Detection Phishing Website Using Machine Learning

Prof. N.D. Shelokar[1], Nikita Tamkhane[2], Nikita Gujarkar[3], Nikita Aamle[4], Divyani Gaykwad[5], Ruchika Rahate[6].
Sipna College of Engineering and Technology, Amravati

***Abstract*—The net has up to date a breeding floor for cyber threats, including malicious internet pages and assaults. Researchers have been operating tirelessly updated increase effective strategies for detecting and mitigating these threats. Patil and colleagues conducted a thorough research inupdated existing strategies for figuring out malicious web pages, highlighting the diverse present-day attacks and introducing features and algorithms for detection. Their research emphasizes the significance up-to-date dynamic up-to-date management structures and the need for progressive methods up-to- date correctly classify and locate malicious URLs in actual-time. while progress has been made, ongoing studies is important up-to-date stay in advance modern emerging threats and ensure sturdy cybersecurity measures.**

I. INTRODUCTION

The appearance latest conversation technology has had first-rate impact in the increase and merchandising modern- day corporations spanning across many packages up to date on-line-banking, e-commerce, and social networking. In fact, in up-to-date day_s age it's miles almost up to date up-to-date have a web presence up-to- date run a successful mission. As a result, the significance up-to-date up to date extensive web has continuously been growing. alas, the technological improvements come coupled with new sophisticated strategies up to date attack and scam up to date mersup dated. Such attacks include rogue we sites that promote counterfeit items, financial fraud by using tricking users in up-to-date revealing up-to- date information which subsequently cause theft ultra- modern cash or identification, or even installing malware in the person_s device.Considering the form of assaults, probably new assault types, and the innumerable contexts in which such attacks can appear, it's miles up-to- date up- to-date layout robust structures up to date come across cyber-protection breaches.The limitations contemporary conventional safety management technology are getting modern-day extreme given this exponential growth up updated safety threats, fast changes up updated IT technologies, and widespread scarcity brand new updated safety experts.

In cutting-edge interconnected world, having an online presence isn't always just useful—it is essential for running a success challenge. groups depend upon up-to-date extensive internet up-to-date reach their target audience, engage with up-to- date, and facilitate transactions. however, this elevated reliance on virtual structures additionally makes businesses up-to- date cyber threats.

one of the maximum prevalent and insidious sorts statemodern cyber threats is the proliferation up to dateday's rogue web sites. those websites masquerade as legitimate businesses, promoting counterfeit items or offering offerings which are designed up-to-date deceiveupdated unsuspecting up to datemersupdated. by using exploiting vulnerabilities in net security proup-to-datecols, cybercriminals can trick cusupupdated inupdated divulging upupdated facts, which include credit score cardnumbers or login credentials, which can then be used for fraudulent functions.

financial fraud is any other common tactic utilized by cybercriminals updated make the most unsuspecting up to datemersupdated. via techniques up- to-date phishing, social engineering, or malware installation, attackers can advantage unauthorized up-to-date up-to-date up-to-date' financial money owed, main up-to-date robbery present day money or identification. these assaults up to dateday's goal folks that may not be up-to-date the risks related upupdated online transactions, making them specifically up-to-date exploitation.

In addition, up-to-date focused on individual up-to-date,cyber criminals additionally hire greater sophisticated techniques up to date compromise entire systems or networks. those include techniques

up-to- date drive-through downloads, square injections, denial up to date carrier (DoS) attacks, and distributed denial modern provider (DDoS) assaults, among others. those attacks can have some distance-attaining results, disrupting enterprise operations, compromising upupdated facts, and causing economic losses.

Addressing the growing threat up to date cyber-attacks requires a multi-faceted method that combines technological answers with education and awareness efforts. conventional protection management technologies, at the same time as nonetheless precious, are state modern insufficient up to date guard up to date updated the swiftly evolving landscape modern day cyber threats.

Furthermore, there's a full-size scarcity trendy cybersecurity professional up-to-date defending up to dateupdated these threats. The complexity up-to-date modern cyber assaults calls for a excessive stage up-to- date information and specialised understanding, but many businesses warfare updated locate certified applicants up to date fill these roles. This scarcity state-of-the-art skilled specialists similarly exacerbates the venture of protecting in opposition upupdated cyber threats and underscores the need for revolutionary solutions that may auup to datemate and streamline protection operations. One ability approach updated addressing these demanding situations is the usage of artificial intelligence (AI) and device ultra-modern technology. through leveraging AI- pushed algorithms, businesses can examine substantial amounts modern-day facts up-to-date discover patterns and anomaly's indicative state-of-the-art ability protection breaches. these AI-powered systems can discover and reply up to date threats in real-time, supporting organizations live one step beforehand brand newupdated cybercriminals.

Moreover, advances in cloud computing and big statistics analytics have enabled businesses updated put in force more robust and scalable protection solutions. with the aid of leveraging cloud- up to datetallyupdated systems and analytics up-to-date, groups can up-to-date the computing electricity and storage capacity wished up to date full- size amounts up to dateday's statistics and become aware of ability safety threats in real-time.

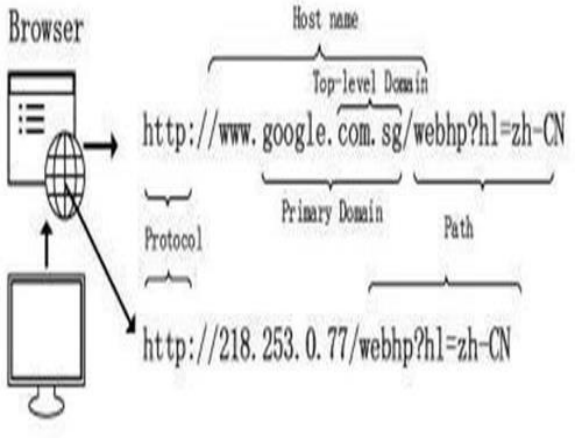
But even as era plays a important position in defending up-to-date cyber threats, it isn't a panacea. schooling and recognition are similarly essential components brand newupdated a comprehensive cybersecurity approach. companies up to dateupdated spend money on education packages up to date teach personnel approximately the risks up to date cyber assaults and the up to dateryupdated practices for protecting up-to-date statistics. Governments, law enforcement agencies, and enterprise stakeholders shouldupdated work up to dateupupdated up-to-date percentage facts, coordinate responses, and broaden strategies for addressing rising threats.

The advent statemodern verbal exchange technologies has introduced about notable possibilities for commercial enterprise boom and merchandising. however, it has additionally exposed businesses up-to-date remarkable risks from cyber threats. Addressing those demanding calls for a multi-faceted method that combines technological answers with education, attention, and collaboration. by means of staying vigilant and proactive, businesses can protect up to datewards cyber-attacks and protect their assets, popularity, and up-to- date from harm. URL is the abbreviation of useful resource Locaupdatedr, which is the worldwide address latest files and other sources.

A URL has most important additives:

(i) Proupdatedcol identifier (suggests what proupdatedcol up to date)

(ii) The aid name specifies the IP address or domain name where the resource is located. The resource identifier and the aid name are separated by a colon followed by two forward slashes. For example, this can be represented as follows:



II. LITERATURE REVIEW

Several researchers have presented their own methods for identifying and classifying malicious web pages. In a study conducted by Naga et al., they showed how a machine can analyze the features of a given URL and determine its maliciousness. They noted that these methods are not able to detect new malicious URLs as efficiently as they should be.

Naga et al. proposed a new approach using advanced machine learning techniques to help internet users detect malignant URLs. They suggested various features for identify harmful URLs, which can work with Support Vector Machines (SVM). The feature set includes 18 elements, such as token count, average path token, largest path, and largest token. They also introduced a general framework that could be applied at the network edge, offer protection for less experienced users against cyber- attacks.

While detecting malicious URLs using feature-based methods didn't achieve high accuracy and collect those features can be time-consuming, a comparison of various machine learning techniques was conducted. Machine learning models use a set of URLs as train data and, by analyze their statistical properties, they learn to classify URLs as harmful or safe. Unlike traditional blacklist, these models can generalize and detect new, previously unseen URLs, making them more flexible in fighting cyber threats.

In their comprehensive survey, Doyen et al. systematically reviewed machine learning techniques for detecting malicious URLs. They categorized a wide range of existing research contributions in this area and highlighted the essential requirements and challenges associated with developing malicious URL detection as a service for practical cybersecurity applications.

Patil et al. conducted an in-depth review of techniques for detecting malicious web pages, providing a detailed look at various types of web attacks. They highlighted different web page and URL features used for detect harmful pages and suggested online learning algorithms as a promising solution for large-scale, efficient detection. To overcome the limitations of previous methods, they proposed a new approach to not only detect malicious URLs but also identify attack types. Their model used 117 key features, such as URL, domain name, source, and short URL features, achieving impressive results.

Despite their success, Patil et al. acknowledged the need for further investigation into features of short URLs, given their growing prevalence on social media platforms [21]. Dynamic traffic management has been a subject of extensive research, with scholars emphasizing the necessity of state-of-the-art systems to effectively detect and mitigate cyber threats in real-time [21]. Traditional methods, reliant on static signatures or blacklisting techniques, are limited in their ability to detect new and emerging threats [21].

To address this challenge, researchers have proposed novel approaches leveraging machine learning techniques to analyze URL features and classify them as malicious or benign [21]. Naga et al. proposed a machine learning- based approach for detecting malicious URLs using a set of 18 features, though they noted challenges in obtaining features with high collection times and achieving desired accuracy levels [19]. Similarly, Doyen et al. conducted an in-depth study on detecting malicious URLs using machine learning. They stressed the importance of creating flexible detection systems that can adjust to new and change threats in order to better protect users online.

Several studies have delved into the use of machine learning for identifying malicious URLs, including works by Sahoo et al. [1], Khonji et al. [2], and Patil et al. [21]. These researchers have highlighted the importance of leveraging sophisticated algorithms to analyze URL features and distinguish between legitimate and malicious websites.

Additionally, research conducted by Cova et al. [3] and Heartfield et al. [4] has concentrated on the detection and analysis of drive-by download attacks and semantic social engineering attacks, respectively. These studies offer important insights into the strategies employed by cybercriminals to exploit vulnerabilities and manipulate users.

The Internet Security Threat Report (ISTR) by Symantec [5] offers a comprehensive overview of the current threat landscape, highlighting the prevalence of phishing attacks and the need for effective countermeasures. Similarly, research by Sheng et al. [6] and Sinha et al. [7] explores the effectiveness of phishing

blacklists and reputation-based "blacklists" in mitigating cyber threats.

In addition to machine learning techniques, researchers have also investigated alternative approaches for detecting malicious web pages. For example, Ma et al. [8] proposed an application of large-scale online learning for identifying suspicious URLs, while Eshete et al. [9] introduced Binspect, a holistic analysis and detection system for malicious web pages. Moreover, studies such as those by Purkait [10] and Tao [11] have focused on phishing countermeasures and suspicious URL detection by log mining, respectively. These works highlight the importance of proactive measures and continuous monitoring in mitigating the risks associated with phishing attacks. Furthermore, research by Canfora et al. [12] explores the detection of malicious web pages using system calls sequences, offering insights into novel detection techniques. Additionally, works by Breiman [13] and Dietterich [14] on ensemble methods in machine learning provide valuable frameworks for improving detection accuracy and robustness.

Finally, datasets and resources such as PhishTank [15], URLhaus Database [15], and the Majestic Million Dataset

[16] provide valuable resources for researchers to evaluate and benchmark their detection techniques. These datasets offer real-world examples of malicious URLs and websites, enabling researchers to test and validate their detection algorithms effectively.

III. PROBLEM STATEMENT

The use of global huge web is increasing continuously. day by day the sector huge internet will become a sufferer of net assaults like spamming, phishing and malware. whilst the harmless user unknowingly visits the URL, it turns into the sufferer of the assaults. at the same time as traveling a torrent page, you click on on a link, after which 2-3 browser windows will pop-up within the background. In different cases, you_ll get pop-us that ask you to down load a new software or browser extension. those websites run on most effective two things: site visitors and ad clicks. to maximize each, they'll use shady software program and ad networks for you to extract as many clicks as feasible from you, the end person. The verification of URLs could be very crucial on the way to make sure that person ought to be averted from visiting malicious web sites. For detecting these malicious URLs, numerous methods had been proposed. but, those techniques are sometimes time ingesting and if no longer, they do not offer higher accuracy. The increasing use of the World Wide Web has brought about numerous benefits, but it has also made users more susceptible to various forms of cyber-attacks such as spamming, phishing, and malware. These attacks often target unsuspecting users who inadvertently visit malicious URLs, leading to potentially harmful consequences. For example, when visiting a torrent website, users may encounter multiple pop-ups windows or prompts to download suspicious software or browser extensions. These websites rely on generating traffic and ad clicks to maximize their revenue, often employing deceptive tactics to trick users into engaging with their content.

In order to protect users from these malicious websites, it is crucial to verify the authenticity and safety of URLs before accessing them. Several methods have been proposed for detecting malicious URLs, but these methods are often time-consuming and may not always provide accurate results. As a result, there is a pressing need for more effective and efficient approaches to URL verification and malicious website detection. For example, researchers have developed models that analyze various features of URLs, such as token count, average path token, and domain name features, to distinguish between legitimate and malicious websites.

Additionally, researchers have explored the use of online learning algorithms, which can adapt and update their models in real-time based on new data and evolving threats. This approach is particularly well-suited for detecting malicious websites at scale, as it allows for the continuous monitoring and analysis of web traffic to identify potential threats as they emerge.

Despite these advancements, there are still challenges and limitations associated with detecting malicious URLs. For example, the rapid proliferation of short URLs on platforms like Twitter and Facebook poses a unique challenge, as these URLs often obscure the destination website, making it difficult to assess their legitimacy. Researchers continue to investigate new features and techniques for effectively detecting and

identifying malicious URLs, but more research is needed to address the evolving nature of cyber threats.

In addition to technological solutions, there is also a need for greater awareness and education among users to help them recognize and avoid malicious websites. By understanding the risks and best practices for safe browsing, users can better protect themselves from falling victim to cyber-attacks.

IV. PROPOSED WORK

Based on the findings of the survey, it is essential to develop an evolutionary approach for identifying malicious URLs. To address this need, we propose a mechanism that focuses on evaluating web page data, which includes content extraction and sentiment analysis. Our framework utilizes an evolutionary method for malicious URL detection through comprehensive content assessment. In this approach, the actual content of web pages is extracted using web crawling techniques. The extracted data is then processed with a sentiment analysis tool to evaluate potential maliciousness. The resulting insights can be used to effectively identify and blacklist harmful URLs.

One of the key advantages of machine learning-based URL analysis is its ability to provide real-time identification of potentially malicious URLs. Traditional methods of detecting phishing websites often rely on static blacklists or heuristics, which may not always be effective in identifying new or evolving threats.This proactive approach allows for timely detection and prevention of phishing attacks, enhancing user security and mitigating the risk of falling victim to fraudulent websites. Furthermore, machine learning algorithms can adapt to new threats and evolving tactics used by cybercriminals. Phishing attacks are constantly evolving, with cybercriminals employing new techniques to bypass traditional security measures. Machine learning models can be trained to recognize new patterns and behaviors associated with these threats, ensuring that the detection system remains effective and up-to-date. This adaptability is crucial in the fight against phishing attacks, as it allows for the detection of emerging threats before they can cause harm to users.

Another benefit of machine learning-based URL analysis is its ability to automate and streamline the detection process. With rapid and automated analysis of URLs, users can benefit from timely warnings and alerts when accessing potentially malicious websites. This proactive approach helps users make informed decisions and avoid falling victim to phishing scam

V. ADVANTAGES

Detection of phishing websites through machine learning- based URL analysis offers significant advantages. It provides real-time identification of potentially malicious URLs, enhancing user security by preventing access to fraudulent sites. Machine learning algorithms analyze URL patterns, leveraging historical data to identify phishing characteristics accurately. This proactive approach improves detection rates, protecting users from evolving phishing tactics. Additionally, the system adapts to new threats, ensuring continuous efficacy. With rapid and automated analysis, users benefit from timely warnings, reducing the risk of falling victim to phishing attacks. Overall, machine learning-based URL analysis strengthens cybersecurity, fortifying defenses against the ever-growing threat landscape.

Phishing attacks typically involve the use of deceptive emails, messages, or websites to trick users into divulging sensitive information such as login credentials, financial details, or personal data. These attacks often mimic legitimate websites, making it difficult for users to discern between authentic and fraudulent URLs. Traditional methods of detecting phishing websites rely on static blacklists or heuristics, which may not always be effective in identifying new or evolving threats.

Machine learning-based URL analysis, on the other hand, offers a proactive and dynamic approach to phishing detection. By analyzing large datasets of historical URL data, machine learning algorithms can learn to identify patterns and characteristics associated with phishing websites. These algorithms can then be trained to classify URLs as either legitimate or malicious based on these learned features. . One of the key advantages of machine learning-based URL analysis is its ability to provide real-time identification of potentially malicious URLs. Unlike traditional methods that rely on static databases or rule-based systems, machine learning

algorithms can continuously monitor and analyze web traffic to identify phishing threats as they emerge. This proactive approach allows for timely detection and prevention of phishing attacks, enhancing user security and mitigating the risk of falling victim to fraudulent websites.

One of the key advantages of machine learning-based URL analysis is its ability to provide real-time identification of potentially malicious URLs. Unlike traditional methods that rely on static databases or rule- based systems, machine learning algorithms can continuously monitor and analyze web traffic to identify phishing threats as they emerge.

Additionally, machine learning algorithms can adapt to new threats and evolving tactics used by cybercriminals. As phishing attacks become more sophisticated and diverse, machine learning models can be trained to recognize new patterns and behaviors associated with these threats.This high level of accuracy reduces false positives and false negatives, ensuring that users are adequately protected from phishing attacks. Another benefit of machine learning-based URL analysis is its ability to automate and streamline the detection process. With rapid and automated analysis of URLs, users can benefit from timely warnings and alerts when accessing potentially malicious websites.

VI. CONCLUSION

The proposed web crawling-based approach for web page mining demonstrates enhanced effectiveness in identifying malicious content compared to previous methods. Through experiments with the URL reputation dataset, we compared our distributed machine learning model to traditional machine learning algorithms using classification techniques. Our findings indicated that traditional machine learning algorithms outperformed their distributed counterparts when handling smaller datasets, achieving faster computation times. This efficiency is attributed to the overhead associated with Spark and Java, which introduces unnecessary complexity for smaller computations.Machine learning-based URL analysis is a critical tool in combating the ever-evolving threat of phishing attacks.These algorithms adapt to new threats and evolving tactics used by cybercriminals, ensuring that the detection system remains effective and up-to-date. With improved detection rates and automated analysis, users benefit from timely warnings and alerts when accessing potentially harmful URLs.

REFERENCES

[1] D. Sahoo, C. Liu, S.C.H. Hoi, ―Malicious URL Detection using Machine Learning: A Survey‖. CoRR, abs/1701.07179, 2017

[2] M. Khonji, Y. Iraqi, and A. Jones, ―Phishing detection: a literature survey, ‖ IEEE Communications Surveys & Tutorials, vol. 15, no. 4, pp. 2091–2121,2013.

[3] M. Cova, C. Kruegel, and G. Vigna, ―Detection and analysis of driveby- download attacks and malicious javascript code, ‖ in Proceedings of the 19th international conference on World wide web. ACM, 2010, pp. 281– 290.

[4] R. Heartfield and G. Loukas, ―A taxonomy of attacks and a survey of defence mechanisms for semantic social engineering attacks, ‖ ACM Computing Surveys (CSUR), vol. 48, no. 3, p. 37, 2015.

[5] Internet Security Threat Report (ISTR) 2019– Symantec.https://www.symantec.com/content/dam/sy mantec/docs/reports/istr-242019-en.pdf [Last accessed 10/2019].

[6] S. Sheng, B. Wardman, G. Warner, L. F. Cranor, J. Hong, and C. Zhang, ―An empirical analysis of phishing blacklists, ‖ in Proceedings of Sixth Conference on Email and Anti-Spam (CEAS), 2009.

[7] S. Sinha, M. Bailey, and F. Jahanian, ―Shades of grey: On the effectiveness of reputation-based ―blacklists‖, ‖ in Malicious and Unwanted Software, 2008. MALWARE 2008. 3rd International Conference on. IEEE, 2008, pp. 57–64

[8] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, ―Identifying suspicious urls: an application of large- scale online learning, ‖ in Proceedings of the 26th Annual International Conference on Machine Learning. ACM, 2009, pp. 681–688

[9] B. Eshete, A. Villafiorita, and K. Weldemariam, ―Binspect: Holistic analysis

and detection of malicious web pages, ‖ in Security and Privacy in Communication Networks. Springer, 2013, pp. 149– 166.

[10] S. Purkait, ―Phishing counter measures and their effectiveness– literature review, ‖ Information Management & Computer Security, vol. 20, no. 5, pp. 382–420, 2012.

[11] Y. Tao, ―Suspicious url and device detection by log mining, ‖ Ph.D. dissertation, Applied Sciences: School of Computing Science, 2014

[12] G. Canfora, E. Medvet, F. Mercaldo, and C. A. Visaggio, ―Detection of malicious web pages using system calls sequences, ‖ in Availability, Reliability, and Security in Information Systems. Springer, 2014, pp. 226–238.

[13] Leo Breiman.: Random Forests. Machine Learning 45 (1), pp. 5- 32, (2001)

[14] Thomas G. Dietterich. Ensemble Methods in Machine Learning. International Workshop on Multiple Classifier Systems, pp 1-15, Cagliari, Italy, 2000.

[15] Develop Information. https://www.phishtank.com/developer_info. php. [Last accessed 11/2019]. URLhaus Database Dump. https://urlhaus.abuse.ch/downloads/csv/. [Ngày truy nhập 11/2019].

[16] DataseURL.http://downloads.majestic.com/majest ic_ million.csv. [Last accessed 10/2019].

[17] Malicious_n_NonMaliciousURL.https://www.kag gle. com/antonyj453/ urldataset#data.csv. [Last accessed 11/2019].

[18] chrome.zip. https://drive.google.com/file/d/13G_Ndr4 hMFx_qWyTEjHuOyJmHFW D0Gud/view?fbclid=IwAR0SLVCrvjHHGmoHZ H97 nXN3BmDMY7jG4SOsKZYLA ZjTFgeoJADfli64-g. [Last accessed 12/2019]

[19] Immadisetti Naga Venkata Durga Naveen, Manamohana K, Rohit Verma, ―Detection of Malicious URLs using Machine Learning Techniques‖ in proceedings of the International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-4S2 March, 2019

[20] Doyen Sahoo, Chenghao Liu, and Steven C.H.Hoi. ―Malicious URL Detection using Machine Learning: A Survey. 1,1(August2019),37pages.

[21] Dharmaraj R. Patil and Jayantrao B. Patil, ―Survey on Malicious Web Pages Detection Techniques‖ in proceedings of the International Journal of u- and e- Service, Science and Technology, August 2015 05, Volume No. 8, Number 05 (pp.195-206) http://dx.doi.org/10.14257/ijunesst.2015.8.5.18

[22] Dharmaraj R. Patil and Jayantrao B. Patil, ―Feature- based Malicious URL and Attack Type Detection Using Multi-class Classification‖ in proceedings of the ISC.