IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134130

Malicious Website Detection Using Machine Learning with Chrome Extension

Mr. Sumanth C M[1], Sumanth H[2], Varun C L[3] , Vijay J D[4], Siddharth B P[5]

Assistant Professor, Department of Computer science and Engineering, Malnad College Of Engineering[1]

UG Student, Department of Computer science and Engineering, Malnad College Of Engineering[2-5]

Abstract: The website security is an important issue that must be pursued to protect Internet users. Traditionally, blacklists of malicious websites are maintained, but they do not help in the detection of new malicious websites. This work proposes a machine learning architecture for intelligent detecting malicious URLs. Forty-one features of malicious URLs are extracted from the data processes of domain, Alexa and obfuscation. ANOVA (Analysis of Variance) test and eXtreme Gradient Boost (eXtreme Gradient Boosting) algorithm are used to identify the 16 most important features. Finally, dataset is used to learn the eXtreme Gradient Boost classifier, which has a detection accuracy of more than 98%.

Keywords: eXtreme Gradient Boosting algorithm; Malicious URL;. Feature Analysis; Chrome Extension

I. INTRODUCTION

In today's interconnected world, the internet serves as a fundamental tool for communication, commerce, and information exchange. However, alongside its myriad benefits, the internet also harbors threats in the form of malicious websites, which aim to deceive users, steal sensitive information, or infect devices with malware. Detecting and mitigating these threats is essential to ensure a safe and secure online experience.

This project focuses on developing a robust solution for the detection of malicious websites using advanced machine learning techniques and integrating it into a Chrome extension. Leveraging the power of XGBoost, a state-of-the-art gradient boosting algorithm, the system analysis various features extracted from URLs, domains, HTML content, and JavaScript behaviour to assess the likelihood of a website being malicious.

By training the model on a diverse dataset comprising both benign and malicious website samples, it learns to recognize patterns indicative of malicious intent, enabling it to make accurate predictions in real-time. The Chrome extension seamlessly integrates this detection capability into users' browsing experience, providing a proactive defense mechanism against potential threats.

Through this project, we aim to enhance users' cybersecurity posture by empowering them with a tool that can identify and flag potentially harmful websites, thereby enabling them to make informed decisions and protect themselves from cyberattacks. By continually refining the model and updating its capabilities, we strive to stay ahead of emerging threats and ensure the ongoing effectiveness of our solution in safeguarding users' online activities.

II. RELATED WORK

"Identifying generic features for malicious url detection system", Malicious URLs pose serious cybersecurity threats to the Internet users. It is critical to detect malicious URLs so that they could be blocked from user access. Several techniques have been proposed to differentiate malicious URLs from benign ones. However, the goal of our work is to find the list of substantial features that can be used to classify most of the malicious URLs. In this paper, we select the most significant lexical features from different datasets using Chi-Square and ANOVA F-value. Later, we apply a voting classifier that combines several machine learning algorithms on those selected features[1].

"What's in a url: Fast feature extraction and malicious url detection". Phishing is an online social engineering attack with the goal of digital identity theft carried out by pretending to be a legitimate entity. The attacker sends an attack vector commonly in the form of an email, chat session, blog post etc., which contains a link (URL) to a malicious website hosted to elicit private information from the victims. We focus on building a system for URL analysis and classification to primarily detect phishing attacks.

DOI: 10.17148/IJARCCE.2024.134130

URL analysis is attractive to maintain distance between the attacker and the victim, rather than visiting the website and getting features from it. It is also faster than Internet search, retrieving content from the destination website and network-level features used in previous research[2].

“Based on URL Feature Extraction Identify Malicious Website Using Machine Learning Techniques”. This propose system deals with machine learning technology for the detection of phishing URLs by extracting and analysing various feature of legitimate and phishing URLs. Decision Trees, random forest and support vector machine algorithms are used to detect phishing websites or unsecure websites. The aim of the paper is to detect phishing URLs as well as cut down to the best machine learning algorithm by comparing the accuracy rate, false positive and false negative rate of each algorithm. This paper analyses the structural feature of the URL of the Phishing websites extracts 12 kinds of features and uses four machine learning algorithms for training and use the best-performing algorithm as our model to identify unknown URLs[3].

“Feature-based Malicious URL and Attack Type Detection Using Multi-class Classification”. Nowadays, malicious URLs are the common threat to the businesses, social networks, net-banking. Existing approaches have focused on binary detection i.e., either the URL is malicious or benign. Very few literature is found which focused on the detection of malicious URLs and their attack types. Hence, it becomes necessary to know the attack type and adopt an effective countermeasure. This paper proposes a methodology to detect malicious URLs and the type of attacks based on multi-class classification. In this work, we propose 42 new features of spam, phishing and malware URLs. These features are not considered in the earlier studies for malicious URLs detection and attack types identification. Binary and multi-class dataset is constructed using 49935 malicious and benign URLs. It consists of 26041 benign and 23894 malicious URLs containing 11297 malware, 8976 phishing and 3621 spam URLs. To evaluate the proposed approach, the state-of-the-art supervised batch and online machine learning classifiers are used. Experiments are performed on the binary and multi-class dataset using the aforementioned machine learning classifiers[4].

“Detecting malicious urls in e-mail–an implementation”. The World Wide Web has become the most essential criterion for information communication and knowledge dissemination. It helps to transact information timely, rapidly and easily. Identity theft and identity fraud are referred as two sides of cybercrime in which hackers and malicious users obtain the personal data of existing legitimate users to attempt fraud or deception motivation for financial gain. E-Mails are used as phishing tools in which legitimate looking emails are sent making the genuine users identity with legitimate content with malicious URLs. It helps to steal consumers’ personal data such as user names, account numbers, passwords and other financial account credentials. Spam E-Mails emerges or transforms as Phishing mails. Spoofed Mails plays a vital role in which the hackers pretends to be a legitimate sender posing to be from a legitimate organization which divulges the user to give his personal credentials. The content may escape from Content based filters or the email may be without any body of the message except malicious URL in it. This paper identifies malicious URLs in email through reduced feature set method[5].

III. IMPLEMENTATION

The implementation of the malicious website detection system using XGBoost and its integration into a Chrome extension is a multifaceted process. It commences with meticulous data collection, where a diverse dataset encompassing URLs labeled as either phishing or legitimate is amassed from reliable sources. This dataset undergoes rigorous preprocessing to ensure uniformity and cleanliness, followed by feature engineering. Features are meticulously extracted from the URLs, delving into address bar characteristics like IP address presence, "@" symbol occurrence, URL length, depth, and redirection indicators. Domain-based features, such as DNS records, website traffic metrics, domain age, and expiration period, are also derived. Additionally, HTML and JavaScript-based features, including iframe redirection, status bar customization, right-click disabling, and website forwarding, are extracted from web pages.

With the features prepared, the XGBoost algorithm emerges as the optimal choice for model training due to its ability to handle structured data effectively and its robust performance in classification tasks. The dataset is partitioned into training and testing sets, with the XGBoost classifier trained on the former. The model undergoes rigorous evaluation, assessing its performance using a suite of metrics such as accuracy, precision, recall, and F1-score to gauge its efficacy in distinguishing between malicious and legitimate URLs.

In parallel, the development of the Chrome extension unfolds. This entails designing and engineering an intuitive user interface that seamlessly integrates with the Chrome browser, allowing users to interact with the extension effortlessly.

DOI: 10.17148/IJARCCE.2024.134130

The extension is imbued with functionality to dynamically analyze URLs as users navigate the web, leveraging the trained XGBoost model to classify them in real-time. Upon classification, the extension promptly provides feedback to users, indicating whether a website is deemed malicious or legitimate, empowering them to make informed browsing decisions.

Following development, the extension undergoes comprehensive testing across various browser environments to ensure compatibility and functionality consistency. User feedback is solicited and meticulously analyzed to identify areas for improvement and refine the extension's performance further. Continuous monitoring and maintenance protocols are established to uphold the extension's efficacy and responsiveness in the face of evolving cyber threats. Regular updates are rolled out to the extension, incorporating enhancements and adapting to emerging threat landscapes to ensure users remain protected during their online endeavors.

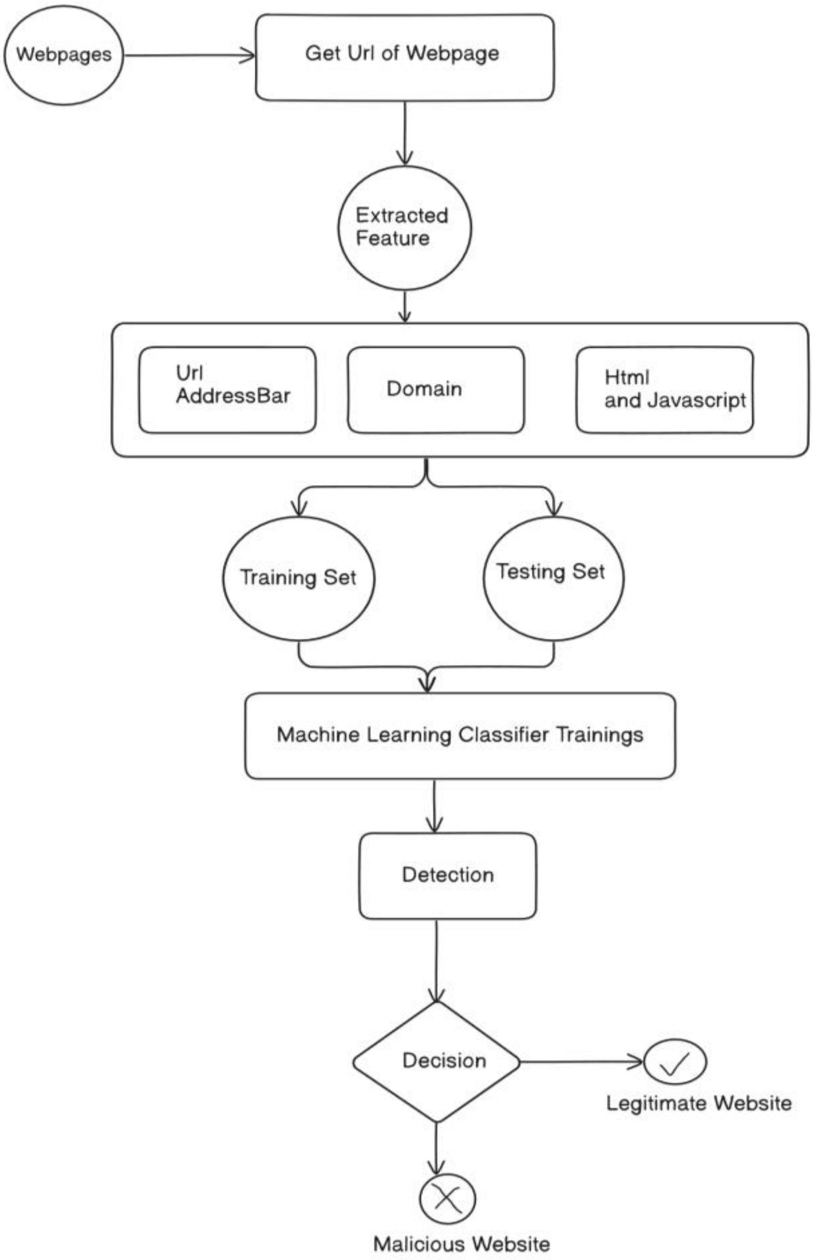


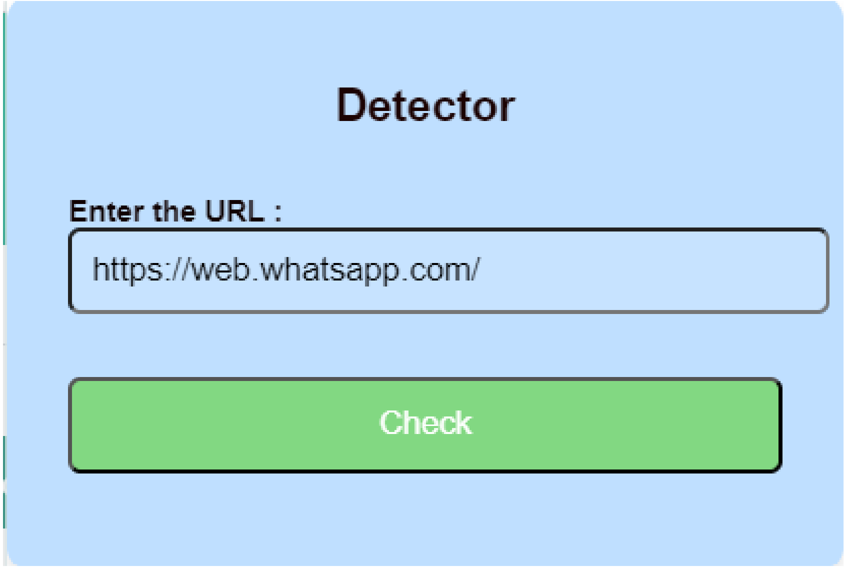
Fig.1. Work Flow For Malicious Website Detection

DOI: 10.17148/IJARCCE.2024.134130

IV. RESULTS

Results of the project demonstrate the efficacy and practicality of integrating machine learning-based malicious website detection into a user-friendly Chrome extension, contributing to a safer and more secure online experience for users. Continued refinement and updates to both the model and the extension ensure ongoing effectiveness in mitigating evolving cyber threats.

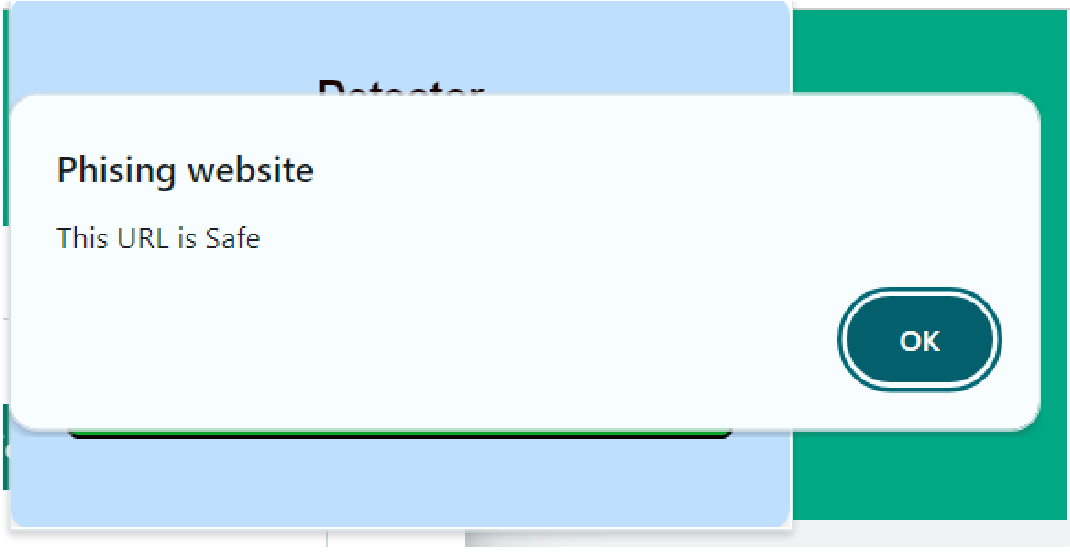
The project's outcomes demonstrate the effectiveness of merging machine learning-based malicious website detection into a user-friendly Chrome extension, enhancing online safety for users. By seamlessly integrating cutting-edge algorithms and intuitive features, the system empowers users to browse confidently, proactively identifying and mitigating potential threats in real-time. Continuous refinement and updates ensure the system's ongoing adaptability to evolving cyber threats, fostering a culture of collective cybersecurity awareness and proactive risk mitigation. In essence, this project sets a new standard for user-centric cybersecurity solutions, contributing to a safer and more secure digital landscape for all internet users.



Detector

Enter the URL : https://web.whatsapp.com/

Fig.2. Chrome Extension Display

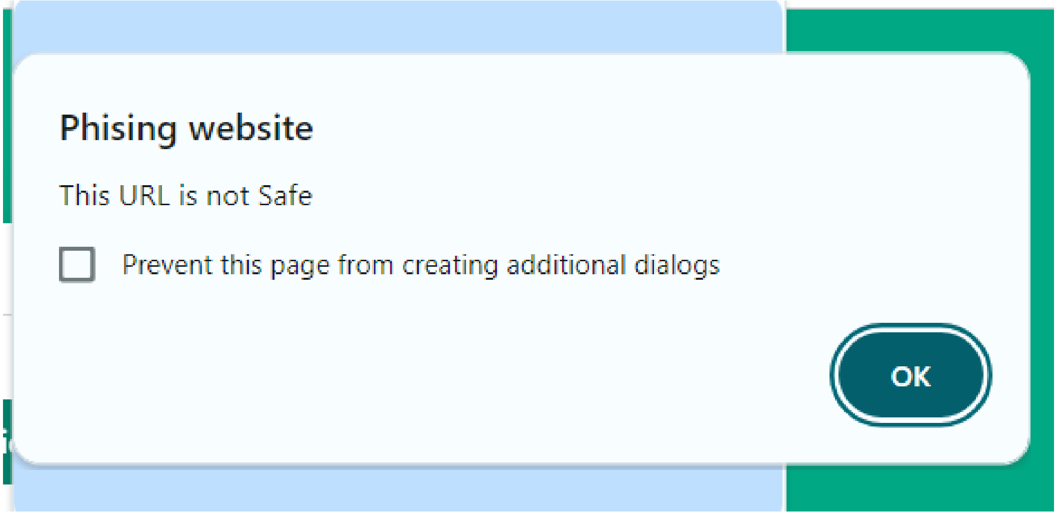


Phising website

This URL is Safe

Fig.3. Alert message for Safe Url

DOI: 10.17148/IJARCCE.2024.134130

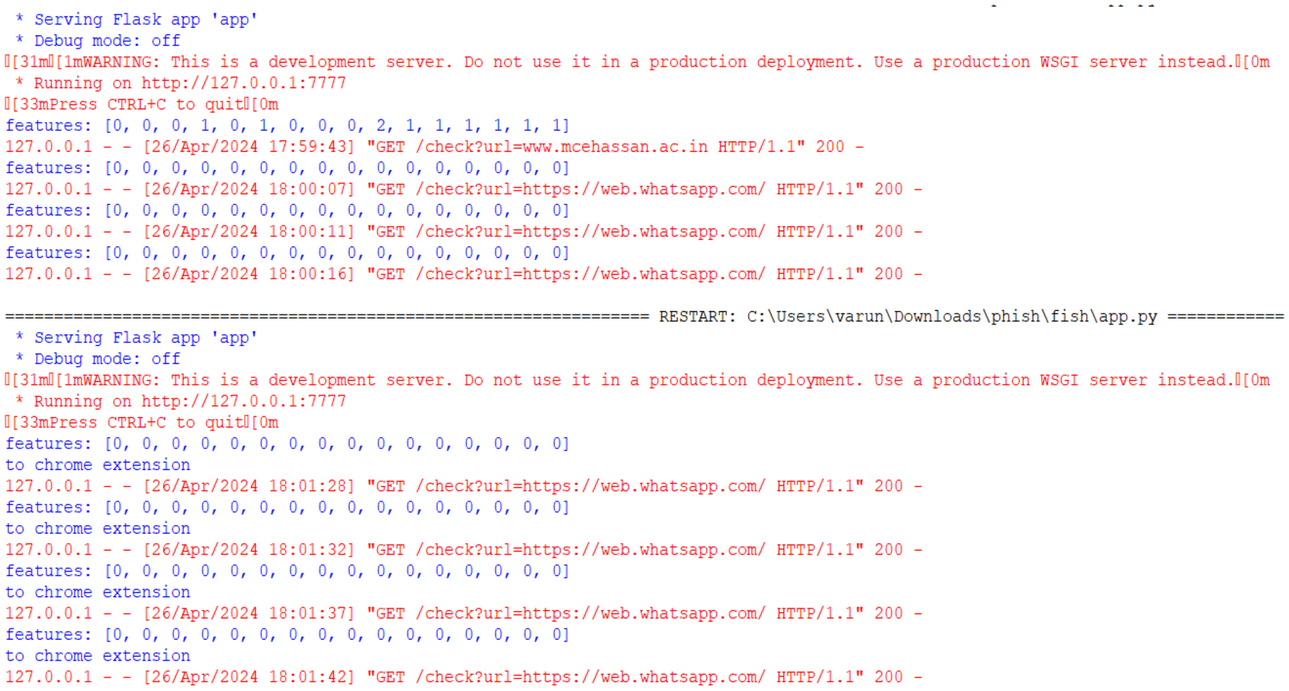


Phising website

This URL is not Safe

□ Prevent this page from creating additional dialogs

Fig.4. Alert message for Malicious Url



```
 * Serving Flask app 'app'
 * Debug mode: off
▯[31m▯[1mWARNING: This is a development server. Do not use it in a production deployment. Use a production WSGI server instead.▯[0m
 * Running on http://127.0.0.1:7777
▯[33mPress CTRL+C to quit▯[0m
features: [0, 0, 0, 1, 0, 1, 0, 0, 0, 2, 1, 1, 1, 1, 1, 1]
127.0.0.1 - - [26/Apr/2024 17:59:43] "GET /check?url=www.mcehassan.ac.in HTTP/1.1" 200 -
features: [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
127.0.0.1 - - [26/Apr/2024 18:00:07] "GET /check?url=https://web.whatsapp.com/ HTTP/1.1" 200 -
features: [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
127.0.0.1 - - [26/Apr/2024 18:00:11] "GET /check?url=https://web.whatsapp.com/ HTTP/1.1" 200 -
features: [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
127.0.0.1 - - [26/Apr/2024 18:00:16] "GET /check?url=https://web.whatsapp.com/ HTTP/1.1" 200 -

================================================================== RESTART: C:\Users\varun\Downloads\phish\fish\app.py ============
 * Serving Flask app 'app'
 * Debug mode: off
▯[31m▯[1mWARNING: This is a development server. Do not use it in a production deployment. Use a production WSGI server instead.▯[0m
 * Running on http://127.0.0.1:7777
▯[33mPress CTRL+C to quit▯[0m
features: [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
to chrome extension
127.0.0.1 - - [26/Apr/2024 18:01:28] "GET /check?url=https://web.whatsapp.com/ HTTP/1.1" 200 -
features: [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
to chrome extension
127.0.0.1 - - [26/Apr/2024 18:01:32] "GET /check?url=https://web.whatsapp.com/ HTTP/1.1" 200 -
features: [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
to chrome extension
127.0.0.1 - - [26/Apr/2024 18:01:37] "GET /check?url=https://web.whatsapp.com/ HTTP/1.1" 200 -
features: [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
to chrome extension
127.0.0.1 - - [26/Apr/2024 18:01:42] "GET /check?url=https://web.whatsapp.com/ HTTP/1.1" 200 -
```

Fig.5. Server Side Display

V. CONCLUSION

In conclusion, the project successfully demonstrates the fusion of machine learning-based malicious website detection with a user-friendly Chrome extension, significantly bolstering online security for users. Through seamless integration of advanced algorithms and intuitive features, the system empowers users to navigate the web with confidence, proactively identifying and mitigating potential threats in real-time.

Continuous refinement and updates ensure the system remains agile and responsive to evolving cyber threats, fostering a culture of collective cybersecurity awareness and proactive risk mitigation. Ultimately, this project exemplifies a paradigm shift towards usercentric cybersecurity solutions, contributing to a safer and more secure digital landscape for all internet users.

DOI: 10.17148/IJARCCE.2024.134130

REFERENCES

[1]. Hafiz Mohammd Junaid Khan, Quamar Niyaz, Vijay K Devabhaktuni, Site Guo, and Umair Shaikh. Identifying generic features for malicious url detection system. In 2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), pages 0347–0352. IEEE, 2019.

[2]. Rakesh Verma and Avisha Das. What's in a url: Fast feature extraction and malicious url detection. In Proceedings of the 3rd ACM on International Workshop on Security and Privacy Analytics, pages 55–63, 2017.

[3]. Khushbu Digesh Vara, Vaibhav Sudhir Dimble, Mansi Mohan Yadav, and Aarti Ashok Thorat. Based on url feature extraction identify malicious website using machine learning techniques. International Research Journal of Innovations in Engineering and Technology, 6(3):144, 2022.

[4]. Dharmaraj R Patil and Jayantrao B Patil. Feature-based malicious url and attack type detection using multi-class classification. ISeCure, 10(2), 2018.

[5]. Dhanalakshmi Ranganayakulu and C Chellappan. Detecting malicious urls in e-mail implementation. AASRI Procedia, 4:125–131, 2013.