IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

MALICIOUS BREACHES URL DETECTION USING MACHINE LEARNING

[1]Harish P, [2]Surya Dharmaraj, [3]Dhanush K, [4]Rajavel M
Department of Computer Science and Engineering
Faculty of Engineering and Technology
SRM Institute of Science and Technology, Vadapalani , Chennai, India

Abstract: The exponential expansion of internet usage has profoundly reshaped global dynamics, ushering in a new era characterized by enhanced knowl- edge dissemination, streamlined goods movement, and interconnected interpersonal relationships. However, this unprecedented connectivity has also spawned a proliferation of malicious activities, particularly in the realm of client-side attacks targeting websites. Traditional mitigation approaches, such as blacklisting, have fallen short in effectively combat- ing these sophisticated threats, prompting our project to embark on the development of a robust solution. Our endeavor involved the integration of host-based, content-based, and lexical features, culminating in the im- plementation of a random forest machine learning model. This powerful amalgamation yielded an impressive accuracy rate of 94.7 percentage un- derscoring the efficacy of advanced machine learning techniques in bol- stering cybersecurity defenses. By leveraging these diverse features, our model exhibited enhanced discriminatory capabilities, adept at detecting anomalies within hosting environments, identifying malicious elements embedded within web content, and scrutinizing the linguistic attributes of URLs.

Keywords: Machine Learning, Deep Learning, SVM, Random Forest, Artificial Intelligence etc

I. INTRODUCTION

As of November 2023, the pervasive connectivity of the internet had woven a digital fabric connecting 5.16 billion individuals globally, rendering it an indispensable facet of our daily lives. While its omnipresence has undoubtedly facilitated seamless communication and access to information, the flip side reveals an alarming vulnera- bility to cyberthreats. In the annals of 2021, a staggering 323,972 users fell victim to phishing attempts, a perilous maneuver even though Google's security measures man- aged to thwart 99.9% of such malicious endeavors. The escalating risks are further compounded by the proliferation of drive-by downloads, wherein unsuspecting users inadvertently download malware while visiting seemingly innocuous websites

The ominous trajectory of malicious URLs being exploited for illegal purposes poses a significant challenge to contemporary cybersecurity measures. Blacklists, once considered a stalwart defense, now find themselves woefully inadequate against the relentless evolution of attack techniques employed by cyber adversaries. Adding a new layer of complexity, the process of appending malicious websites to blacklists has be- come a formidable challenge, with attackers employing sophisticated methods to evade detection.In this era of heightened digital interdependence, the imperative to develop practical and robust methods to identify and halt malicious activities becomes increasIngly pronounced. As we navigate the intricate landscape of an internet-dependent society, a paradigm shift in cybersecurity strategies is imperative to confront the mul- tifaceted challenges posed by cyberthreats. The quest for innovative and adaptive solutions remains at the forefront, echoing the critical need for proactive measures to safeguard the integrity of our digital interactions in the face of evolving cyber dangers.

Artificial intelligence is the use of computer science programming to imitate human thought and action by analysing data and surroundings, solving or anticipating prob- lems and learning or self-teaching to adapt to a variety of tasks.

One of the key facets of AI is its ability to adapt and evolve over time. Through techniques such as machine learning and neural networks, AI systems can ingest vast amounts of data, recognize patterns, and make predictions or decisions based on that information. This capacity for continuous learning allows AI to refine its capabilities and improve performance with experience, much like the human brain.

Artificial intelligence (AI) is the ability of a computer program or a machine to think andlearn. It is also a field of study which tries to make computers "smart". As machines become increasingly capable, mental facilities once thought to require intelligence are removed from the definition. AI is an area of computer sciences that emphasizes the creation of intelligent machines that work and reacts like humans. Some of the activ- ities computers with artificial intelligence are designed for include: Face recognition, Learning, Planning, Decision making etc.,

II. Literature Survey

[1] **Title: Machine learning based phishing detection from URLs Authors: Ozgur Koray Sahingoz** Due to the rapid growth of the Internet, users change their preference from traditional shopping to the electronic commerce. Instead of bank/shop robbery, nowadays, criminals try to find their victims in the cyberspace with some specific tricks. By using the anonymous structure of the Internet, attackers set out new techniques, such as phishing, to deceive victims with the use of false websites to collect their sensitive information such as account IDs, usernames, passwords, etc. Understanding whether a web page is legitimate or phishing is a very challenging problem, due to its semantics-based attack structure, which mainly exploits the computer users' vulnerabilities. Although software companies launch new anti-phishing products, which use blacklists, heuris- tics, visual and machine learning-based approaches, these products cannot prevent all of the phishing attacks. In this paper, a real-time anti-phishing system, which uses seven different classification algorithms and natural language processing (NLP) based features, is proposed. The system has the following distinguishing properties from other studies in the literature: language independence, use of a huge size of phishing and legitimate data, real-time execution, detection of new websites, independence from third-party services and use of feature-rich classifiers. For measuring the performance of the system, a new dataset is constructed, and the experimental results are tested on it. According to the experimental and comparative results from the implemented clas- sification algorithms, Random Forest algorithm with only NLP based features gives the best performance with the 97.98% accuracy rate for detection of phishing URLs.)

[2] **Title: Detection of phishing websites using an effi- cient feature-based machine learning framework Authors: Routhu Srinivasa Rao** In their paper, Phishing is a cyber-attack which targets naive online users tricking into revealing sensitive information such as username, password, social security number or credit card number etc. Attackers fool the Internet users by masking webpage as a trustworthy or legitimate page to retrieve personal information. There are many anti-phishing solutions such as blacklist or whitelist, heuristic and visual similarity-based methods proposed to date, but online users are still getting trapped into revealing sensitive information in phishing websites. In this paper, we propose a novel classification model, based on heuristic features that are extracted from URL, source code, and third-party services to overcome the disadvantages of existing anti-phishing techniques. Our model has been evaluated using eight different machine learning algorithms and out of which, the Ran- dom Forest (RF) algorithm performed the best with an accuracy of 99.31%. The experi- ments were repeated with different (orthogonal and oblique) random forest classifiers to find the best classifier for the phishing website detection. Principal component analysis Random Forest (PCA-RF) performed the best out of all oblique Random Forests (oRFs) with an accuracy of 99.55%. We have also tested our model with the third-party-based features and without third-party-based features to determine the effectiveness of third- party services in the classification of suspicious

websites. We also compared our results with the baseline models (CANTINA and CANTINA+). Our proposed technique out- performed these methods and also detected zero-day phishing attacks.

[3] **Title: A machine learning based approach for phishing detection using hyperlinks information Authors: Ankit Kumar Jain** This paper presents a novel approach that can detect phishing attack by analysing the hyperlinks found in the HTML source code of the website. The proposed approach incorporates various new outstanding hyperlink specific features to detect phishing at- tack. The proposed approach has divided the hyperlink specific features into 12 dif- ferent categories and used these features to train the machine learning algorithms. We have evaluated the performance of our proposed phishing detection approach on vari- ous classification algorithms using the phishing and non-phishing websites dataset. The proposed approach is an entirely client-side solution, and does not require any services from the third party. Moreover, the proposed approach is language independent and it can detect the website written in any textual language. Compared to other methods, the proposed approach has relatively high accuracy in detection of phishing websites as it achieved more than 98.4% accuracy on logistic regression classifier.

[4] **Title: CatchPhish: detection of phishing websites by inspecting URLs Authors: Routhu Srinivasa Rao** In There exists many anti-phishing techniques which use source code-based features and third party services to detect the phishing sites. These techniques have some lim- itations and one of them is that they fail to handle drive-by-downloads. They also use third-party services for the detection of phishing URLs which delay the classification process. Hence, in this paper, we propose a light-weight application, CatchPhish which predicts the URL legitimacy without visiting the website. The proposed technique uses hostname, full URL, Term Frequency-Inverse Document Frequency (TF-IDF) features and phish-hinted words from the suspicious URL for the classification using the Ran- 6 Authors: Ankit Kumar Jain Authors: Routhu Srinivasa Rao dom forest classifier. The proposed model with only TF-IDF features on our dataset achieved an accuracy of 93.25%. Experiment with TF-IDF and hand-crafted features achieved a significant accuracy of 94.26% on our dataset and an accuracy of 98.25%, 97.49% on benchmark datasets which is much better than the existing baseline models..

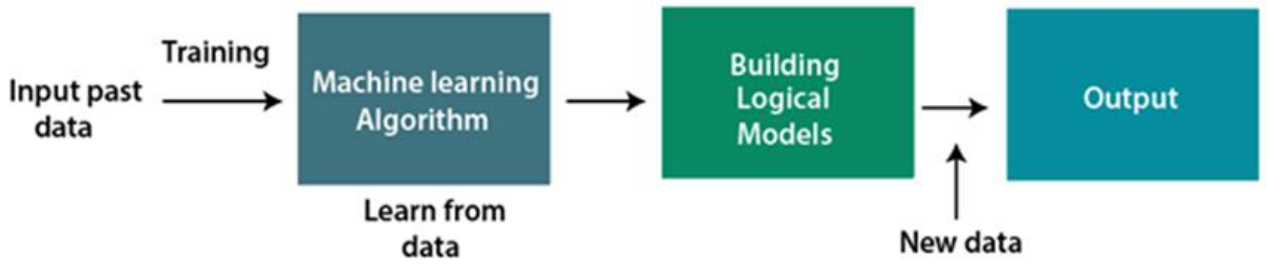
[5] **Title: An effective detection approach for phishing websites using URL and HTML features Authors: Ali Aljofey** Today's growing phishing websites pose significant threats due to their extremely undetectable risk. They anticipate internet users to mistake them as genuine ones in order to reveal user information and privacy, such as login ids, pass-words, credit card numbers, etc. without notice. This paper proposes a new approach to solve the anti- phishing problem. The new features of this approach can be represented by URL char- acter sequence without phishing prior knowledge, various hyperlink information, and textual content of the webpage, which are combined and fed to train the XGBoost clas- sifier. One of the major contributions of this paper is the selection of different new features and these features do not depend on any third-party services. In particular, we extract character level Term Frequency-Inverse Document Frequency (TF-IDF) features from noisy parts of HTML and plaintext of the given webpage. Moreover, our proposed hyperlink features determine the relationship between the content and the URL of a webpage. Due to the absence of publicly available large phishing data sets, we needed to create our own data set with 60,252 webpages to validate the proposed solution. This data contains 32,972 benign webpages and 27,280 phishing webpages. For evaluations, the performance of each category of the proposed feature set is evaluated, and various classification algorithms are employed. From the empirical results, it was observed that the proposed individual features are valuable for phishing detection. However, the integration of all the features improves the detection of phishing sites with significant accuracy. The proposed approach achieved an accuracy of 96.76% with only 1.39% false-positive rate on our dataset, and an accuracy of 98.48% with 2.09% false-positive rate on benchmark dataset, which outperforms the existing baseline approaches.

III. Methodology

The methodology for developing an AI-Powered Trauma Chat Assistance platform, aimed at identifying trauma symptoms from voice and text communications while integrating direct doctor connection functionality, is a multifaceted process that requires careful consideration of various technical, ethical, and practical aspects. This methodology encompasses several key stages, including data collection and preprocessing, algorithm development, model training, platform integration, and ethical considerations..

3.1 Machine Learning

Machine learning finds applications across a wide range of domains, including im- age and speech recognition, natural language processing, recommendation systems, and autonomous vehicles, among others. In healthcare, machine learning models analyze medical images to assist in diagnosis and treatment planning. In e-commerce, recom- mendation algorithms personalize product suggestions based on user preferences and behavior. In finance, predictive models forecast market trends and optimize investment strategies. Machine learning encompasses various techniques and algorithms, each suited to different types of tasks and data. Supervised learning involves training a model on la- beled data, where each example is paired with a corresponding target or output. The model learns to map inputs to outputs by minimizing the difference between its predic- tions and the true labels.



3.2 Random Forest

Random Forest is a powerful ensemble learning algorithm used for both classification and regression tasks in machine learning. It operates by constructing a multitude of decision trees during the training phase and outputs the mode of the classes for classi- fication tasks or the mean prediction for regression tasks. The name "Random Forest" stems from the idea that each decision tree is built using a random subset of the features and a random subset of the training data. Random Forest is a powerful machine learn- ing algorithm that belongs to the ensemble learning family. It is widely used for both classification and regression tasks due to its robustness and versatility. The key idea behind Random Forest is to build multiple decision trees during the training phase and combine their predictions to obtain more accurate and stable results.

3.3 Existing System

Traditional security measures heavily rely on signature-based methods to detect known malicious URLs, revealing inherent limitations. Their effectiveness diminishes against novel or zero-day threats, as they lack adaptability. Vulnerable to evasion techniques, these methods falter when attackers modify URLs to bypass existing signatures. Fur- thermore, their inefficiency in handling large-scale datasets results in slower detection rates. Dependency on manual rule creation and updates introduces delays in response, hindering real-time threat mitigation. The struggle to detect polymorphic URLs further compounds the issue, potentially leading to high false positives. The shortcomings of signature-based approaches underscore the need for more adaptive and scalable strate- gies in the ever-evolving landscape of cybersecurity. Moreover, traditional methods struggle to handle large-scale datasets efficiently. The sheer volume of data to be processed can overwhelm these systems, leading to slower detection rates and delayed responses to emerging threats. This inefficiency is exacerbated by the manual nature of rule creation and updates, which introduce further delays in adapting to evolving attack patterns. Another challenge faced by signature-based approaches is the detection of polymor- phic URLs. These URLs dynamically change their structure or content to evade detec- tion, making them particularly difficult to identify using static signatures. As a result, signature-based methods may generate high false positive rates, leading to unnecessary alerts and resource wastage

3.4 RESOURCE REQUIREMENTS:

A. SOFTWARE REQUIREMENTS:

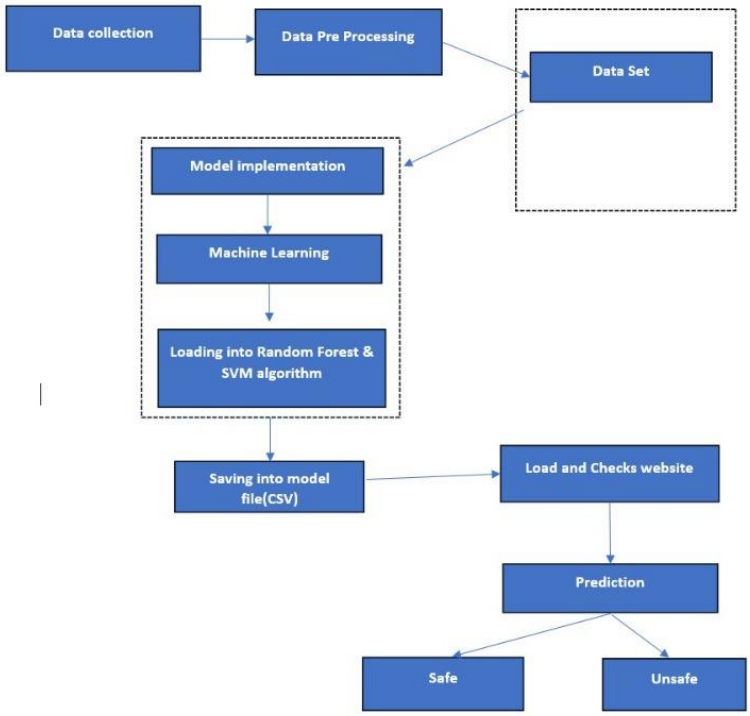
Operating System	Windows 7or later
Simulation Tool	Visual Studio Code + Browser
Documentation	Ms – Office

B. HARDWARE REQUIREMENTS:

CPU type	I5 and Above
Ram size	4GB
Hard disk capacity	80 GB
Keyboard type	Internet keyboard
Monitor type	15 Inch colour monitor
CD -drive type	52xmax

IV. System Architecture :

General Architecture:



Description

Machine learning techniques, notably Random Forest and Support Vector Machines (SVM), were utilized for the detection of malicious URLs. These algorithms were em- ployed to forecast the likelihood of websites engaging in nefarious activities. Following the training phase, the model was converted into a CSV file, facilitating its application in identifying and addressing potential security breaches on websites. This process rep- resents a significant advancement in bolstering cybersecurity measures, as it enables automated detection and mitigation of malicious activities, thereby enhancing overall web security.

vi. Testing:

The primary objective of testing is to uncover errors and ensure that a software system functions as intended. It involves a systematic process of scrutinizing every aspect of a work product to identify any potential faults or weaknesses. Testing serves as a critical mechanism to verify the functionality of components, sub-assemblies, assemblies, and the final product, ensuring it aligns with requirements and user expectations while avoiding unacceptable failures. There exist various types of tests, each addressing specific testing requirements.

Testing plays a pivotal role in software development, with its primary objective being to uncover errors and ensure that a software system functions as intended. It involves a systematic and thorough process of scrutinizing every aspect of a work product, in- cluding components, sub-assemblies, assemblies, and the final product, to identify any potential faults or weaknesses One of the key features of the platform is its integration of both voice and text communication channels, allowing users to seek help through their preferred medium. Whether it's speaking with a chatbot via text or engaging in a conversation with a voice-enabled assistant, users have the flexibility to communicate in a manner that feels most comfortable and accessible to them. This inclusivity ensures that individuals with varying communication preferences or accessibility needs can benefit from the support offered by the platform.

The importance of testing lies in its ability to verify the functionality of the software system and ensure that it aligns with specified requirements and user expectations. By detecting and addressing defects early in the development process, testing helps mit- igate risks and prevents unacceptable failures that could impact the reliability, perfor- mance, and security of the software.

Unit testing entails the creation of test cases designed to validate the internal logic of a program, ensuring that program inputs yield valid outputs. It involves testing individual software units after their completion but before integration. This form of testing, cate- gorized as structural, relies on a deep understanding of the software's construction and is intrusive in nature. Unit tests verify basic functionality at the component level, ensur- ing that each unique path of a business process conforms to documented specifications.

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

$$specificity = \frac{TN}{TN + FP}$$

- Training set - Total length of training samples divided by 100 for every trained sample
  - `int((2144 * 8)/100) = int(171.52) = 171`
- Testing set - Total length of testing samples divided by 100 for every testing sample
  - `int((460 * 8)/100) = int(36.8) = 36`

vii. Conclusion:

In conclusion, The escalating digital storage of personal information on mobile devices underscores the critical need for robust cybersecurity measures, particularly in detecting malicious URLs. The prevalence of phishing attacks, leading to substantial data and financial losses, necessitates effective detection mechanisms. This project employs supervised learning algorithms, specifically random forest and SVM, revealing that Random for- est achieves a commendable 90.61percentage accuracy in identifying malicious URLs. The research emphasizes the significance of URL attributes in distinguishing between malicious and benign entities, highlighting specific characteristics crucial for encryp- tion and disguise. The visualization of relationships between these attributes provides valuable insights into patterns associated with malicious URLs, contributing to the on- going efforts to enhance cybersecurity in the digital era.

Overall, this research contributes significantly to the ongoing efforts to enhance cyber- security in the digital era. By leveraging supervised learning algorithms and analyzing URL attributes, the project provides valuable insights and tools for detecting and miti- gating the growing threat of malicious URLs. As cyber threats

continue to evolve, the findings of this research serve as a foundation for future advancements in cybersecurity technology and practice.

viii. Future Work:

Further investigation into the performance of the supervised learning algorithms demonstrates their efficacy in distinguishing between malicious and benign URLs. Random forest and Support Vector Machine (SVM) algorithms are chosen for their ability to handle complex data and provide accurate classification results. Through rigorous experimentation and analysis, the research team identifies Random forest as the most effective algorithm, achieving an impressive accuracy rate of 90.61precent. This high accuracy rate underscores the potential of machine learning approaches in bolstering cybersecurity measures against malicious URL threats.

REFERENCES

1. Patil, Dharmaraj R., and Jayantro B. Patil. ”Malicious URLs detection using deci- sion tree classifiers and majority voting technique.” Cybernetics and Information Technologies 18, no. 1 (2018): 11-29.

2. Vinayakumar, R., K. P. Soman, and Prabaharan Poornachandran. ”Evaluating deep learning approaches to characterize and classify malicious URL’s.” Journal of Intelligent Fuzzy Systems 34.3 (2018): 1333-1343.

3. Garera, Sujata, et al. ”A framework for detection and measurement of phishing attacks.” Proceedings of the 2007 ACM workshop on Recurring malcode. 2007.

4. Mohammad, Rami M., Fadi Thabtah, and Lee McCluskey. ”An assessment of features related to phishing websites using an automated technique.” 2012 in- ternational conference for internet technology and secured transactions. IEEE, 2012.

5. Zhiwang, Cen, Xu Jungang, and Sun Jian. ”A multi-layer bloom filter for dupli- cated URL detection.” 2010 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE). Vol. 1. IEEE, 2010.

6. Khonji, Mahmoud, Youssef Iraqi, and Andrew Jones. ”Phishing detection: a literature survey.” IEEE Communications Surveys Tutorials 15.4 (2013): 2091- 2121.

7. Sahoo, Doyen, Chenghao Liu, and Steven CH Hoi. ”Malicious URL detection using machine learning: a survey. CoRR abs/1701.07179 (2017).” (2017).

8. Vazhayil, Anu, R. Vinayakumar, and K. P. Soman. ”Comparative study of the de- tection of malicious URLs using shallow and deep networks.” 2018 9th Interna- tional Conference on Computing, Communication and Networking Technologies (ICCCNT). IEEE, 2018

9. Srinivasan, S., Vinayakumar, R., Arunachalam, A., Alazab, M., Soman, K. P. (2021). DURLD: Malicious URL detection using deep learningbased charac- ter level representations. Malware analysis using artificial intelligence and deep learning, 535-554

10. Menon, R.R.K., Akhil Dev, R., Bhattathiri, S.G., ”An insight into the relevance of word ordering for text data analysis.” 2020 fourth international conference on computing methodologies and communication (ICCMC). IEEE, 2020