IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 6, June 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306199|

Beyond Blacklists: Building Scalable Malicious URL Detection with Advanced Machine Learning

Dr. A. R. Jeyasudha, Prawin Raj S P

Head of the Department, Department of Master of Computer Applications, Hindusthan College of Engineering and Technology, Coimbatore, India

II MCA Student, Department of Master of Computer Applications, Hindusthan College of Engineering and Technology, Coimbatore, India

ABSTRACT: The proliferation of malicious URLs poses a significant threat to online security. To mitigate this threat, this project proposes the development of a Machine Learning (ML) model designed to detect and prevent access to malicious URLs. The objective of this project is to enhance internet security by proactively identifying and blocking potentially harmful URLs. The significance of this project lies in its potential to revolutionize the conventional reactive approach to cybersecurity. Rather than responding to known threats after their manifestation, the proposed ML model establishes a proactive defence mechanism. By continuously learning and adapting to the dynamic landscape of online threats, it enables the real-time identification of malicious URLs, allowing for swift and effective preventive action.

KEYWORDS Malicious URL Detection, Machine Learning for Cybersecurity, Proactive Threat Detection, Real-Time URL Filtering.

I. INTRODUCTION

The escalating prevalence of malicious Uniform Resource Locators (URLs) constitutes a formidable menace to online security, necessitating robust measures to safeguard users and systems. In response to this critical challenge, this project seeks to address the burgeoning threat by proposing the creation of a sophisticated Machine Learning (ML) model specifically engineered for the detection and prevention of access to malicious URLs. By leveraging advanced ML algorithms, the objective is to fortify internet security protocols, thereby proactively identifying and thwarting potential cyber threats associated with harmful URLs.

II. RELATED WORK

In the recent past, we have witnessed a significant increase in cybersecurity attacks such as ransomware, phishing, injection of malware, etc. on different websites all over the world. As a result of this, various financial institutions, e-commerce companies, and individuals incurred huge financial losses.

In such type of scenario containing a cyber security attack is a major challenge for cyber security professionals as different types of new attacks are coming day by day.

Hackers create 300,000 new pieces of malware daily. - Source: McAfee
On average 30,000 new websites are hacked every day. - Source: Forbes
The above ones are determining how rapidly these malicious websites spread all over the world on a daily basis.

Kevser Sahinbas & Volkan Dortkardes – Istanbul Medipol University , Turkey [1] - The increasing vulnerability of users' sensitive information, such as usernames and passwords, due to the pervasive use of the Internet for various daily activities. As technology advances, the convenience offered by online services is accompanied by an escalating threat landscape, with web pages and applications becoming prime targets for attackers. The paragraph

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306199|

underscores the fact that the rapid proliferation of web pages has resulted in a substantial increase in malicious websites, posing a significant risk to both individuals and institutions.

The reference to the malicious behavior of users, whether intentional or unintentional, emphasizes the potential dangers associated with simply visiting harmful web pages. The paragraph also notes the prevalence of attackers exploiting web environments by embedding or uploading malicious code, making it easier for them to compromise user security. Citing Google Research Center's data that over 10% of web pages contain malicious code underscores the scale of the challenge in maintaining a secure online environment. The paragraph concludes by emphasizing the critical importance of detecting malicious URLs in cybersecurity applications. It acknowledges that the detection of malicious URLs is paramount in safeguarding users from the evolving threats present on the internet. Additionally, the mention of Cyber-Physical Systems (CPS) highlights the integral role they play in the development of secure data-accessing and data-processing services on the internet, further emphasizing the need for robust cybersecurity measures. Overall, the paragraph underscores the urgency of addressing cybersecurity concerns in the context of the growing threat landscape associated with web pages and applications.

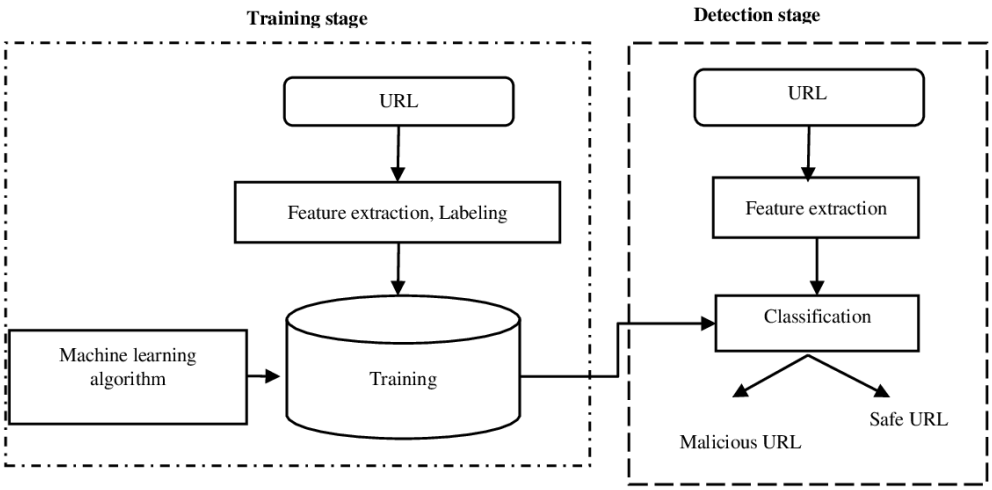
III. METHODOLOGY

3.1 Data Preprocessing:

For training and testing machine learning algorithms, we have used a huge dataset of 651,191 URLs, out of which 428103 benign or safe URLs, 96457 defacement URLs, 94111 phishing URLs, and 32520 malware URLs. Figure 2 depicts their distribution in terms of percentage. As we know one of the most crucial tasks is to curate the dataset for a machine learning project. We have curated this dataset from five different sources.

For collecting benign, phishing, malware and defacement URLs we have used URL dataset (ISCX-URL-2016) For increasing phishing and malware URLs, we have used Malware domain black list dataset. We have increased benign URLs using faizan git repo At last, we have increased more number of phishing URLs using Phishtank dataset and PhishStorm dataset As we have told you that dataset is collected from different sources. So firstly, we have collected the URLs from different sources into separate dataframe and finally merge them to retain only URLs and their class type.

The final pre-processed dataset is available at kaggle https://www.kaggle.com/sid321axn/malicious



3.2 Data Loading and Preprocessing:

- The script loads a CSV file named "malicious_phish.csv" containing URLs and their types (benign, phishing, malware, defacement).
- It checks for missing values and performs feature engineering to create new features based on the existing URL data. This includes:

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306199|

- Extracting information like presence of IP address, abnormalities in hostname matching, presence in Google search index.
- Counting occurrences of dots, "www", "@", etc., and analyzing URL length and suspicious words.
- Identifying features like directory depth, subdomains, shortening service usage, protocol type (http/https), and Top-Level Domain (TLD).

3.3 Exploratory Data Analysis (EDA):

It uses libraries like seaborn to visualize the distribution of some features for different URL types. This helps understand how features differ between malicious and benign URLs. This is shown in figure 3.1.1

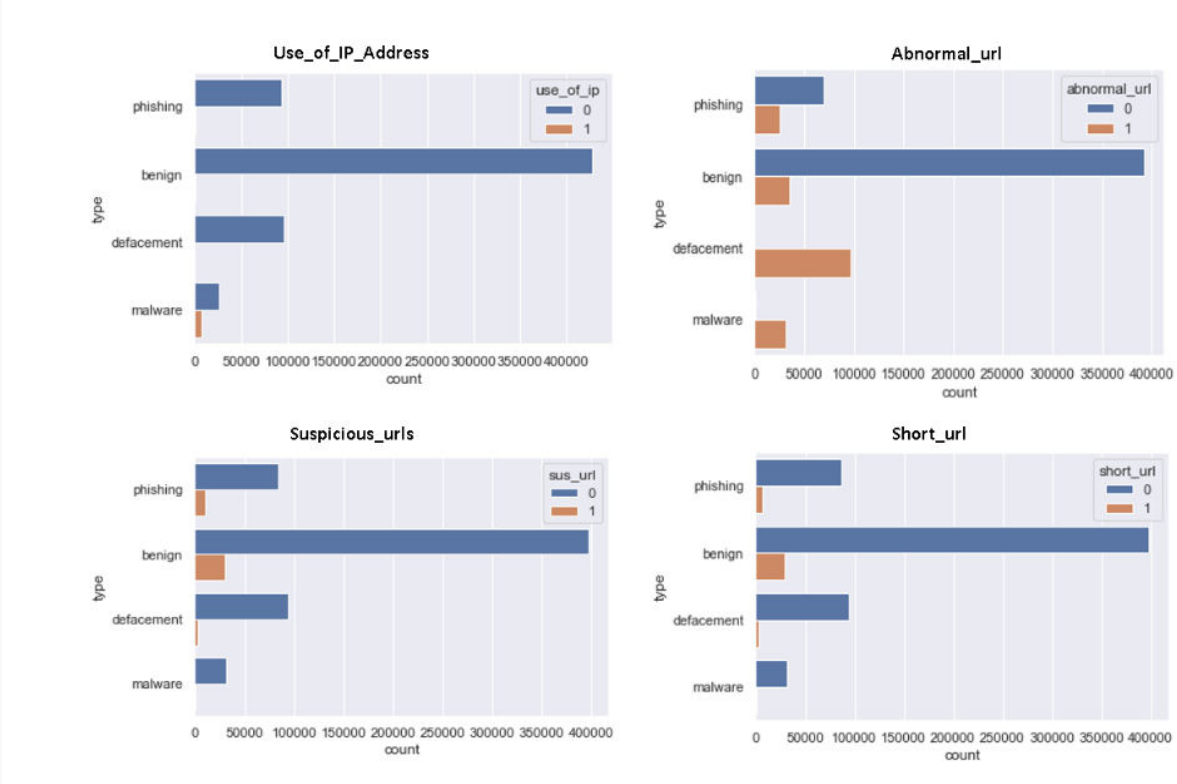


figure 3.1.1

3.4 Target Encoding:

The script selects relevant features for model training based on domain knowledge or feature importance analysis (not shown in this code).

It then splits the data into training and testing sets. The training set is used to train the model, and the testing set is used to evaluate its performance on unseen data.

3.5 Model Building and Evaluation:

The script builds a Random Forest classification model with 100 trees and trains it on the training data. Random Forest is a popular ensemble method that combines decision trees for improved accuracy and robustness.

The model is then evaluated on the testing data using various metrics like classification report, accuracy score, and confusion matrix. These metrics help assess how well the model can distinguish between different URL types.

3.6 Feature Importance:

The script analyzes feature importances to understand which features contribute most to the model's predictions. This helps identify the most informative features for URL classification.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306199|

3.7 Prediction Function:

The script defines functions to prepare features from a new URL and make predictions using a trained model (assumed to be saved as "lgb").
It extracts similar features from the new URL as done during preprocessing.
It predicts the URL type (safe, defacement, phishing, malware) based on the model.

IV. CONCLUSION

here we have demonstrated a machine learning approach to detect malicious urls. we have created 22 lexical features from raw urls and trained three machine learning models xg boost, light gbm, and random forest. further, we have compared the performance of the 3 machine learning models and found that random forest outperformed others by attaining the highest accuracy of 96.6%. by plotting the feature importance of random forest we found that hostname_length, count_dir, count-www, fd_length, and url_length are the top 5 features for detecting the malicious urls. at last, we have coded the prediction function for classifying any raw url using our saved model i.e., random forest.categories, continuously enhance the crop categories the model can identify, and raise the identification accuracy.

REFERENCES

1. "A Survey of Blacklists and Graylists for Web Spam Filtering" by M. L. Franz & B. J. Levine (2000)
2. "Analysis of Blacklist-Based Malware URL Detection" by R. S. Chhikara & S. S. Narula (2014)
3. "Evolving Threat: A Survey of URL Obfuscation Techniques" by S. Stajkovic & J. Zwitter (2013)
4. "Survey of Machine Learning Techniques for Malware Classification" by A. Vinayakumar & K. P. Soman (2017)
5. "Machine Learning-Based Phishing Detection: A Review" by S. V. Sharma & S. V. Sharma (2019)
6. "Deep Learning for Malicious URL Detection: A Survey" by S. J. Shen & H. T. Ma (2020)
7. https://wisdomml.in/malicious-url-detection-using-machine-learning-in-python
8. Adversarial Machine Learning in Online Fraud Detection" by R. K. Sharma & V. Bajpai (2020)
9. "Domain Adaptation for Spam Filtering" by H. L. Engstrom & P. P. Smyth (2006)
10. "Towards Explainable AI for Cybersecurity Decision-Making" by M. L. T. Yap & Y. Tan (2019)

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details