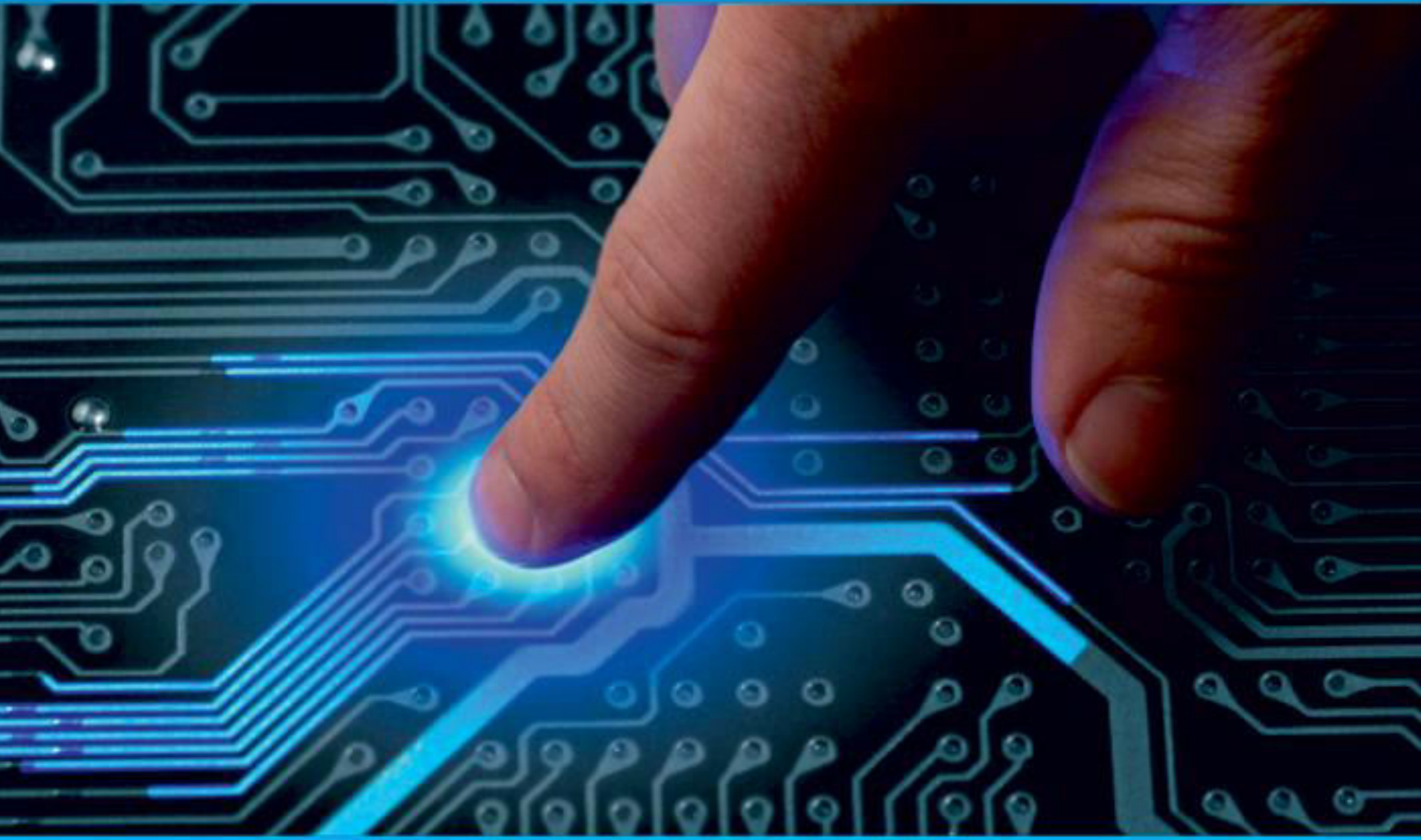
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 2, February 2023

INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 8.165

Phishing URL Detection to Avoid Intruders with ML

Dr. Deepali Sale[1], Prof. Anita Shinkar[1], Sanskruti A. Agale[2], Arya K. Aglawe[2], Pranav P. Bansod[2], Sakshi P. Chaudhari [2]

Professor, Department of Computer Technology, Dr. D.Y.Patil College of Engineering & Innovation Talegaon Pune, India[1]

Student, Department of Computer Technology, Dr. D.Y.Patil College of Engineering & Innovation Talegaon Pune, India[2]

ABSTRACT: Internet users have suffered from phishing attacks for a long time. Attackers deceive users through malicious constructed phishing websites to steal sensitive information, such as bank account numbers, website usernames, and passwords. In recent years, many phishing detection solutions have been proposed, which mainly leverage whitelists or blacklists, website content, or side channel-based techniques. However, with the continuous improvement of phishing technology, current methods have difficulty in achieving effective detection. Hence, in this paper, we propose an effective phishing website detection approach, which we call HinPhish. HinPhish extracts various link relationships from webpages and uses domains and resource objects to construct a heterogeneous information network. HinPhish applies a modified algorithm to leverage the characteristics of different link types in order to calculate the phish-score of the target domain on the webpage. Moreover, HinPhish not only improves the accuracy of detection, but also can increase the phishing cost for attackers. Extensive experimental results demonstrate that HinPhish can achieve an accuracy of 0.9856 and F1- score of 0.9858.

KEYWORDS: Malicious domain detection; phishing; Heterogeneous information network

I. INTRODUCTION

Collaborative filtering is a technology to recommend items based on similarity. There are two types of collaborative filtering: User-based collaborative filtering and Item-based collaborative filtering. User-based collaborative filtering algorithm is an effective way of recommending useful contents to users by exploiting the intuition that a user will likely prefer the items preferred by similar users. Therefore, at first, the algorithm tries to find the user's neighbours based on user similarities and then combines the neighbour user's rating score by using supervised learning like genetic algo. Item-based collaborative filtering algorithm fundamentally has the same scheme with user-based collaborative filtering in terms of using user's rating score. Instead of the nearest neighbours, it looks into a set of items; the target user has already rated items and this algorithm computes how similar items are to the target item under recommendation. After that it also combines the customer's previous preferences based on these item similaritie To enhance the customer experience and to boost up the sales of products, almost all of the companies are trying to make some sort of mechanism that is nothing but a recommendation system. So to finalize this task recommender system comes into the light. The system works in two steps, first, it analyses the user search for an item and simply user interests, and secondly, it tries to find a similar set of items that the user may be interested in. This in turns lead to better choices among the products.

II. LITERATURE SURVEY

Phishing is a very common and dangerous attack campaign. With the popularization of network technology, the cost of phishing attacks has been reduced significantly. Phishers can deploy a mimic website with low cost and high efficiency by utilizing various phish kits [1]. Then, these websites convince victims to leak their sensitive credentials, such as passwords, phone numbers, bank accounts, credit card numbers, and other private information. Phishing attacks have been become one of the major cybersecurity threats. They may not only cause the victim to suffer economic losses, but can also destroy the reputation of the imitated website. According to the report published by the Anti-Phishing Working Group (APWG) [2], after experiencing a rapid growth in 2020 (doubling over the course of the year), the number of phishing attacks in January 2021 reached its peak in the APWG's records, as shown in Figure 1. The trend of phishing attacks is becoming worse.

III. PROBLEM DEFINITION

The Internet has become an indispensable part of our life, However, It also has provided opportunities to anonymously perform malicious activities like Phishing. Phishers try to deceive their victims by social engineering or creating mock-up websites to steal information such as account ID, username, password from individuals and organizations.

IV. ARCHITECTURE DIAGRAM

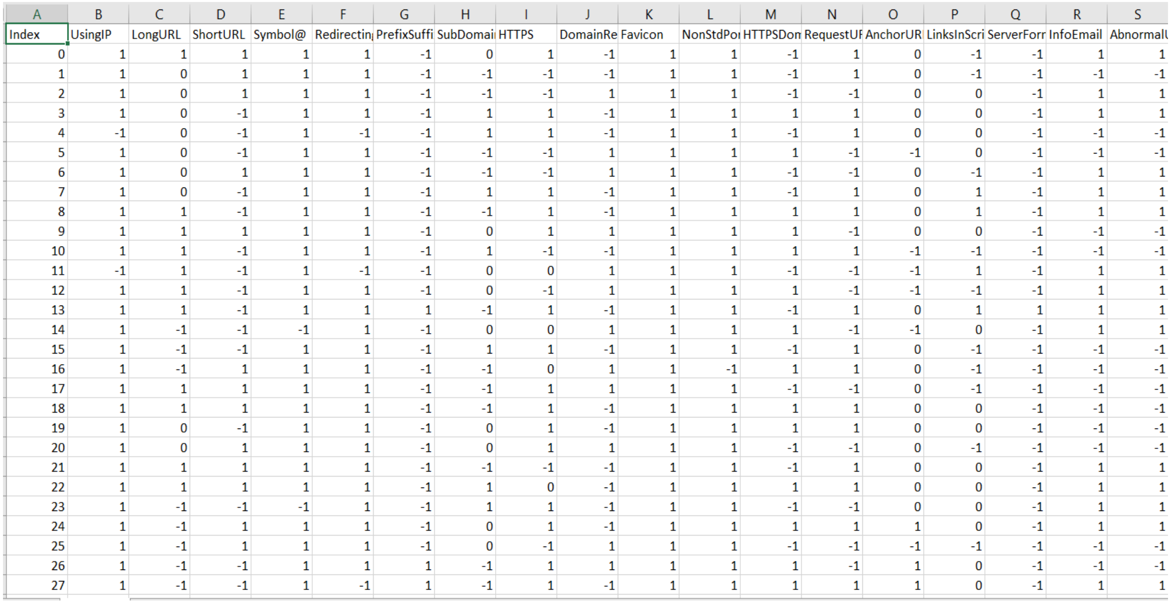
Dataset: It contains features from 1353 URLs. Out of these, 548 are legitimate, 702 are phishing, and 103 are suspicious. The data set also contains nine features that were extracted from each URL. The attributes provide information such as the URL anchor, popup window, age of the domain, URL length, IP address, web traffic, etc. Each feature value holds categorical values, either binary or ternary.

Decision tree: Decision trees are non-parametric classifiers. As its name indicates, a decision tree is a tree structure, where each non-terminal node denotes a test on an attribute, each branch represents an outcome of the test, and the leaf nodes denote classes. The basic algorithm for decision tree induction is a greedy algorithm that constructs the decision tree in top-down recursive divide-and-conquer manner.
Step 1: Initialize weights with small random values.

Step 2: Present an input vector and make a forward pass to compute weighted sums i S and activations of a S i i= ( ) for each unit, where f (.) represents the activation function.
Step 3: Backpropagation: Starting with the output units, make a backward pass through output units and hidden layer units.

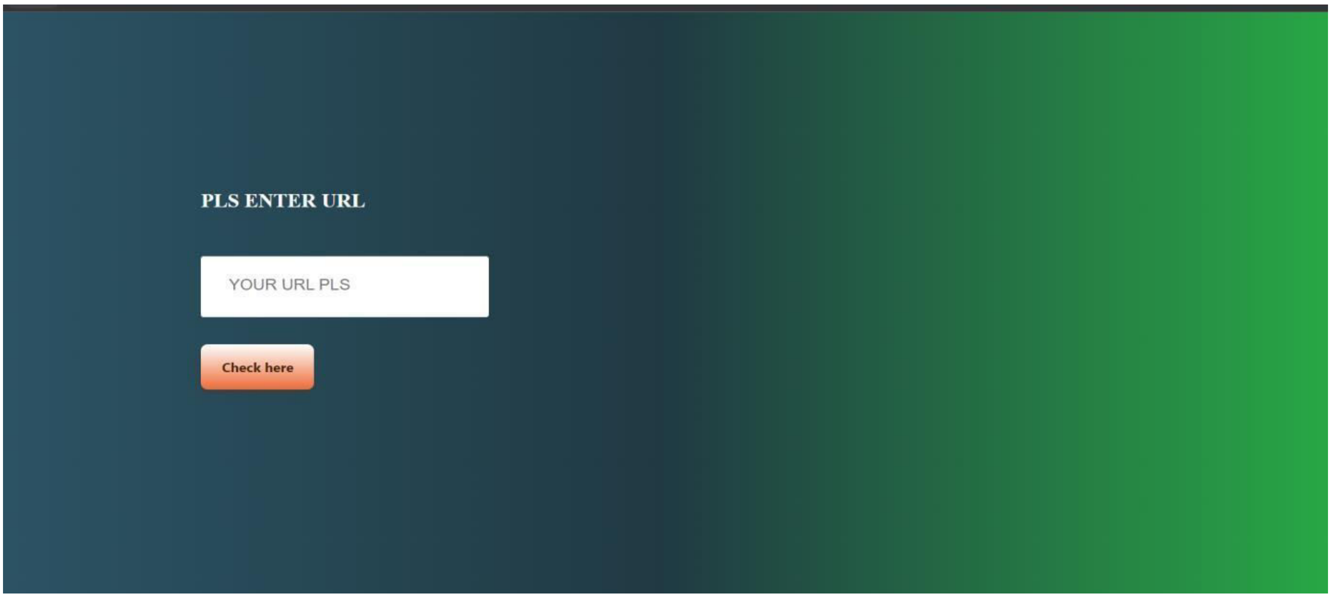
V. RESULT

DATASET:

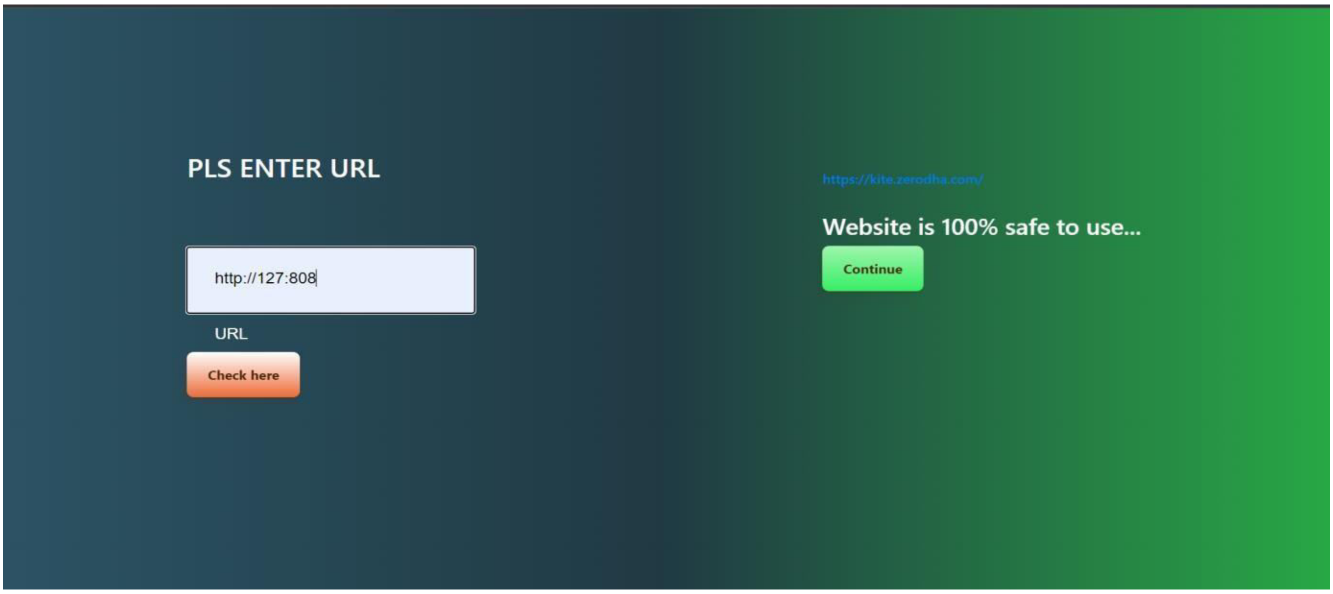


A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
Index	UsingIP	LongURL	ShortURL	Symbol@	Redirectin	PrefixSuffi	SubDomai	HTTPS	DomainRe	Favicon	NonStdPo	HTTPSDom	RequestUF	AnchorUR	LinksInScri	ServerForr	InfoEmail	AbnormalU
0	1	1	1	1	1	-1	0	1	-1	1	1	-1	1	0	-1	-1	1	1
1	1	0	1	1	1	-1	-1	-1	-1	1	1	-1	1	0	-1	-1	-1	-1
2	1	0	1	1	1	-1	-1	-1	1	1	1	-1	-1	0	0	-1	1	1
3	1	0	-1	1	1	-1	1	1	-1	1	1	1	1	0	0	-1	1	1
4	-1	0	-1	1	-1	-1	1	1	-1	1	1	-1	1	0	0	-1	-1	-1
5	1	0	-1	1	1	-1	-1	-1	1	1	1	1	-1	-1	0	-1	-1	-1
6	1	0	1	1	1	-1	-1	-1	1	1	1	-1	-1	0	-1	-1	1	1
7	1	0	-1	1	1	-1	1	1	-1	1	1	-1	1	0	1	-1	1	1
8	1	1	-1	1	1	-1	-1	1	-1	1	1	1	1	0	1	-1	1	1
9	1	1	1	1	1	-1	0	1	1	1	1	1	-1	0	0	-1	-1	-1
10	1	1	-1	1	1	-1	1	-1	-1	1	1	1	1	-1	-1	-1	-1	-1
11	-1	1	-1	1	-1	-1	0	0	1	1	1	-1	-1	-1	1	-1	1	1
12	1	1	-1	1	1	-1	0	-1	1	1	1	1	-1	-1	-1	-1	1	1
13	1	1	-1	1	1	1	-1	1	-1	1	1	-1	1	0	1	1	1	1
14	1	-1	-1	-1	1	-1	0	0	1	1	1	1	-1	-1	0	-1	1	1
15	1	-1	-1	1	1	-1	1	1	-1	1	1	-1	1	0	-1	-1	-1	-1
16	1	-1	1	1	1	-1	-1	0	1	1	-1	1	1	0	-1	-1	-1	-1
17	1	1	1	1	1	-1	-1	1	1	1	1	-1	-1	0	-1	-1	-1	-1
18	1	1	1	1	1	-1	-1	1	-1	1	1	1	1	0	0	-1	-1	-1
19	1	0	-1	1	1	-1	0	1	-1	1	1	1	1	0	0	-1	-1	-1
20	1	0	1	1	1	-1	0	1	1	1	1	-1	-1	0	-1	-1	-1	-1
21	1	1	1	1	1	-1	-1	-1	-1	1	1	-1	1	0	0	-1	1	1
22	1	1	1	1	1	-1	1	0	-1	1	1	1	1	0	0	-1	1	1
23	1	-1	-1	-1	1	-1	1	1	-1	1	1	-1	-1	0	0	-1	1	1
24	1	-1	1	1	1	-1	0	1	-1	1	1	1	1	1	0	-1	1	1
25	1	-1	1	1	1	-1	0	-1	1	1	1	-1	-1	-1	-1	-1	-1	-1
26	1	-1	-1	1	1	1	-1	1	1	1	1	1	-1	1	0	-1	-1	-1
27	1	-1	-1	1	-1	1	-1	1	-1	1	1	1	1	1	0	-1	1	1

Homepage:



Phishing Websites:

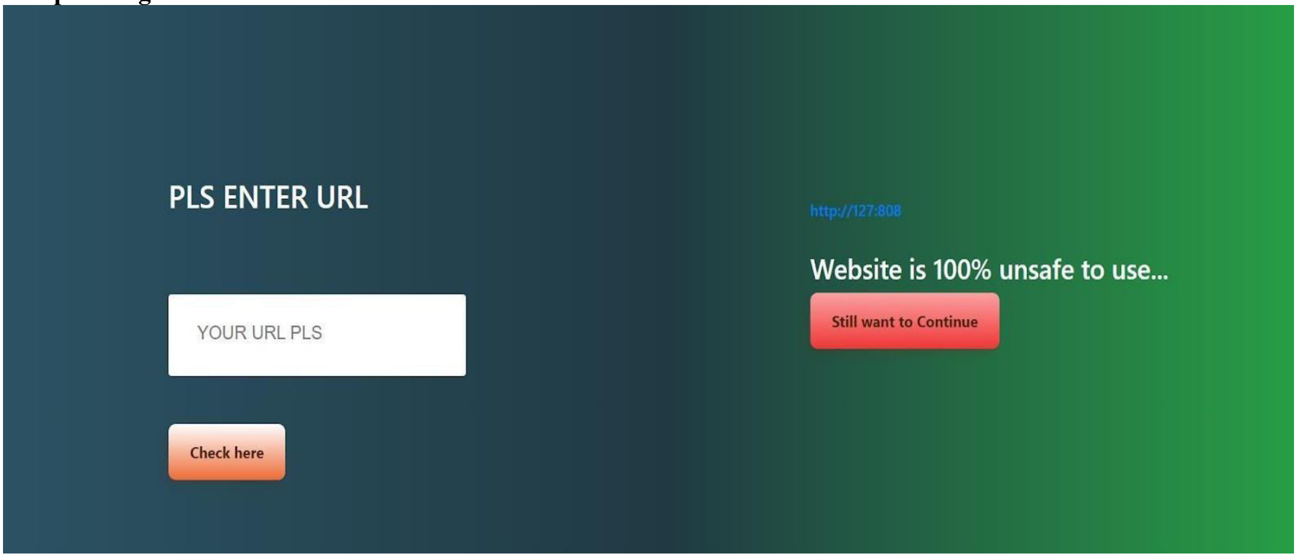


http://127:808

https://kite.zerodha.com/

Website is 100% safe to use...

Non phishing Websites:



PLS ENTER URL

YOUR URL PLS

Check here

http://127:808

Website is 100% unsafe to use...

Still want to Continue

Diagram:

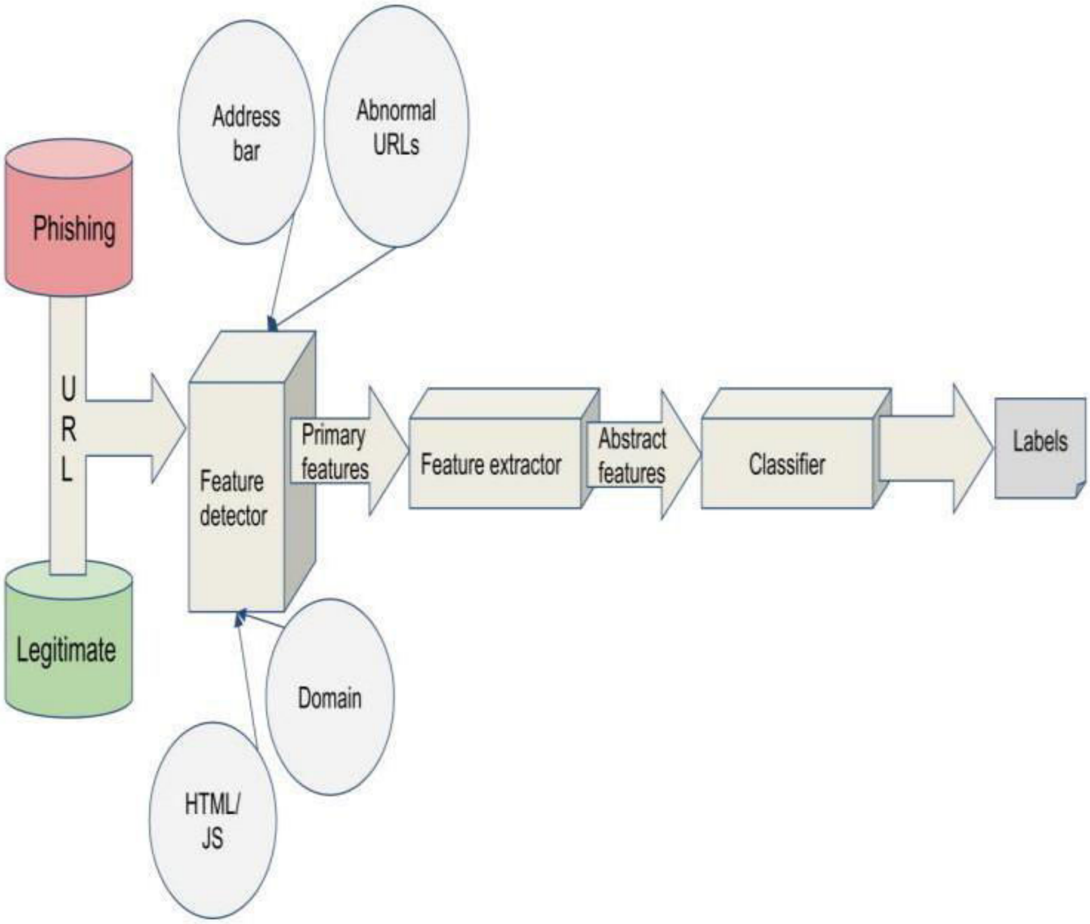


Fig 1. Architecture Diagram

VI. CONCLUSION

Phishing is a very common and dangerous attack campaign. With the popularization of network technology, the cost of phishing attacks has been reduced significantly. Phishers can deploy a mimic website with low cost and high efficiency by utilizing various phish kits [1]. Then, these websites convince victims to leak their sensitive credentials, such as passwords, phone numbers, bank accounts, credit card numbers, and other private information. Phishing attacks have been become one of the major cybersecurity threats. They may not only cause the victim to suffer economic losses, but can also destroy the reputation of the imitated website. According to the report published by the Anti- Phishing Working Group (APWG) [2], after experiencing a rapid growth in 2020 (doubling over the course of the year), the number of phishing attacks in January 2021 reached its peak in the APWG's records, as shown in Figure 1. The trend of phishing attacks is becoming worse.

REFERENCES

1. N. Lord, "What is a Phishing Attack? Defining and Identifying Different Types of Phishing Attacks". https://digitalguardian.com/blog/whatphishing-attack-defining-and-identifying-different- types-phishingattacks, 2018.
2. Vemula VR. Adaptive Threat Detection in DevOps: Leveraging Machine Learning for Real-Time Security Monitoring. International Machine learning journal and Computer Engineering. 2022 Nov 17;5(5):1-7.
3. N. Sadeh, A. Tomasic, and I Fette, "Learning to detect phishing emails", Proceedings ofthe16thinternational conference on world wide web, pp.649–656, 2007.
4. Srinivasa Rao Thumala. (2022), "Importance of Business Continuity and Disaster Recovery (BCDR) Methodologies for Organizations: A Comparison Study between AWS and Azure". International Journal of Science and Research (IJSR), 11(12): 1406-1415.
5. J. Ma, S. S. Savag, G. M. Voelker, "Learning to detect malicious URLs", ACM Transactions on Intelligent Systems and technology, vol. 2, no. 9, pp 30:1-30:24, 2011.
6. S. Purkait, "Phishing counter measures and their effectiveness–literature review", Information Management & Computer Security, vol. 20, no. 5, pp. 382–420, 2012.
7. N. Abdelhamid, A. Ayesh, F. Thabtah, "Phishing Detection based Associative Classification", Data Mining. Expert Systems with Applications (ESWA), vol. 41, pp 5948-5959, 2014.

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.165

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

☎ 9940 572 462 6381 907 438 ✉ ijircce@gmail.com

Scan to save the contact details

www.ijircce.com