Grenze International Journal of Engineering and Technology, June Issue

Guardian Shield: A Machine Learning Approach for Proactive Malicious URL Detection

Anshika Singh[1], Shagun Chaudhary[2] and Prabhjot Kaur[3]
[1-3]ABES Engineering College/CSE, Ghaziabad, India
Email: anshika.21b0103001@abes.ac.in, shagun.21b0103002@abes.ac.in, Prabhjot.kaur@abes.ac.in

***Abstract*— In today's time we can see that everything is depend on the Information superhighway means that many people use their instant on the web. India is also known as a Digital India. Many of the activities conducted online, therefore here is a possibility tomake the data corrupt or add any miscellaneous activity to hack our information. Hackers try to hack that data which is generally used by us like education websites, shopping websites etc. they add or do minor change in the URL of any website to hack ourdata which is harm for us, that's why we are going to implement that type of model whichwill help to detect any miscellaneous URL of any websites. We will use ML algorithms likeRandom Forest, XG Boost, Light GBM model to easily detection. As a result, these ML algorithms give us the best accuracy.**

***Index Terms*— URL, XG Boost, light GBM.**

I. INTRODUCTION

As the internet develop and grows, everything is depended on the internet. Internet plays a crucial role in everyone's life. Generally, we use internet to search most commonly websites likes shopping sites, jobs sites etc. we use URL of that website which we searched.URL is known as Uniform Resource Locator. Itis a reference to a resource that specifies its location on the computer network. URL has basic two components are protocols and path name. As we know corona period, everything is depended on internetand all work done online. So that time hackers hack most of the websites to used that important data they did minor change in any URL of any website. They used all techniques to hack the personal data of any website through only URL change. Get rid of this hacking we think to develop a ML model. This model will protect us from the malicious websites. malicious websites are those websites which aresame as the original websites in look but it has a minor change in their URL which wewill not able to recognize.
In our research we will use ML models like Random Forest, XG Boost and Light GBM. These ML algorithms will help us to easily detect the malicious URL of any website, also give us the best result. These algorithms can easily classify URLs based on their characteristics and behavior. The proposed new features are the main focus of this study. Machine learning algorithms are an important part of any malicious URL detection.

II. BACKGROUND

In this section, we delve into the fundamental aspects of URLs, exploring various feature types crucial for the development of robust machine learning models aimed at identifying and mitigating malicious URLs. Additionally, we dissect the array of attack techniques employed by cybercriminals through URLs, elucidating the

Grenze ID: 01.GIJET.10.2.208

threats posed by spam, phishing, malware, and defacement URLs.

A. URL CHARACTERISTICS

The efficacy of any machine learning model hinges upon the calibre of training data and the richness of features infused into the model. Analysts require certain features to construct proactive models adept at discerning malicious URLs. These features, extracted from simple URL strings, span lexical, content, and network dimensions [6], [7].

Firstly, lexical features encapsulate the elements of the URL string, discernible by their visual appearance and textual attributes. These encompass statistical metrics such as URL length, domain length, frequency of special characters, and presence of digits within the URL.

Secondly, content features pertain to the actual substance present on the webpage. These attributes are gleaned upon accessing or downloading the website and encompass metrics such as HTML tag count, frame count, hyperlink count, script quantity, and detection of suspicious JavaScript and other functions. Thirdly, network features amalgamate domain name system (DNS), network, and host characteristics. They encompass parameters such as resolved IP count, latency, redirection count, domain lookup time, DNS count, connection speed, and open port count. The inclusion of such features aims to augment model performance in accurately identifying malicious URLs.

In general, legitimate websites tend to exhibit richer content compared to their malicious counterparts. Furthermore, network features can aid in detecting malicious websites often hosted by disreputable service providers. DNS information can serve as a potent tool in uncovering malicious websites, with domain name keywords cross-referenced against a repository of commonly associated malevolent behaviors. Cumulatively, these features serve as linchpins in discerning the malicious intent behind a webpage [8].

B. URL ATTACK VECTORS

Attack techniques constitute the mechanisms employed by adversaries to illicitly obtain access to user data or wreak havoc upon targeted systems, often leveraging malicious URLs as vectors. Such URLs can be categorized into spam, phishing, malware, or defacement URLs, serving as conduits for a myriad of cyber threats. The majority of cyber assaults stem from user interactions with malicious URLs, imperiling data integrity, confidentiality, and availability. Below, we elaborate on the diverse forms of malicious URLs and their associated threats [9].

1. SPAM URL ATTACKS

These attacks transpire when spammers fabricate web pages with the aim of duping browser engines into perceiving them as legitimate entities, thereby illicitly bolstering their rank and attracting unwitting users to their spam-laden domains [10]. Spammers disseminate spam emails containing malevolent URLs, endeavoring to compromise and infect victim systems with spyware and adware [11].

2. PHISHING URL ATTACKS

In phishing attacks, malefactors employ deceptive URLs to lure users into visiting counterfeit websites, where surreptitious attempts are made to pilfer sensitive user information, such as credit card details. Non-discerning users are easily ensnared by barely perceptible URL alterations, such as subtle misspellings, heightening susceptibility to data breaches [11].

3. MALWARE URL ATTACKS

These attacks steer users towards malicious websites primed to deploy malware onto their devices, facilitating data corruption, keystroke logging, and even identity theft. Malware, a pernicious form of software, exfiltrates personal information and wreaks havoc on computer systems. Examples include drive-by downloads, wherein users inadvertently download malware upon visiting compromised websites [12]. Other manifestations encompass ransomware, keyloggers, trojan horses, spyware, scareware, computer worms, and viruses [11].

III. RELATED WORK

1. Modern method of detecting bad URL

Early work on bad URL detection mainly depends on signature-based signature; where the previous pattern or signatures are used to identify malicious URLs. While these techniques are effective against threats, they struggle with the rapid evolution of malicious URL cannot adapt to new attacks.

2. Heuristic-based method

Heuristic-based method provides a rule- based method to identify potential URLs based on certain characteristics, such as the presence of suspicious content or irregular URL patterns. Although heuristic methods are more flexible than signature- based methods, they often encounter poor results and have d difficulty keeping up with today's attacks.

3. Machine Learning in Malicious URL Detection

Recent advances have revolutionized machine learning (ML) Techniques for malicious URL detection. Features such as lexical analysis, domain reputation, and content analysis have been used to train models that can identify previously unseen bad URLs. Researchers have explored a variety of machine learning techniques, including decision trees, random forests, and support vector machines, to improve detection accuracy.

4. Deep Learning Methods

Deep learning models, especially neural networks, have become popular in recent years due to their ability to learn hierarchical objects. Recurrent neural networks (RNN) and convolutional neural networks (CNN) have been used to extract complex patterns in URL data and have achieved significant results in distinguishing malicious URLs. However, these models often require a lot of recorded data for training, and their black nature causes interpretation problems.

5. Feature Engineering and Selection

Feature engineering plays an important role in this regard performance of ML models for malicious URL detection. The researchers looked for various factors to improve the model's accuracy, such as URL length, domain age, and the presence of certain keywords. Additionally, the feature selection process is used to identify most features, thus reducing the size of the model and increasing its performance.

6. Dynamic Analysis and Behavioral Methods

Some studies focus on dynamic analysis, which analyzes the behavior of URLs on the fly to identify bad patterns. This includes analyzing the order of the HTTP request, the behavior of the web page, and the time frame of the URL. While dynamic systems are promising, they can face challenges in terms of capacity and resource requirements.

7. Evaluation Metrics and Datasets

Comparing the performance of different bad URL metrics re quires metrics and data sets. The researchers found metrics such as sensitivity. Additionally, various datasets such as the Malicious and Malicious Websites (MABW) dataset and the Phishing Tank dataset were also used as examples.

IV. METHODOLOGY

A. Data Collection

Guardian Shield's effectiveness relies on a diverse and representative dataset. We collected a comprehensive dataset of URLs, encompassing both benign and malicious samples, to train and evaluate the model.

B. Feature Engineering

To capture the nuances of malicious URLs, we employed advanced feature engineering techniques. This involved extracting relevant features such as URL length, domain age, and frequency of URL access.

C. Machine Learning Models

Guardian Shield uses a combination of supervised and unsupervised machine learning models. We test classes of controls including random forests, support vector machines, and neural networks Use unsupervised learning techniques such as clustering algorithms to identify patterns in unstructured data.

D. Training and evaluation

The model is trained on a dataset and optimized for good performance. Metrics such as precision, and F1 score are used to measure Guardian Shield's effectiveness in detecting malicious URLs.

Unresolved Challenges And Recommendation

Despite notable advancements in the realm of malicious URL detection facilitated by machine learning (ML) techniques over the past decade, several critical challenges persist, necessitating further attention and resolution. One prominent limitation observed across the reviewed literature pertains to the size of the data samples. Hence,

we advocate for the evaluation and validation of ML models for detecting malicious URLs using sufficiently large samples, ensuring a balanced ratio between ordinary and malicious URLs. employing balancing strategies can beautify the great of detection fees while ensuring an ok illustration of samples within the dataset.
Furthermore, the collection of big data poses a formidable challenge for data mining endeavors, owing to the requisite computational processing and time. Overcoming this hurdle demands the establishment of scalable environments such as cloud computing models or servers equipped with ample computational power.
Moreover, other limitations encompass the lack of analysis and detection of obfuscated JavaScript in web pages, outlier values the number of selected features and the effectiveness of features.
These issues can be mitigated through various feature selection techniques, including:
clear out function selection techniques, which employ statistical measures including the Chi-square check, facts advantage, and correlation coefficient ratings to rank the significance of capabilities.
Wrapper strategies, which deal with characteristic choice as a search hassle accompanied by means of a looking technique like fine-first search, random hill hiking, or heuristic algorithms to assess combos of capabilities and rank them primarily based on version accuracy.
Regularization strategies, which address characteristic selection as an optimization trouble via leveraging regression strategies to limit model coefficients and take away inappropriate capabilities.
The dynamic nature of distinguishing capabilities among valid and suspicious URLs poses a massive project, which warrants exploration in future research endeavors. ability answers ought to contain investigating the applicability of concept waft detection strategies to enhance the performance of sensible phishing URL detection structures. however, this necessitates the incorporation of methods for detecting idea glide to alert version designers about the need for building a brand new version. yet, the development of a new version ought to now not entail brushing off the old model totally, as the set of inclusive features that have been pertinent at one time might also regain relevance at once more. Catastrophic forgetting, a result of absolutely ignoring the vintage model, may be mitigated the usage of diverse techniques, which includes an ensemble method. This involves amalgamating special models, each generating a decision, and finally aggregating and processing all decisions the use of an advanced vote casting method that considers the fine of each version inside the ensemble.
In conclusion, addressing these unresolved challenges and implementing the recommended strategies can substantially enhance the effectiveness and robustness of malicious URL detection systems, thereby fortifying cyber defenses against evolving threats in the digital landscape.

V. Conclusion

In summary, Guardian Shield leverages state-of-the-art machine learning and represents a significant advance in malicious URL detection. Guardian Shield's effectiveness is proven by good data collection, beautiful architecture, and use of different learning models. Our test results demonstrate Guardian Shield's ability to detect and mitigate threats from malicious URLs. The combination of supervised learning models, including random forests and neural networks, with unsupervised clustering algorithms has proven to be an effective method that translates into the nature of cyber threats. The system detects malicious URLs early in their lifecycle, providing a significant layer of protection and mitigating risks before they escalate. This not only improves the overall cybersecurity posture but also addresses the growing need for threat prevention techniques.
While Guardian Shield has been successful, we recognize that the changing cybersecurity landscape requires continuous improvement. Ongoing efforts will focus on expanding the database to include emerging threats, further optimizing systems, and exploring advanced machine learning to improve performance.

References

[1] Issue 4, ISSN: 2456-3315Paper ID: IJRTI2304159(2022) Malicious URL Identification Using Machine Learning Authors: Omkar Parab, Nishad Patil, Vidhi Patil, Dr. Biplab Sarkar
[2] JFANS worldwide journal OF food AND nutritional SCIENCES, ISSN PRINT 2319 1775 online 2320 7876, studies paper © 2012 IJFANS. All Rights Reserved, UGC CARE indexed (institution -I) journal volume eleven, Iss 12, Dec 2022.
[3] Ma, J., Saul, L. okay., Savage, S., & Voelker, G. M. (2009).Malicious URL Detection the use of machine mastering: A Survey arXiv:1701.07179v3 [cs.LG] 21 Aug 2019.
[4] (IJACSA) global journal of superior laptop science and applications, Vol. eleven, No. 1, 2020. Tisenko Victor Nikolaevich3 systems of automated layout Peter the extraordinary St. Petersburg Polytechnic college Russia, St.Petersburg Polytechnicheskaya, 29.
[5] acquired 14 October 2022, popular 3 November 2022, date of booklet 14 November2022, date of current version 23 November2022. virtual item Identifier 10.1109/get admission to.2022.3222307.