e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

MALICIOUS URL DETECTION USING MACHINE LEARNING

Sanika Ambadkar[*1], Shraddha Deshmukh[*2], Dhanshree Agham[*3], Anuj Mahure[*4], Prof. Kalyani H. Deshmukh[*5]

[*1,2,3,4]U.G. Student, Department Of Computer Science Engineering, PRMIT&R, Amravati, Maharashtra, India.

[*5]Asst. Professor, Department Of Computer Science Engineering, PRMIT&R, Amravati, Maharashtra, India.

DOI : https://www.doi.org/10.56726/IRJMETS52962

ABSTRACT

Malicious URLs are links that when clicked, direct users to a web page or website that could be potentially hazardous or fraudulent. As the name suggests, a malicious URL never leads to anything good. The issue of cybersecurity is widespread because it has the ability to steal sensitive information and cause financial losses. Viruses such as phishing, spam, drive-by vulnerabilities, and other unwanted content are hosted on malicious URLs. Typically, a data breach will cost $4.24 million. Detecting and responding to such threats as soon as possible is of utmost importance as a result of this. The main tool used for detection in the past has been blacklists. However, blacklists are incomplete and unable to identify recently created harmful URLs. The generality of malicious URL detectors has been improved through the use of machine learning techniques. The goal of this study is to provide a comprehensive survey and a structural understanding of the methods used to detect harmful URLs using machine learning. A different approach to problem solving and machine learning approaches that yield a more accurate result are presented. The use of random forest and support vector machine algorithms is a popular approach when dealing with malicious URLs. Additionally, this paper provides a timely and comprehensive review of different strategies to tackle this issue in the cybersecurity industry, including coverage for future research.

Keywords- Malicious URL, Cybersecurity, Machine Learning, Cybercrime

I. INTRODUCTION

New communication technologies have been introduced increased the growth of various industries, including social Connecting for the purpose of e-commerce and online banking. Thousands of new websites are established daily that collect user data via [1] is the login function. Determining which websites is a challenging task. They are safe and reputable due to the large number of networks. In this context, cybersecurity is of utmost importance.It is difficult to decide which websites are secure and reliable because of the big range of networks. Cybersecurity is vital on this context.A malicious URL is a hyperlink that directs customers to a faux or risky internet web page or internet site whilst clicked [4]. Nothing high-quality can ever come from a malicious URL, because the call implies [5]. This is due to the fact the intention of producing those horrible internet pages is generally to in addition a criminal agenda, scouse borrow private or corporation facts, or make rapid money [6]. The ability to store facts on technological gadgets raises the opportunity of being attacked via way of means of intruders. The World Wide Web`s uniform useful resource locator, or URL, is the worldwide deal with for files and other resources [7]. A URL is made of parts: the protocol identifier which represents the protocol for use and the useful resource call which specifies the area call of the useful resource [8]. Malware has modified over time, making it tougher to become aware of those files.

II. LITERATURE SURVEY

Akamai`s file now no longer simplest blocks malware-associated queries however additionally famous that its internet safety device has blocked 6.258,597 queries associated with phishing, in addition to sharing its personal revel in with phishing. The file located that Platform, Finance, Global Services, CIO Office, Web Sales and Marketing devices had been the maximum centered commercial enterprise devices in phrases of phishing attempts, even as Support, Media and Carrier groups had been additionally targets. This variety is much less than the variety of Malware-associated queries that Akamai has blocked, however it's miles nevertheless significant. Google Safe Browsing statistics suggests that there at the moment are almost seventy five instances extra phishing webweb sites at the net than there are malware webweb sites. A latest take a look at located that the biggest class of phishing, 34.7%, is centered at webmail and SaaS users. In the beyond year, the variety of BEC

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

(commercial enterprise e mail compromise) assaults on unfastened webmail carriers has expanded from 61% to 72%, with extra than 1/2 of of those assaults the usage of Gmail. 25% of all breaches are social assaults, along with phishing. [3].

The motive of this phase is to show that the distinction between the loading time of the profile pages of the blocked and the unblocked person may be used for a timing attack. In the subsequent sections, we can first speak about the traits of RTTs measured for the blocked and unblocked person accounts. Then, we can speak about some technical strategies that could make the RTT greater distinguishable. Finally, after enforcing the RTT enlargement technique, we can take a look at if the RTT is statistically distinguishable the usage of exclusive social net offerings like Twitter, Facebook, eBay, and Xbox Live. he writes GET requests in Java to a web page that blocks A and every other web page that doesn`t block A to look the distinction withinside the measured RTT measures for the 2 accounts: blockading and non-blockading the subsequent describes the measured RTT measurements the usage of 3 social net offerings: Facebook, Twitter, Tumblr [12].

Patgiri et al. [17] consciousness on evaluating ML fashions for fixing the malicious URL problem. To decide which version is suitable, they take a look at becoming primarily based totally on accuracy and computational time. The dataset is gathered from an internet database and becomes separated into 3 groups: 80:20, 70:30, and 60:40. A binary type is used to decide the gadget`s overall performance via way of means of indicating whether or not a website is malicious or safe. SVM and Random Forest are the 2 distinguished ML fashions, and their effectiveness standards had been decided primarily based totally on the accuracy, minimum, and most values of the records examined.

III. METHODOLOY

Data Acquisition and Preparation: Collect categorized datasets containing fraudulent and non-fraudulent URLs from dependable sources. Clean the facts with the aid of using dealing with lacking values and doing away with beside-the-point columns, then encode specific variables into numerical layout for analysis.

Classical Machine Learning Approach: Feature Engineering: Enhance the dataset with the aid of using developing extra informative functions that describe the facts, along with URL period and server type.

Model Training: Utilize classical ML algorithms like logistic regression, selection trees, guide vector machines (SVM), and neural networks for education at the organized dataset.

Model Evaluation: Evaluate the skilled fashions in the usage of metrics like accuracy, precision, recall, and F1-rating on each education and trying out datasets.

Quantum Machine Learning Approach: Dataset Adaptation: Encode the dataset`s functions right into a layout appropriate for quantum algorithms, prioritizing numeric illustration for compatibility.

Algorithm Selection: Choose suitable quantum device mastering algorithms, along with a Variational Quantum Classifier (VQC), well suited to the tailored dataset.

Model Training and Evaluation: Train the chosen QML version in the usage of quantum hardware or simulators, then compare its overall performance with the usage of set-up metrics.

Comparison and Analysis: Compare the overall performance of classical ML fashions with QML fashions in phrases of accuracy, education time, and useful resource utilization. Analyze the strengths and weaknesses of every technique and pick out eventualities wherein QML would possibly provide benefits over classical ML.

Further Experimentation and Optimization: Experiment with special mixtures of characteristic encoding, quantum algorithms, and optimization strategies to enhance version overall performance.

Optimize parameters and hyperparameters of the fashions primarily based totally on assessment consequences to beautify accuracy and efficiency.

Documentation and Reporting: Document all steps of the methodology, inclusive of facts preprocessing, version education, assessment, and consequences.

Prepare a complete file outlining the findings, insights, and suggestions for destiny studies or applications.

Future Research Directions: Identify capacity regions for addition studies, along with exploring novel quantum algorithms, enhancing facts encoding strategies, or investigating the effect of various hardware systems on QML overall performance.

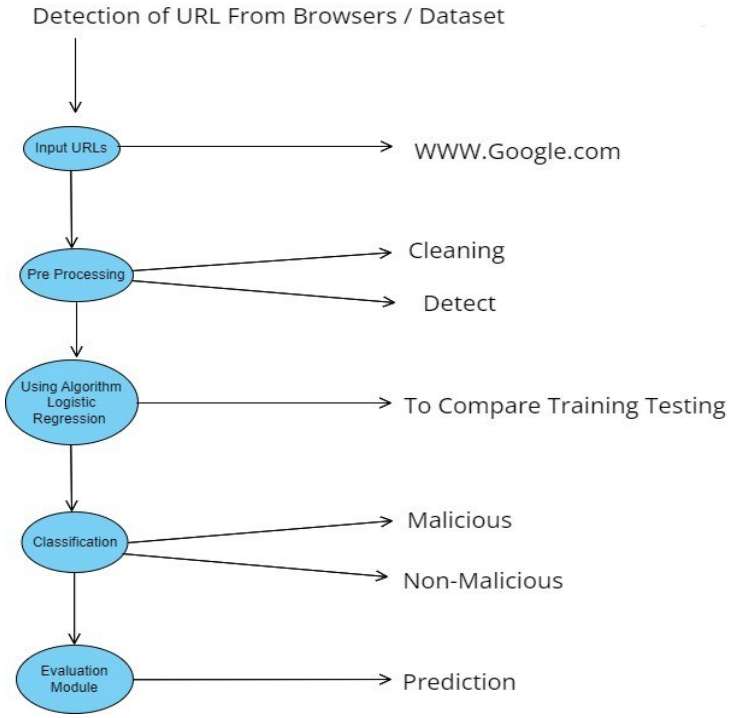
e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

IV. PROBLEM STATEMENT

Malicious URL detection represent of malicious URL detection involves indentifing and categorizing URLs as either benign or malicious based on various features such as domain reputation, URL structure, contain analysis, and behavior analysis. The goal is to develop algorithms or systems that can accurately differentiate between safe and harmful URLs to protect users from cyber threats such as phishing malware and scam.

The proliferation of malicious URLs poses a significant threat to individuals and organizations, exploiting human behaviour and trust in online interactions. Users frequently encounter deceptive web addresses through social media,email, and messaging platforms, leading to a range of cyber threats,including phishing attacks, malware infections and identity theft. The problem is exacerabated by the integration of social elements on these malicious sites, making it challenging for users to distinguish between legitimate and harmfulURLs

V. DATA FLOW DIAGRAM



VI. TOOLS AND TECHNOLOGY

1. Data Processing and Analysis:

Python: A programming language for records processing, analysis, and version implementation.

Pandas: Python library for records manipulation and analysis.

NumPy: Library for numerical computing, used for coping with arrays and matrices.

Matplotlib and Seaborn: Python libraries for records visualization to visualize the dataset and version performance.

2. Classical Machine Learning:

Scikit-learn: Python library for classical device getting to know algorithms which includes selection trees, help vector machines, logistic regression, and neural networks.

TensorFlow / Keras: Deep getting-to-know frameworks for imposing neural networks and deep getting-to-know models.

3. Quantum Machine Learning:

Qiskit: IBM`s open-supply quantum computing framework for operating with quantum circuits, simulators, and actual quantum computer systems.

Cirq: Google's open-supply framework for quantum computing, providing equipment for creating, simulating, and jogging quantum circuits.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

PennyLane: Quantum device getting-to-know library well matched with diverse quantum computing backends, facilitating integration with classical device getting-to-know frameworks like TensorFlow and PyTorch.

4. Quantum Computing Hardware:

IBM Quantum Experience: Access to IBM's cloud-primarily based quantum computer systems and simulators through Qiskit.

Google Quantum AI: Access to Google's quantum processors and simulators through the Cirq framework.

Amazon Bracket: Amazon's quantum computing carrier provides admission to quantum computer systems from a couple of hardware providers.

5. Development Environment:

Jupyter Notebook / JupyterLab: Interactive computing environments for jogging Python code, facilitating experimentation and analysis.

PyCharm / VS Code: Integrated improvement environments (IDEs) for Python improvement, presenting capabilities for code editing, debugging, and model manipulation.

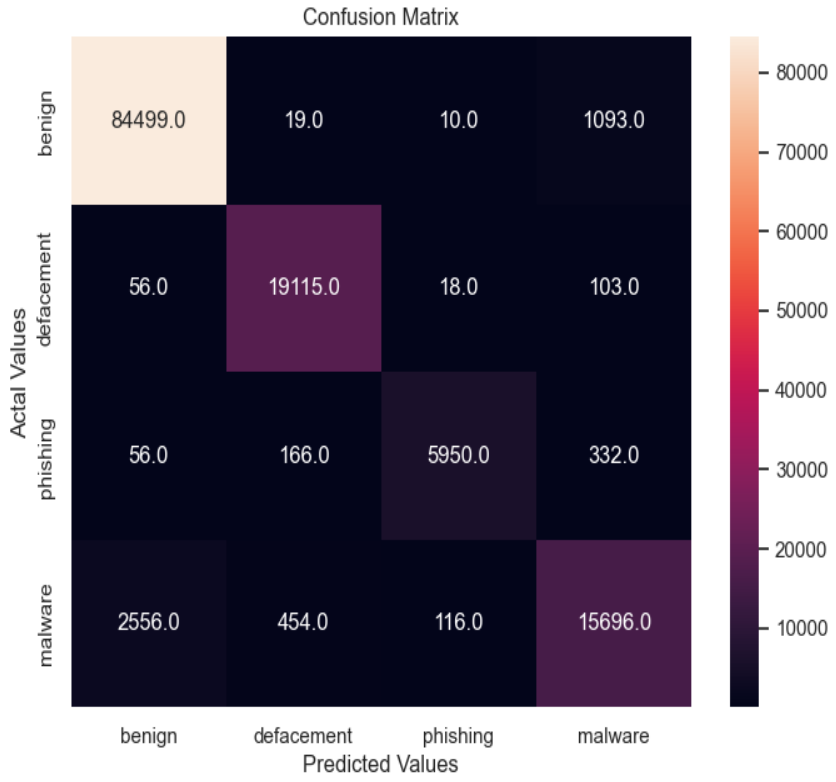
6. Miscellaneous:

GitHub: Version manipulation platform for collaboration, sharing code, and handling undertaking repositories.

VII. FUTURE SCOPE

In this art work, the cybersecurity problem of detecting fraud-using tool learning in every its traditional version and every about the dataset itself and about the usefulness of On the handiest hand, regarding the dataset, at some stage in the appearance at False Negatives, the final quit is that the exceptional measure learning models is that one of the three neural networks pro-posed in this art work modified into surely identified due to the fact the most premier ate the usefulness of QML models in cybersecurity, in an effective combinations that produce outcomes much like classical When comparing the outcomes obtained with ML and MLitt's some distance smooth that traditional models produce better outcomes, From the studies done on the prevailing literature, it is without delay clean that the software of QML is a totally recent field and therefore its effects are nevertheless very theoretical.

VIII. OUTPUTS AND ANALYSIS



e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

ANALYSIS

```
urls = ['br-icloud.com.br','en.wikipedia.org/wiki/North_Dakota']
for url in urls:
    print(get_prediction_from_url(url))
```

```
[LightGBM] [Warning] Unknown parameter: silent
MALWARE
[LightGBM] [Warning] Unknown parameter: silent
SAFE
SAFE
```

OUTPUT (Result)

IX. CONCLUSION

In conclusion, our implementation of Random Forest and XGBoost classifiers for malicious URL detection offers robust protection against evolving cyber threats. Through meticulous feature selection and model optimization, our approach achieves commendable performance metrics, enhancing cybersecurity defenses. While recognizing inherent limitations such as dataset biases, our findings underscore the effectiveness of machine learning in bolstering proactive threat detection. This work not only contributes to advancing cybersecurity measures but also highlights the practical applicability of sophisticated algorithms in real-world scenarios. Moving forward, continued research in this area promises further enhancements in safeguarding digital ecosystems against malicious URL infiltration.

VI. REFERENCES

[1] S. Abad, H. Gholamy, and M. Aslani, 'Classification of malicious URLs using machine learning', Sensors (Basel), vol. 23, no. 18, 2023.

[2] C. Hu et al., 'Recent trends in internet threats: Common industries impersonated in phishing attacks, web skimmer analysis and more', Unit 42, 28-Apr-2023.

[3] C. Jones and C. Jones, '50 web security stats you should know in 2023', Expert Insights, 14-May-2021. [Online]. Available: https://expertinsights.com/insights/50-web-security-stats-you-shouldknow/. [Accessed: 01-Nov-2023].

[4] S. Wen, Z. Zhao, and H. Yan, 'Detecting malicious websites in depth through analyzing topics and web-pages', in Proceedings of the 2nd International Conference on Cryptography, Security and Privacy, 2018.

[5] P. Shi, X. Yao, S. He, and B. Cui, 'Malicious URL detection with feature extraction based on machine learning', Int. J. High Perform. Comput. Netw., vol. 12, no. 2, p. 166, 2018.

[6] S. Selvaganapathy, M. Nivaashini, and H. Natarajan, 'Deep belief network based detection and categorization of malicious URLs', Inf. Secur. J. Glob. Perspect., vol. 27, no. 3, pp. 145–161, 2018.

[7] Shantanu, B. Janet, and R. Joshua Arul Kumar, 'Malicious URL detection: A comparative study', in 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021.

[8] M. Abramson and D. W. Aha, 'What's in a URL? Genre Classification from URLs', in Conference on Artificial Intelligence, 2012, pp. 262–263.

[9] B. Somerville et al., 'JavaScript and CSS. Beginning Rails 6: From Novice to Professional', pp. 281–298, 2020.

[10] C.-A. Staicu, D. Schoepe, M. Balliu, M. Pradel, and A. Sabelfeld, 'An empirical study of information flows in real-world JavaScript', arXiv [cs.CR], 2019.

[11] R. A. Mezei, 'Brief Introduction to Html', in Introduction to the Development of Web Applications Using ASP. Net (Core) MVC (pp, Cham: Springer Nature Switzerland, 2023, pp. 9–27.

[12] T. Watanabe, E. Shioji, M. Akiyama, K. Sasaoka, T. Yagi, and T. Mori, 'User blocking considered harmful? An attacker-controllable side channel to identify social accounts', in 2018 IEEE European Symposium on Security and Privacy (EuroS&P), 2018.

[13] J. Yuan, Y. Liu, and L. Yu, 'A novel approach for malicious URL detection based on the joint model', Secur. Commun. Netw., vol. 2021, pp. 1–12, 2021.

[14] G. Carleo et al., 'Machine learning and the physical sciences', Rev. Mod. Phys., vol. 91, no. 4, 2019.

[15] B. Lampe and W. Meng, 'Intrusion Detection in the Automotive Domain: A Comprehensive Review', IEEE Communications Surveys & Tutorials, 2023.

[16] C. Do Xuan, H. D. Nguyen, and V. N. Tisenko, 'Malicious URL detection based on machine learning', International Journal of Advanced Computer Science and Applications, vol. 11, no. 1, 2020.

[17] R. Patgiri, H. Katari, R. Kumar, and D. Sharma, 'Empirical study on malicious URL detection using machine learning', in Distributed Computing and Internet Technology, Cham: Springer International Publishing, 2019, pp. 380–388.