshot of the dataset.

7.2 Data Set and Input Values



```
In [7]: import matplotlib.pyplot as plt
        import pandas as pd
        import numpy as np
        from sklearn.ensemble import RandomForestClassifier
        from sklearn.model_selection import train_test_split

In [8]: df=pd.read_csv("dataset.csv")
        df.head()
```

Out[8]:

	id	having_IP_Address	URL_Length	Shortining_Service	having_At_Symbol	double_slash_redirecting	Prefix_Suffix	having_Sub_Domain	SSLfinal_State	Domai
0	1	-1	1	1	1	-1	-1	-1	-1	
1	2	1	1	1	1	1	-1	0	1	
2	3	1	0	1	1	1	-1	-1	-1	
3	4	1	0	1	1	1	-1	-1	-1	
4	5	1	0	-1	1	1	-1	1	1	

5 rows × 32 columns

```
In [9]: df.describe()
```

Out[9]:

	id	having_IP_Address	URL_Length	Shortining_Service	having_At_Symbol	double_slash_redirecting	Prefix_Suffix	having_Sub_Domain	SSL
count	11055.000000	11055.000000	11055.000000	11055.000000	11055.000000	11055.000000	11055.000000	11055.000000	110
mean	5528.000000	0.313795	-0.633168	0.738761	0.700588	0.741474	-0.734962	0.063953	
std	3191.447947	0.949534	0.766095	0.673998	0.713598	0.671011	0.678139	0.817518	
min	1.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000	
25%	2764.500000	-1.000000	-1.000000	1.000000	1.000000	1.000000	-1.000000	-1.000000	
50%	5528.000000	1.000000	-1.000000	1.000000	1.000000	1.000000	-1.000000	0.000000	
75%	8291.500000	1.000000	-1.000000	1.000000	1.000000	1.000000	-1.000000	1.000000	
max	11055.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	

8 rows × 32 columns

Figure 7.2.1 Dataset

7.3 Result

GUI Screenshot

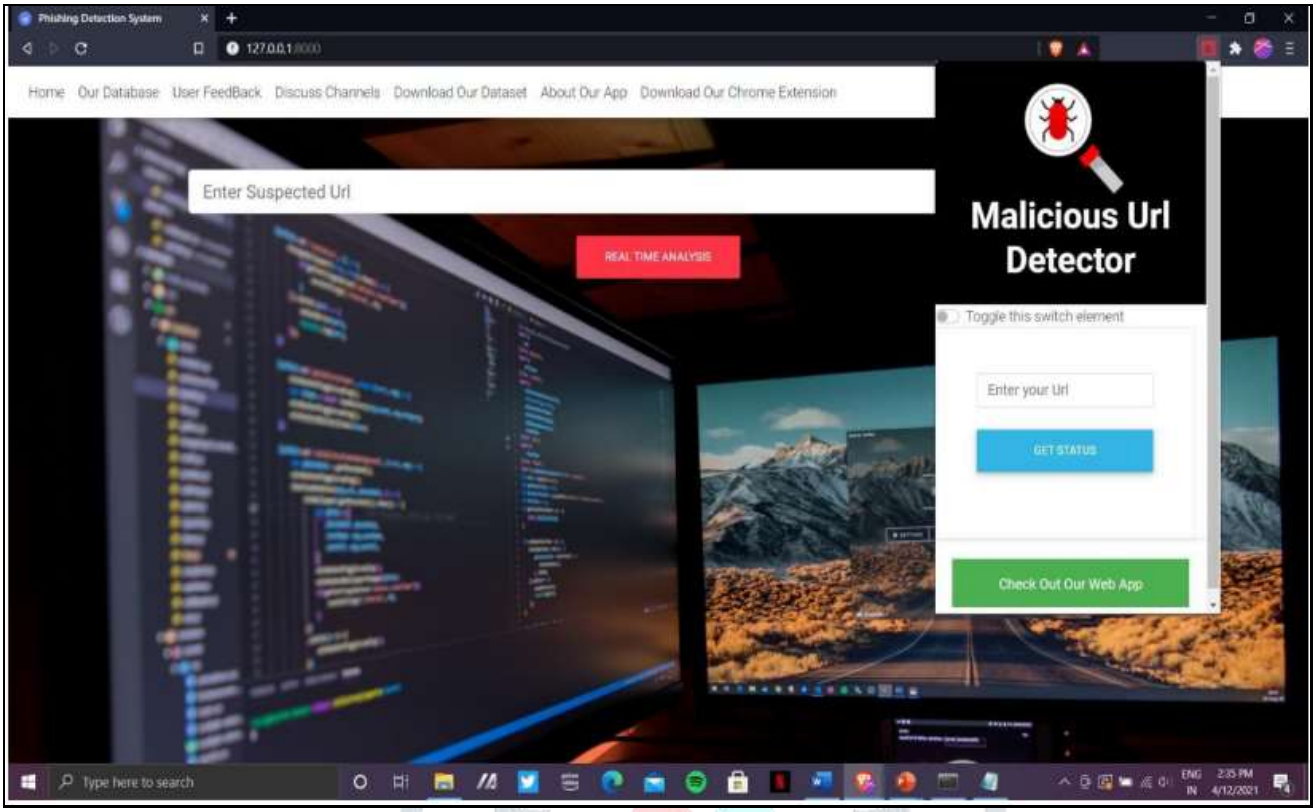


Figure 7.3.1 Extension**.**

Extension: The user can get the status of the URL by entering it into the extension. By enabling the extension, the user can also come to know about the site directly from the browser.

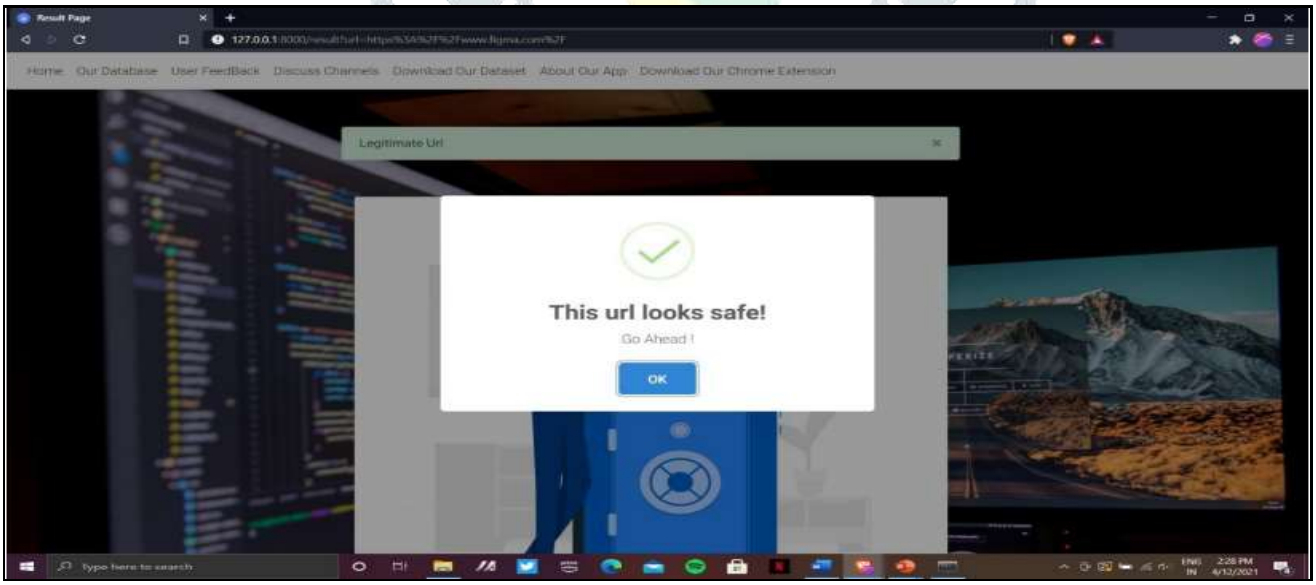


Figure 7.3.2 Legitimate URL.

Legitimate: The above figure will show status as legitimate if the website is safe to visit for the user.

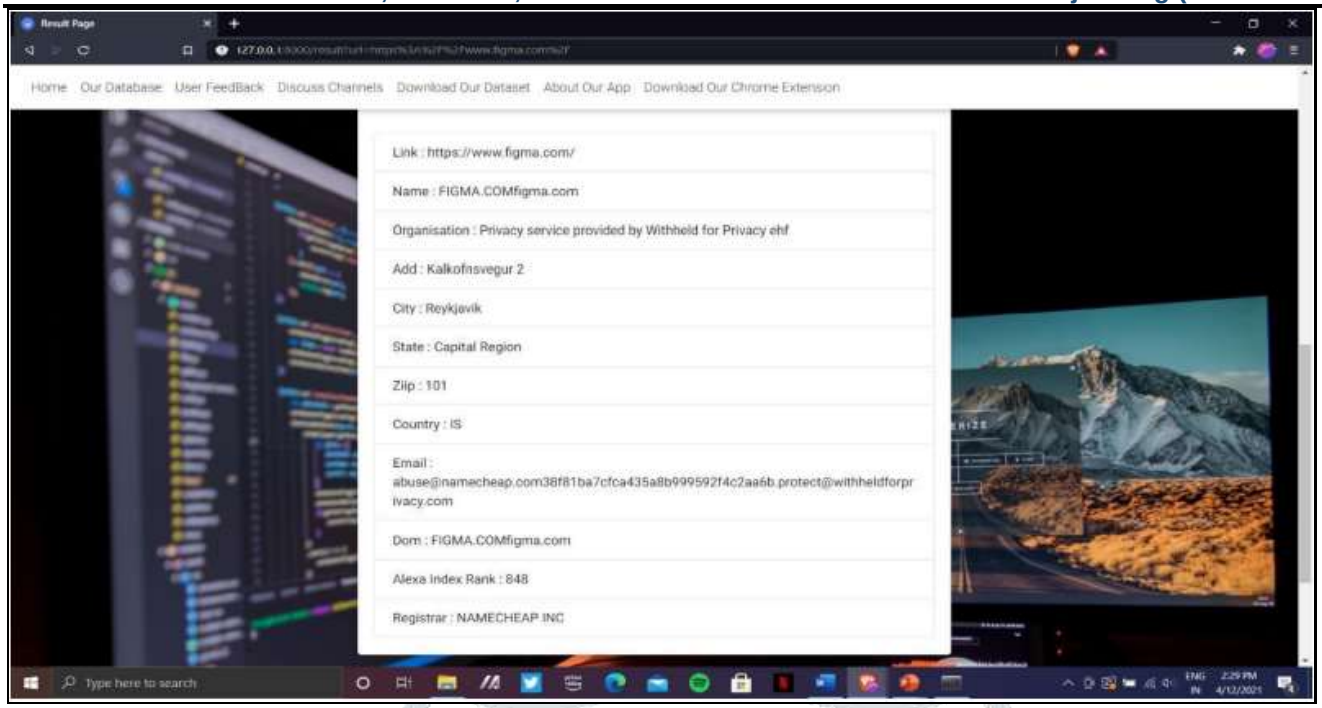


Figure 7.3.3 URL Information.

URL Information: The information of the website will be displayed such as name, organization, country, city, domain, email after the status of the URL.

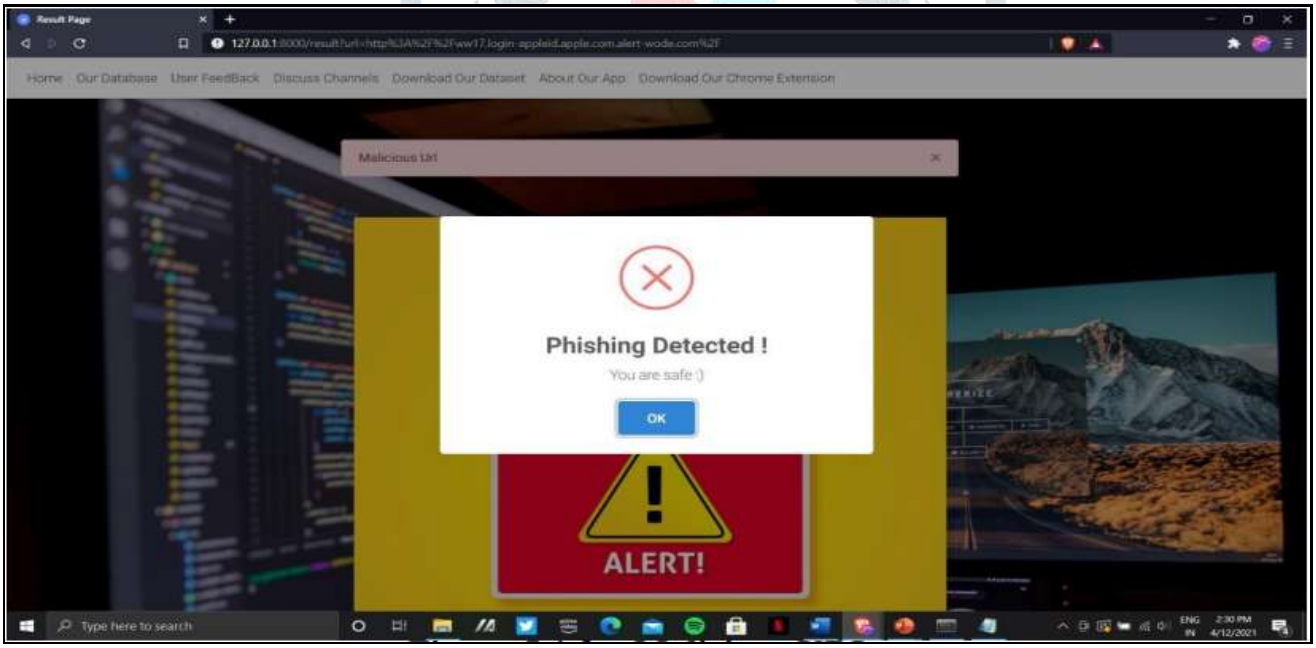


Figure 7.3.4 Malicious URL.

Malicious: The above figure will show status as malicious if the website is risky and dangerous to visit for the user.

7.4 Database Screenshot

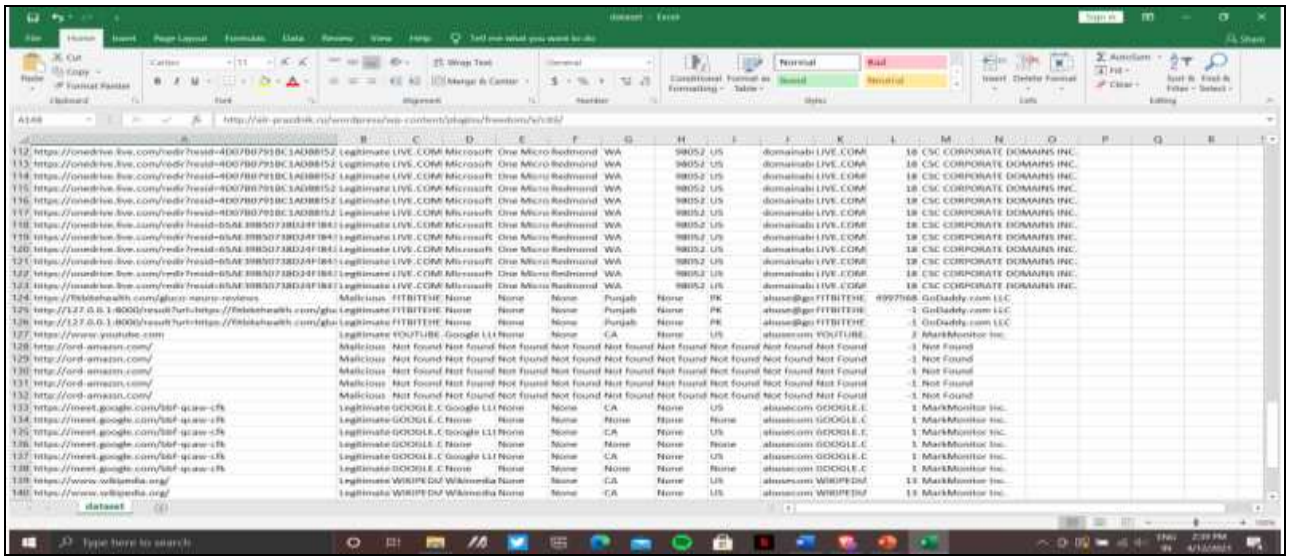


Figure 7.4.1 – Dataset Information.

7.5 Outcomes:

Our solution will detect malicious links and its origin signature (first uploaded person-profile URL, name, email, number etc.) on a real time basis through Chrome Extension and provide advisory reports to the public and corresponding agencies about those links source credibility. We determine whether a URL is Malicious or Legitimate using Machine Learning. Additionally, using Alexa ranking and domain registration for verifying results. Using Whois Database to get all the details of URLs.

VIII. FUTURE SCOPE

Attacks now rely on progressive forms of infiltration that better mask malicious intent. Today, phishing attacks can be difficult to detect, and grant malevolent individuals extensive permissions over user data and online services. Attacks now rely on innovative forms of penetration that better camouflage malicious intent. The means by which the threat actor access user information have quickly evolved beyond traditional phishing emails. Phishing has always had the goal of enticing users to take an action or share a piece of sensitive information by acting as a non-threat actor but awareness has since grown. Impulsive password reset emails, while once effective, no longer drives the same volume of user action and is often detected by spam filters.

Attacks now depend on on advanced forms of intrusion that better mask malicious intent. Phishing Detection System aims to control the security of data and to prevent breaches, to test whether junk mail is to be had from the present-day database, to enable the person to create his very own unwanted mail listing, and to test whether the arriving URL has risky content. This is the future of phishing. The ability to spoof cloud apps while masking the true identity of the sender in order to steal personal information – an alarming trend given the rapid increase of cloud adoption in verticals around the world. Phishing has always had the aim of baiting users to take an action or share a piece of sensitive information by appearing as a non-threat – but awareness has since grown. Unprompted password reset emails, while once effective, no longer drives the same volume of user action and is often detected by spam filters.

Attacks now rely on advanced forms of infiltration that better disguise malicious intent. Phishing Detection System aims to control the security of data and to prevent infringements, to test whether junk mail is to be had from the present-day database, to enable the person to create his very own unsolicited mail listing, and to test whether the incoming URL has risky content. The Gmail phishing attacks shows us just exactly how advanced these techniques have become as it is hard to detect and hard to prevent. A critical takeaway is that the attack was able to clear the mental trust hurdle. Users were deceived into giving approvals to a third party application because they trusted it; they believed the application to be a Google-approved service. A minute change in how the application domain was camouflaged successfully convinced users that the application was trustworthy. This is the future of phishing. The ability to spoof cloud apps while concealing the true identity of the sender in order to take personal information is an alarming trend given the rise of cloud adoption in verticals around the world.

IX. CONCLUSION

Phishing is becoming an ever-growing threat to users as the attacks advance and become more difficult to differentiate. The offenders who carry out these attacks are increasingly hard to catch. Nowadays it has emerged as Very serious. There are many methods to solve those Problems. But humans may also don't aware of the seriousness of phishing. Periodical updating of anti-phishing gear or software in their personal structures may be useful to relax their personal data and credentials. To combat these

challenges, we have proposed a three-Pronged approach project. In a Phishing detection system using Machine learning in Cyber world application, the use of a filtration system helps lessen the number of phishing emails that reach the user, reducing the chances that they will be phished. The user interface model provides users with warnings when the site they are visiting is not trusted, therefore defending against the chance that a convincing email has led them to a phishing site.

REFERENCES

JOURNAL REFERENCES

[1] P. Liu and T. S. Moh, "Content Based Spam Email Filtering," 2016 International Conference on Collaboration Technologies and Systems (CTS), Orlando, FL, pp. 218-224, 2016.

[2] N. Agrawal and S. Singh, "Origin (dynamic blacklisting) based spammer detection and spam mail filtering approach," 2016 Third International Conference on Digital Information Processing, Data Mining, and Wireless Communications (DIPDMWC), Moscow, pp. 99-104,2016.

[3] J. V. Chandra, N. Challa and S. K. Pasupuleti, "A practical approach to E-mail spam filters to protect data from advanced persistent threat," 2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT), Nagercoil, pp. 1-5, 2016.

[4] A. K. Sharma and R. Yadav, "Spam Mails Filtering Using Different Classifiers with Feature Selection and Reduction Technique," 2015 Fifth International Conference on Communication Systems and Network Technologies, Gwalior, pp. 1089-1093, 2015.

[5] Greg Aaron APWG and Ronnie Manning. 2019. APWG Phishing Activity Trends Report, 1st Quarter 2019

[6] Qian Cui, Guy-Vincent Jourdan, Gregor V Bochmann, Russell Couturier, 2017. Tracking phishing attacks over time. In Proceedings of the 26th International Conference on WorldWide Web. International World.

[7] Rachna Dhamija and J Doug Tygar. 2005. The battle against phishing: Dynamic security skins. In Proceedings of the 2005 symposium on Usable

[8] Mahmoud Khonji, Youssef Iraqi, and Andrew Jones. 2013. Phishing detection: a literature 45 survey. IEEE Communications Surveys & Tutorials 15, 4 (2013),

WEB REFERENCES

[1] Phishing URL Detection with ML-Ebubekir Büber – https://towardsdatascience.com/phishing-domain-detection-with-ml-5be9c99293e5.

[2] Phishing Detection and Prevention- Steve Davidson- https://www.edgewave.com/solutions/phishing/

[3] Phishing Attack Prevention: How to Identify & Avoid Phishing Scams in 2019- https://digitalguardian.com/blog/phishing-attack-prevention-how-identify-avoid- phishing-scams survey.

[4] Christina Bonnington. 2018. Twitter is promoting a 'get verified' phishing scam. - https://www.dailydot.com/debug/twitter-promoted-phishing-site/.

[5] Safe Browsing protection from even more deceptive attacks.- https://security.googleblog.com/2015/11/safe-browsing-protection-from-even-more.html

[6] OpenPhish. 2018. OpenPhish - Phishing Intelligence. https://openphish.com/.

[7]Anti Phishing Work Group (2014) Phishing attacks trends report.
http://docs.apwg.org/reports/apwg_trends_report_q2_2014.pdf