

Malicious URL Detection using Machine Learning

Prof. Rajeshree Sonawale¹, Tanmay Khedekar², Amar Kamble³, Juhi Kumavat⁴

Professor, Department of Computer Engineering¹

Students, Department of Computer Engineering^{2,3,4}

Mahatma Gandhi College of Engineering And Technology, Navi Mumbai, India

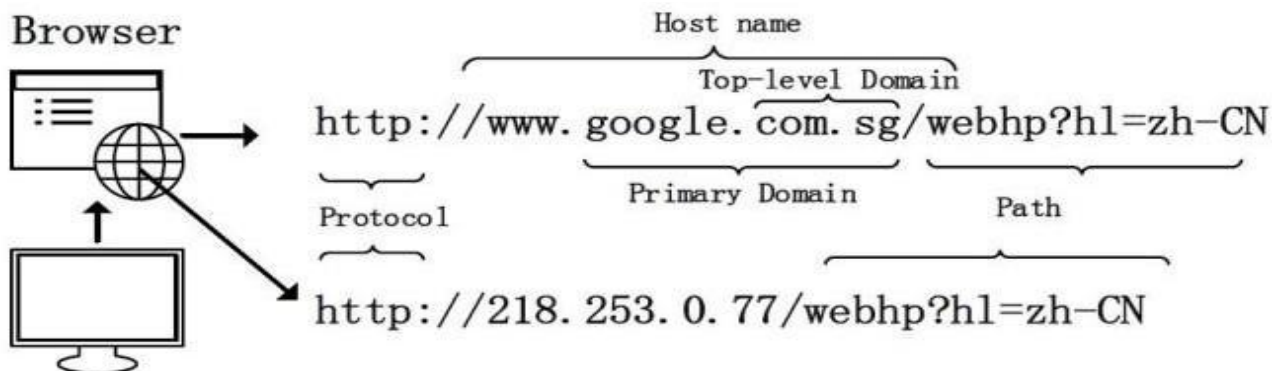
Abstract: Currently, the risk of network information insecurity is increasing rapidly in number and level of danger. The methods mostly used by hackers today is to attack end-to-end technology and exploit human vulnerabilities. These techniques include social engineering, phishing, pharming, etc. One of the steps in conducting these attacks is to deceive users with malicious Uniform Resource Locators (URLs). As a results, malicious URL detection is of great interest nowadays. There have been several scientific studies showing a number of methods to detect malicious URLs based on machine learning and deep learning techniques. In this paper, we propose a malicious URL detection method using machine learning techniques based on our proposed URL behaviors and attributes.

Keywords: Machine Learning, URLs, Malicious, Phishing

I. INTRODUCTION

Uniform Resource Locator (URL) is used to refer to resource on the Internet. In [1], Sahoo et al. Presented about the characteristics and two basic components of the URL as: protocol identifier, which indicates what protocol to use, and resource name, which specifies the IP address or the domain name where the resource is located, it can be seen that each URL has a specific structure and format. Attacker often try to change one or more components of the URL's structure to deceive user for spreading their malicious URL. Malicious URLs are known as links that adversely affect users. These URLs will redirect users to resources or pages on which attackers can execute codes on user's computers, redirect user to unwanted sites, malicious website, or another phishing site, or malware download. Malicious URLs can also be hidden in download links that are deemed safe and can spread quickly through file and message sharing in shared networks. Some attack techniques that use malicious URLs include [2,3,4]: Drive -by Download, Phishing and Social Engineering, and Spam..

1.1 URL Structure and Format

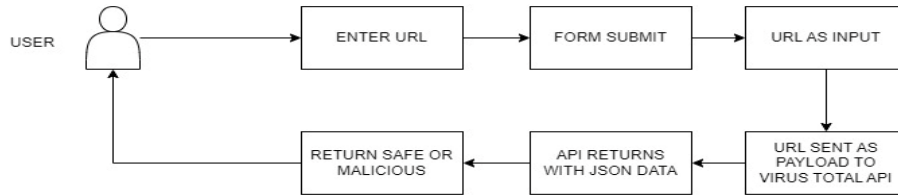


1.2 Data Flow Diagram of Malicious URL Detection

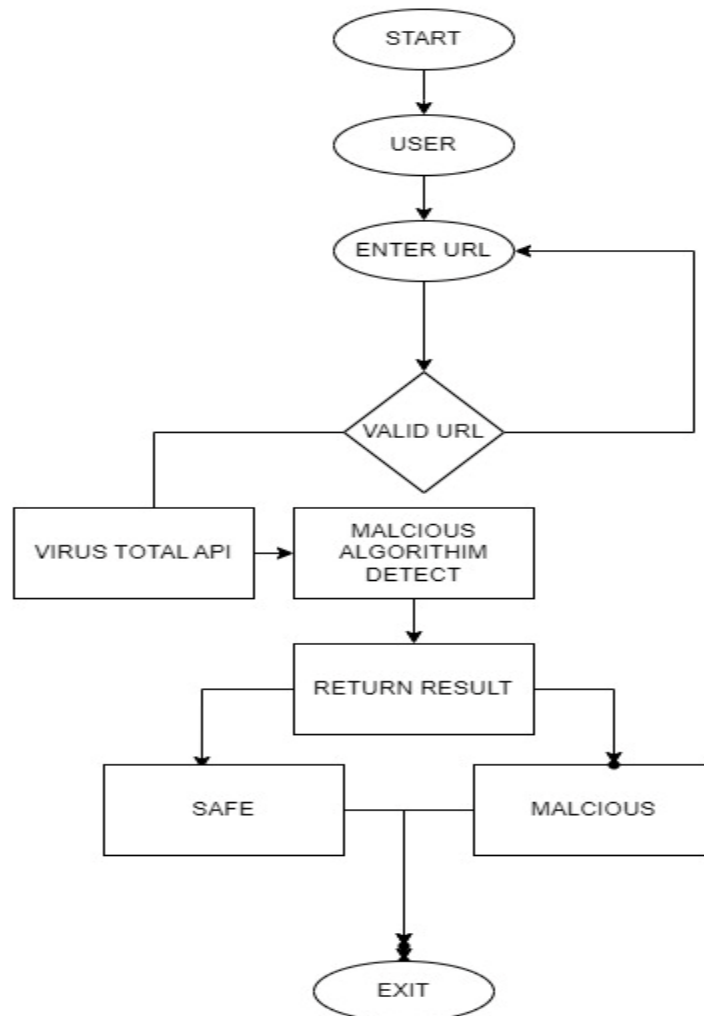
Regarding the problem of detecting malicious URLs, there are two main trends at present as malicious URL detection based on signs or sets of rules, and malicious URL detection based on behavior analysis techniques [1, 2]. The method of detecting malicious URLs based on a set of markers or rules can quickly and accurately detect malicious URLs. However, this method is not capable of detecting new malicious URLs that are not in the set of predefined signs or rules. The method of detecting malicious URLs based on behavior analysis techniques adopt machine learning or deep

learning algorithms to classify URLs based on their behaviors. In this paper, machine learning algorithms are utilized to classify URLs based on their attributes. The paper also includes a new URL attribute extraction method.

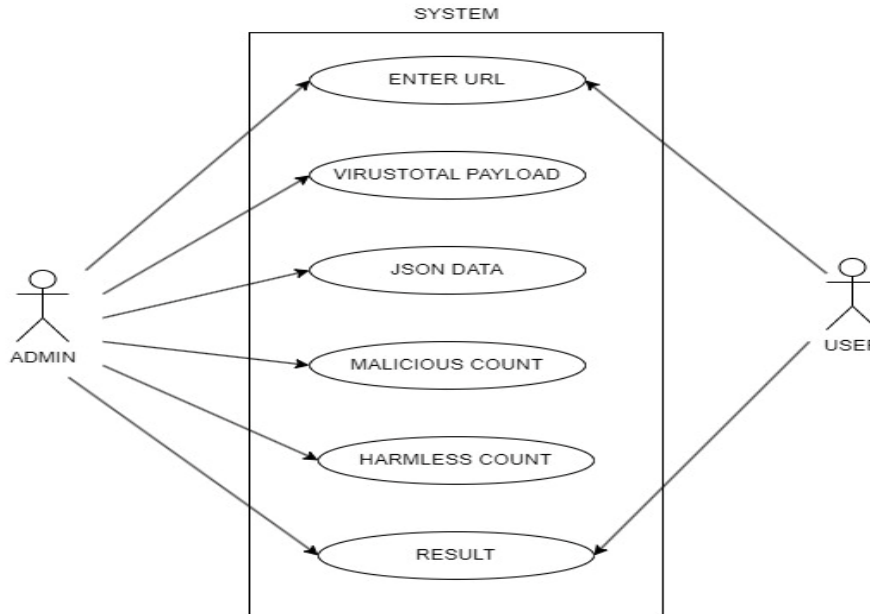
Data Flow Diagram -1



Data Flow Diagram -2



Data Flow Diagram -3 (User Case Diagram)



II. LITERATURE SURVEY

Title Name	Year	Publisher	Description
Phishing URL Detection: A Real-Case Scenario Through Login URLs	2022	IEEE	Phishing detection mechanism aims to improve current blacklist methods, protecting users from malicious login forms.
Malicious URL Detection using Logistic Regression	2022	IJRES	To check that if the website are malicious or not.

III. PROPOSED MACHINE LEARNING -BASED SOLUTION FOR MALICIOUS URL DETECTION

There are two- types of features that can be used - static features, and dynamic features. In static analysis, we perform the analysis of a webpage based on information available without executing the URL. The features extracted include lexical features from the URL string, information about the host, and sometimes even HTML and JavaScript content. Since no execution is required, these methods are safer than the Dynamic approaches. The underlying assumption is that the distribution of these features is different for malicious and benign URLs.

A. Signature based Malicious URL Detection:-

Studies on malicious URL detection using the signature sets had been investigated and applied long time ago [6, 7, 8]. Most of these studies often use lists of known malicious URLs. Whenever a new URL is accessed, a database query is executed. If the URL is blacklisted, it is considered as malicious, and then, a warning will be generated; otherwise, URLs will be considered as safe. The main disadvantage of this approach is that it will be very difficult to detect new malicious URLs that are not in the given list

B. Malicious URL Detection Tools:-

URL Void: URL Void is a URL checking program using multiple engines and blacklists of domains. Some examples of URL Void are Google Safe Browsing, Norton Safe Web and MyWOT. The advantage of the Void URL tool is its compatibility with many different browsers as well as it can support many other testing services. The main

disadvantage of the Void URL tool is that the malicious URL detection process relies heavily on a given set of signatures.

IV. IMPLEMENTATION DETAILS

The malicious URL detection model using machine learning contains two stages: training and detection.

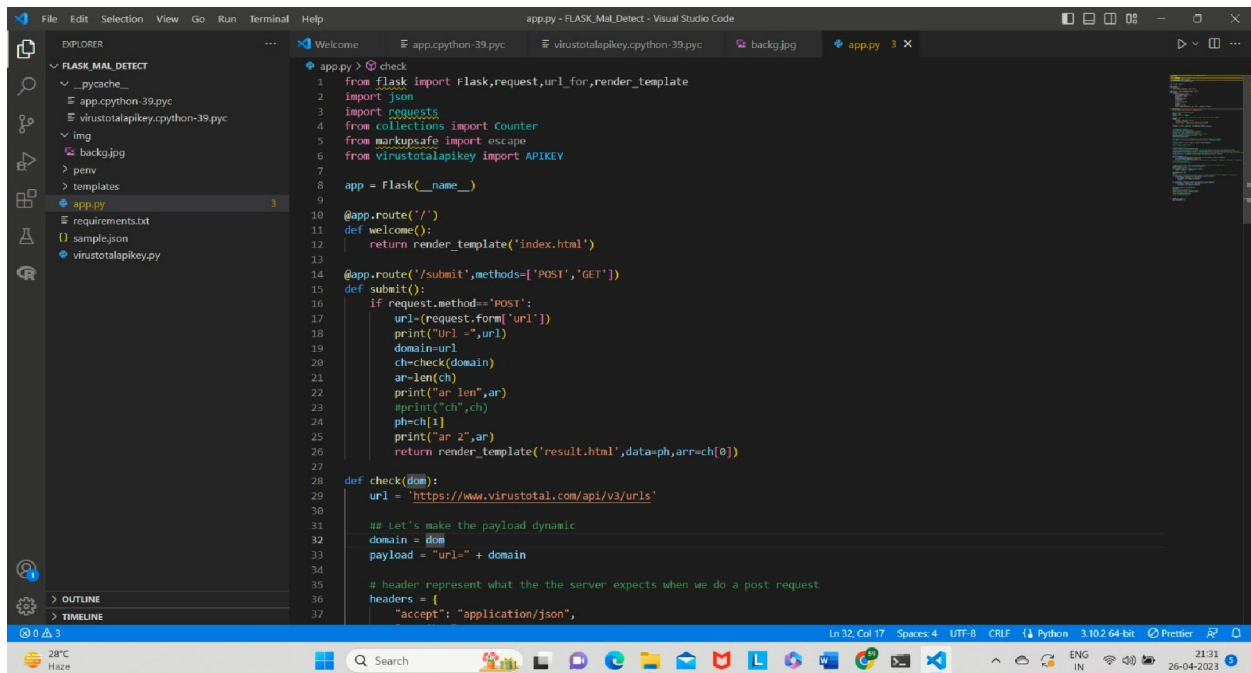
Training stage: To detect malicious URLs, it is necessary to collect both malicious URLs and clean URLs. Then, all the malicious and clean URLs are correctly labeled and proceeded to attribute extraction. These attributes will be the best basis for determining which URLs are clean and which are malicious. Details of these attributes will be presented in details in this paper. Finally, this dataset is divided into 2 subsets: training data used for training machine learning algorithms, and testing data used for testing process. If the classification performance of the machine learning model is good (high classification accuracy), the model will be used in the detection phase.

Detection phase: The detection phase is performed on each input URL. First, the URL will go through attribute extraction process. Next, these attributes are input to the classifier to classify whether the URL is clean or malicious.

- **Lexical features:** These features include URL length, main domain length, maximum token domain length, path average length, average token length in domain.
- **Host-based Features:** These features are extracted from the host characteristics of the URLs. These attributes indicate the location of malicious servers, the identity of malicious servers, the degree of impact of several host-based features that contribute the URL's malicious level.
- **Content-based Features:** These features are acquired when a whole web page is downloaded. The workload of these features is quite heavy, since a lot of information needs to be extracted, and there may be security concerns about accessing that URL. However, with more information available about a particular site, it is expected to create a better prediction model. The content-based features of a website can be extracted primarily from its **HTML** content and the use of **PYTHON**.

V. OUTPUT

Malicious URL Detection Code

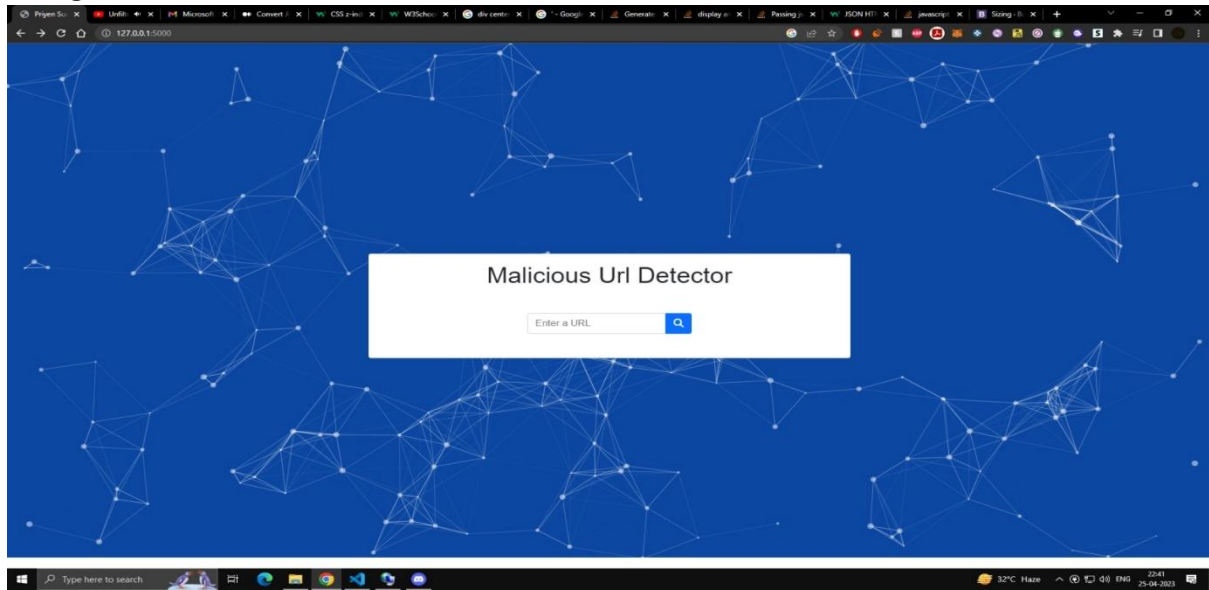


```

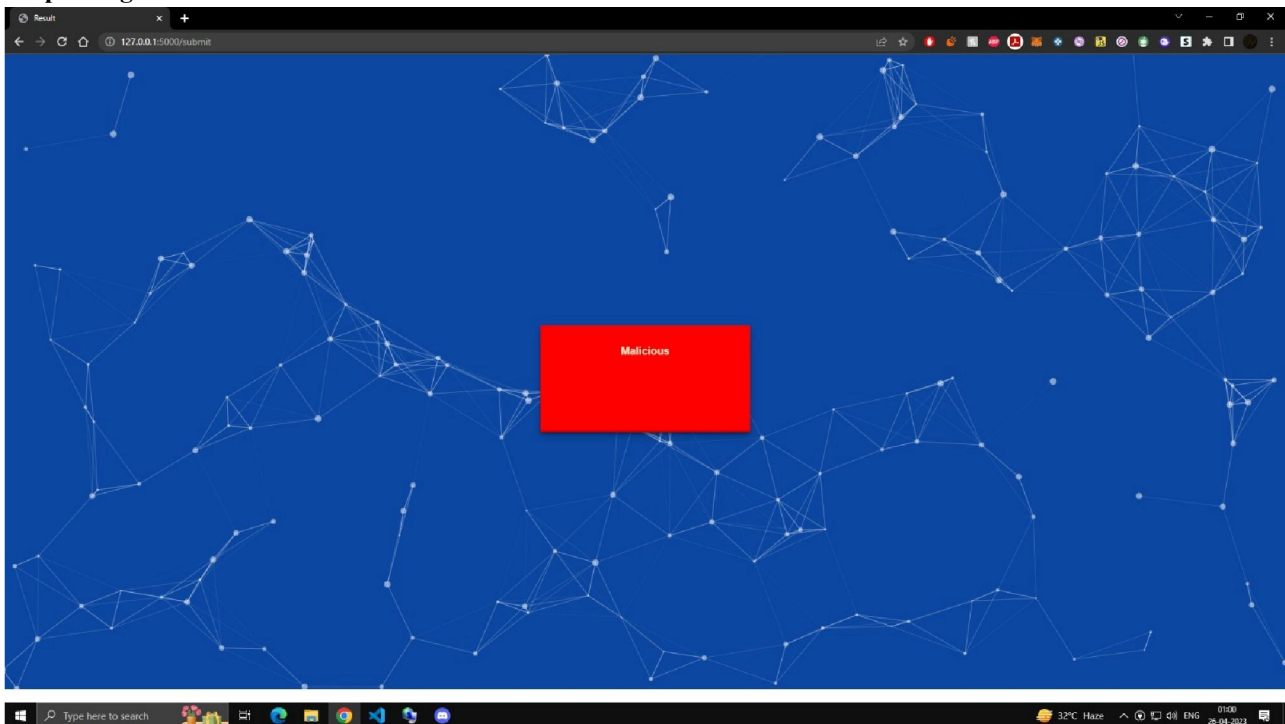
1 from flask import Flask, request, url_for, render_template
2 import json
3 import requests
4 from collections import Counter
5 from markupsafe import escape
6 from virustotalapikey import APIKEY
7
8 app = Flask(__name__)
9
10 @app.route('/')
11 def welcome():
12     return render_template('index.html')
13
14 @app.route('/submit', methods=['POST', 'GET'])
15 def submit():
16     if request.method == 'POST':
17         url = (request.form['url'])
18         print("url -", url)
19         domain = url
20         ch = check(domain)
21         ar = len(ch)
22         print("ar len", ar)
23         #print("ch", ch)
24         ph = ch[1]
25         print("ar 2", ar)
26         return render_template('result.html', data=ph, arr=ch[0])
27
28 def check(dom):
29     url = 'https://www.virustotal.com/api/v3/urls'
30
31     # Let's make the payload dynamic
32     domain = dom
33     payload = "url=" + domain
34
35     # header represent what the server expects when we do a post request
36     headers = {
37         "accept": "application/json",

```

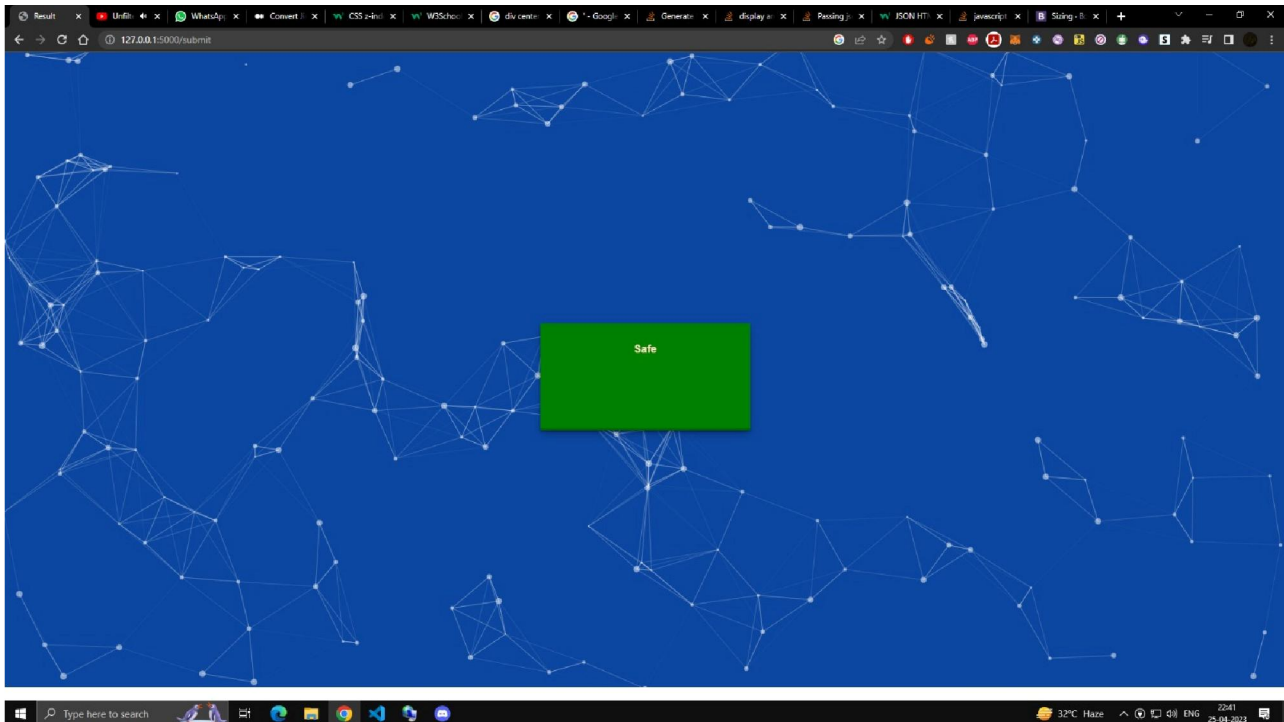
Home Page:-



Output Page:



OUTPUT SHOWING URL IS MALICIOUS



OUTPUT SHOWING URL IS HARMLESS

VI. DISCUSSION

High Scalability

Increasing huge amounts of URLs, a real-world malicious URL detection system must scale up for training the models of training data (millions or billions). To achieve the high scalability desire, first explore more efficient and scalable algorithms and then the second is to build scalable learning systems in distributed computing environments:

Strong Adaptation

A real-world malicious URL detection system has to deal with a variety of practical complexity, including adversarial patterns such as concept drifting where the distribution of malicious URLs may change over time or even change in adversarial way to bypass the detection system, missing values, increasing number of new features, etc. A real-world malicious URL detection system must have a strong adaptation ability to work Effectively and robustly under most circumstances.

High Accuracy

This is one of the important goals to be achieved for any malicious URL detection. We want to maximize the detection of all the threats of by minimizing the detection of classifying benign URLs into malicious. Since no system is capable of perfect detection accuracy it has to differentiate between the ratios of benign and malicious by setting different levels of detection thresholds

VII. CONCLUSION

In this we showed a broad and organized study on malicious URL detection using machine learning techniques. We also presented an efficient design of malicious URL detection from machine learning perspective and then later we analyze the existing system for malicious URL detection particularly In the form of developing new feature representation and designing the new learning algorithms for determining malicious URL detection task. We also identify the requirements and challenges for developing malicious URL detection as a service of real-world cybersecurity applications.

VIII. ACKNOWLEDGEMENT

We would like to express our gratitude to the **M.GM College of Engineering and Technology** for providing us with the necessary resources to conduct this research. We would also like to thank **Prof. Rajashree Sonawale** maam for her guidance and support throughout the project. Additionally, we are grateful to project coordinator **Prof. S.P. Vidya Bharde** ma'am, Head of the Computer Department, and all other faculty members who provided us with valuable insights and feedback. Finally, we extend our thanks to all the participants who willingly contributed their time and data to this study.

REFERENCES

- [1]. Phishing URL Detection: A Real-Case Scenario Through Login URLs (April 18 IEEE) from <https://ieeexplore.ieee.org/document/9759382>
- [2]. Malicious URL Detection using Logistic Regression (2022) from International Journal of Research in Engineering and Science (IJRES)
- [3]. Malicious URL Detection based on Machine Learning (2020) from (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 11, No. 1, 2020.
- [4]. Malicious URL Detection using Machine Learning: A Survey from (24 August 2019) School of Information Systems, Singapore Management University.
- [5]. Phishing Detection: A Literature Survey by Mahmoud Khonji, Youssef Iraqi, Senior Member, IEEE, and Andrew Jones from IEEE COMMUNICATIONS SURVEYS & TUTORIALS, VOL. 15, NO. 4, FOURTH QUARTER 2013