iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** V **Month of publication:** May 2024

DOI: https://doi.org/10.22214/ijraset.2024.62083

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue V May 2024- Available at www.ijraset.com

Detection of Phishing Websites by Using URL Analysis

Utsav Jagdale[1], Siddhant Pawar[2], Prof. Moushmi Kuri[3]
Computer Science Department, MIT ADT University, Pune

Abstract: ***In recent years, with the increasing use of mobile devices, there is a growing trend to move almost all real-world operations to the cyberworld. Although this makes easy our daily lives, it also brings many security breaches due to the anonymous structure of the Internet. Used antivirus programs and firewall systems can prevent most of the attacks. However, experienced attackers target on the weakness of the computer users by trying to phish them with bogus webpages. These pages imitate some popular banking, social media, e-commerce, etc. sites to steal some sensitive information such as, user-ids, passwords, bank account, credit card numbers, etc. Phishing detection is a challenging problem, and many different solutions are proposed in the market as a blacklist, rule-based detection, anomaly-based detection, etc. In the literature, it is seen that current works tend on the use of machine learning-based anomaly detection due to its dynamic structure, especially for catching the "zero-day" attacks. In this paper, we proposed a machine learning-based phishing detection system by using eight different algorithms to analyse the URLs, and three different datasets to compare the results with other works. The experimental results depict that the proposed models have an outstanding performance with a success rate.***

I. INTRODUCTION

Phishing remains one of the most prevalent cyber threats, targeting unsuspecting online users to obtain confidential information for fraudulent purposes. Despite efforts to raise awareness and maintain blacklists of known phishing websites, users continue to fall victim to these deceptive tactics. To address this challenge, machine learning (ML) and deep neural network algorithms have emerged as effective tools for early detection of phishing attempts. This project focuses on training ML models and deep neural networks using datasets comprising both phishing and benign URLs, extracting relevant features from URLs and website content to predict phishing websites. The evolution of phishing as a major cybersecurity threat underscores the urgency of finding solutions to mitigate its impact, given the potential for financial loss, compromised personal information, and reputational damage suffered by victims. Traditional blacklisting methods have limitations, prompting the exploration of ML-based approaches for more comprehensive and timely detection of phishing attempts.

The increasing sophistication of phishing attacks, coupled with their ease of execution, highlights the need for robust detection systems capable of identifying fraudulent websites early in their deployment. Phishing attacks have evolved from simple email-based tactics to more elaborate schemes involving fake websites designed to mimic trusted sources. Attackers capitalize on the trustworthiness of familiar brands and communication channels to deceive users into divulging sensitive information. ML-based systems offer a promising avenue for enhancing cybersecurity by enabling rapid detection of phishing attempts without reliance on outdated blacklists. By leveraging ML algorithms to analyse URL structures and website content, this project aims to provide a proactive defense against phishing attacks, mitigating the financial and reputational risks faced by individuals and organizations.

II. LITERATURE SURVEY

1) Altyeb Taha," Intelligent Ensemble Learning Approach for Phishing Website Detection Based on Weighted Soft Voting"

The rise of network technologies has facilitated various aspects of daily life, including e-commerce, online banking, and social media. However, this advancement has also led to an increase in phishing websites, posing a significant cybersecurity threat. Phishing sites mimic legitimate web pages to deceive users into divulging sensitive information. Detecting these sites accurately is challenging due to their dynamic nature. To address this, ensemble learning, which combines predictions from multiple classifiers, offers a state-of-the-art solution. Altyeb Taha proposes an intelligent ensemble learning approach for phishing website detection using weighted soft voting. By employing four heterogeneous machine-learning algorithms as base classifiers and a novel weighted soft voting method, this approach achieved a remarkable accuracy of 95% and an Area Under the Curve (AUC) of 98.8% in experiments conducted on a dataset from the UCI Machine Learning Repository.

©IJRASET: All Rights are Reserved

The proliferation of online services and e-commerce has made internet users more susceptible to phishing attacks, where fraudulent websites imitate official pages to steal personal information. With phishing attacks resulting in significant financial losses and reputational damage, the need for robust anti-phishing methods is paramount. Reports indicate a steady increase in phishing attacks, with over 146,994 phishing websites discovered in the second quarter of 2020 alone. Kaspersky Lab's anti-phishing systems stopped over 482 million phishing threats in 2018, highlighting the scale of the problem. Given the anticipated average cost of a business breach caused by phishing attacks in 2020 was 2.8 million USD, effective anti-phishing measures are crucial to mitigate such losses.

2) Shwetha. Kavitha," Detection of Phishing Websites Using Machine Learning"

Phishing, a type of social engineering attack, exploits vulnerabilities in system processes induced by system users. Despite technical security measures, such as protection against password theft, end users remain susceptible to manipulation. Attackers may employ tactics like sending false password update requests through fake websites, exploiting users' trust to compromise system security. Addressing this multifaceted issue requires interventions at both technical and human layers. While technical vulnerabilities like DNS cache poisoning can enhance the persuasiveness of socially-engineered messages, human factors make phishing attacks difficult to mitigate. Even with user awareness programs, a significant percentage of phishing attacks go undetected, underscoring the challenge of defending against these threats.

In response to the evolving nature of phishing attacks, definitions are broadening to encompass semantic attacks across various electronic communication channels. Phishing is now understood as a tactic that communicates socially engineered messages to persuade victims into actions benefiting attackers. These actions, ranging from submitting login credentials to accessing malicious links, exploit users' trust and create a perceived urgency, compelling them to comply. This expanded definition reflects the nuanced strategies employed by attackers, who may not always impersonate third parties but manipulate users into compromising actions for their benefit.

3) Pratik Patil , Devale," A Literature Survey of Phishing Attack Technique"

This paper delves into the detection and prevention of phishing attacks, emphasizing the need to protect users from fraudulent websites aiming to steal data, money, or personal information. It proposes using real-time URL analysis based on distinctive properties extracted from the URL's components. Machine learning classification is employed to identify phishing URLs from a dataset. Additionally, techniques like AntiPhish are suggested to warn users about untrusted websites. The paper acknowledges the challenges of phishing detection due to evolving tactics but proposes using machine reliability and data mining techniques for effective identification. It concludes by presenting an intelligent model based on fuzzy logic and data mining algorithms to detect e-banking phishing websites, aiming to address the complexities of assessing such threats.

III. SYSTEM ANALYSIS

A. Existing System

Existing phishing detection methods based on text or visual similarities are easily bypassed. A proposed technique analyses website signatures, comprising distinct texts and images, to identify real domain names, claiming high accuracy. Aaron Blum et al. suggest confidence-weighted classification with content-based URL detection for dynamic phishing domain detection, claiming improved protection against zero-hour threats compared to reactive blacklisting. However, reliance on certain characteristics in zero-hour attacks leads to high false positives. Phishers exploit techniques like invisible "iframe" tags to insert webpages, allowing input of sensitive data. These challenges underscore the need for robust, adaptive phishing detection systems that can effectively counter evolving threats in real-time.

B. Proposed System

The most common phishing attack involves cybercriminals impersonating reputable entities or organizations to steal sensitive information like login credentials, bank account details, or credit card information. These attacks lack sophistication, often targeting large numbers of users with bulk emails containing phishing URLs or infected attachments. The goal is to deceive and impersonate, creating panic and urgency in victims to divulge personal information upon opening such emails or visiting malicious URLs.

Victims of these scams suffer monetary losses, privacy breaches, and reputational damage. Consequently, finding timely solutions to mitigate such security threats is essential. While traditional methods rely on blacklists to detect phishing websites, they have limitations. Blacklists may not cover all malicious sites and struggle to identify newly generated ones.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue V May 2024- Available at www.ijraset.com

In response, recent advancements in machine learning have been employed for more effective classification and detection of phishing websites, offering promising avenues for enhancing cybersecurity measures against these prevalent threats.

IV. SYSTEM SPECIFICATION

All computer software needs certain hardware components or other software resources to be present on a computer to be used efficiently. These prerequisites are known as (computer) system requirements and are often used as a guideline as opposed to an absolute rule.

A. Hardware Requirements

1)	Processor	:	i3
2)	RAM	:	4G or more
3)	Hard-Disk Drive	:	500 GB

B. Software Requirements

1)	Development Platform	:	Windows 10
2)	Front End	:	Python, Pandas, NumPy, Matplot, Sklearn
3)	Back-End	:	Dataset

V. PROJECT DESCRIPTION

A. Problem Definition

Phishing attacks, seeking to obtain sensitive user data like usernames and passwords, remain a significant cybersecurity concern. This study focuses on employing machine learning to detect phishing URLs by analyzing various features. Decision Tree, Random Forest, and Support Vector Machine algorithms are utilized with the aim of identifying the most effective approach based on accuracy rates and false positives. Phishing presents a substantial threat, particularly as fake websites closely mimic legitimate ones, making it difficult for users to discern the difference. The financial impact is significant, with businesses in the United States alone losing billions annually to phishing attacks. Enhancing detection techniques is crucial due to the success of these attacks, often facilitated by users' lack of awareness about phishing threats. Since phishing attack exploits the weaknesses found in users, it is very difficult to mitigate them but it is very important to enhance phishing detection techniques.

B. Overview Of The Project

The widespread adoption of mobile devices has led to a significant shift towards online operations, introducing convenience but also exposing vulnerabilities inherent in cyberspace. While antivirus programs and firewalls offer protection against many cyber threats, sophisticated attackers exploit user weaknesses through phishing schemes, utilizing fake webpages that mimic reputable sites to steal sensitive information such as login credentials and financial details. Current research leans towards machine learning-based anomaly detection, offering dynamic solutions capable of identifying "zero-day" attacks. This paper proposes a machine learning-driven phishing detection system employing eight algorithms to analyse URLs, demonstrating exceptional performance compared to existing solutions across three datasets. Phishing remains a prevalent cybersecurity threat, with attackers employing deceptive tactics via email or social media to trick users into divulging personal information under the guise of reputable entities. Exploiting users' lack of awareness, phishing attacks often lead to financial losses and reputational damage for both individuals and organizations. Despite advancements in digital platforms facilitating various aspects of daily life, the prevalence of cyber threats underscores the importance of user vigilance and robust security measures to mitigate potential attacks orchestrated by cybercriminals seeking financial gain or data exploitation.

C. Module

- Presence of IP address in URL Presence of @ symbol in URL
- Number of dots in Hostname
- Prefix or Suffix separated by (-) to domain
- URL redirection
- Information submission to email
- URL Shortening Services “TinyURL”

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

- Length of Host name
- Presence of sensitive words in URL

1) Presence of IP address in URL: To determine if a URL contains an IP address, we set a feature to 1 if an IP address is present and 0 otherwise. Typically, benign websites do not employ IP addresses in their URLs for webpage retrieval. The presence of an IP address in a URL often signifies malicious intent, suggesting an attempt to pilfer sensitive data.

2) Presence of @ symbol in URL: We designate a feature as 1 if the URL contains an "@" symbol, and 0 otherwise. Phishers frequently incorporate the "@" symbol in URLs, causing browsers to disregard preceding content, often leading to the real address following the "@" symbol.

3) Number of dots in Hostname: We set a feature to 1 if a URL contains more than three dots, indicating a potentially suspicious URL, such as http://shop.fun.amazon.phishing.com. Here, "phishing.com" constitutes the actual domain name, while the inclusion of "amazon" aims to deceive users. In benign URLs, the average number of dots is typically three or fewer.

4) Prefix or Suffix separated by (-) to domain : We designate a feature as 1 if the domain name is separated by a dash (-) symbol, indicating a potential phishing attempt. Legitimate URLs seldom utilize the dash symbol in their domain names. Phishers incorporate dashes to deceive users into believing they are interacting with a legitimate webpage. For instance, while the actual site may be http://www.onlineamazon.com, a phisher could create a fake website like http://www.online-amazon.com to confuse unsuspecting users.

5) URL redirection: A feature is assigned a value of 1 if "//" is found in the URL path, indicating a redirection to another website. Otherwise, it is set to 0. The presence of "//" in the URL path signifies an impending redirection away from the current webpage.

6) Information submission to Email: If the URL contains functions like "mail()" or "mailto:", indicating potential redirection of user information to a personal email by a phisher, the feature is set to 1; otherwise, it remains at 0.

7) URL Shortening Services “TinyURL”: The feature is set to 1 if the URL is generated using URL shortening services such as bit.ly, as this is often employed by phishers to obscure lengthy phishing URLs and redirect users to malicious websites. Otherwise, it is set to 0.

8) Length of Host name: If the length of a URL exceeds 25 characters, the feature is set to 1; otherwise, it remains at 0. This criterion is based on the observation that the average length of benign URLs is 25 characters.

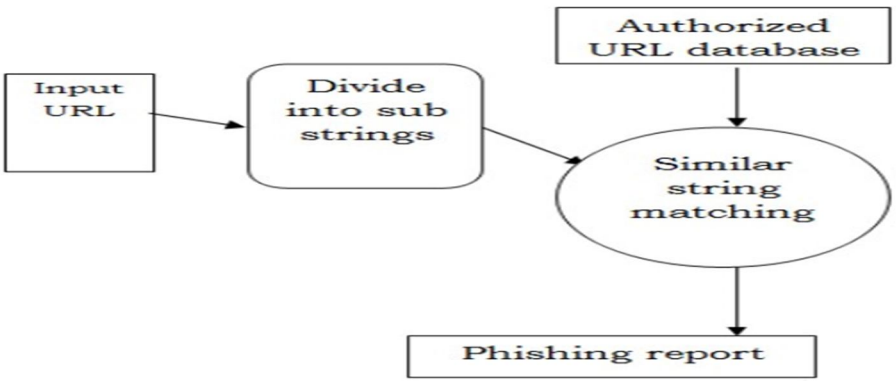
9) Presence of sensitive words in URL: Phishing websites often incorporate specific sensitive words into their URLs to create an illusion of legitimacy. These words include 'confirm', 'account', 'banking', 'secure', 'ebyisapi', 'webscr', 'signin', 'mail', 'install', 'toolbar', 'backup', 'paypal', 'password', 'username', and others.

VI. SYSTEM SPECIFICATION

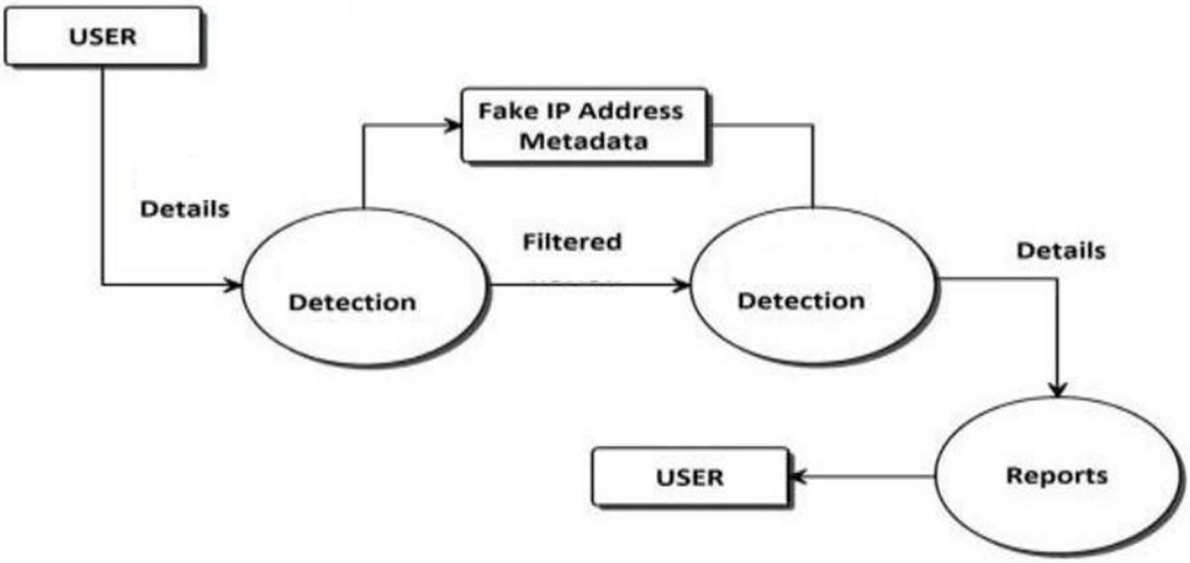
A. Data Flow Diagram

A database is an orderly compilation of interconnected data, providing a foundation for retrieving desired information or processing data. The fundamental aspect of constructing an application system lies in designing tables. The data flow diagram categorizes system requirements into significant transformations, which later evolve into programs during system design. It serves as the initial stage of the design phase, breaking down required specifications into finer levels of detail. It comprises a sequence of interconnected bubbles linked by lines.

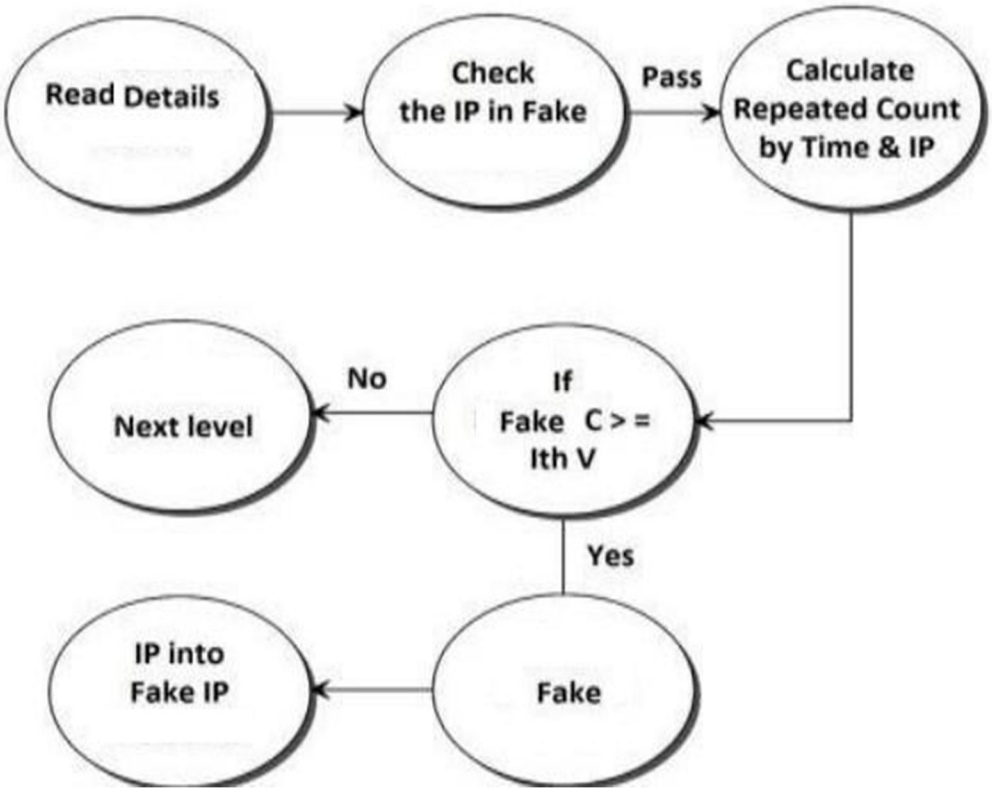
1) LEVEL 0 DATA FLOW DIAGRAM:



2) LEVEL 1 DATA FLOW DIAGRAM :



3) LEVEL 2 DATA FLOW DIAGRAM:

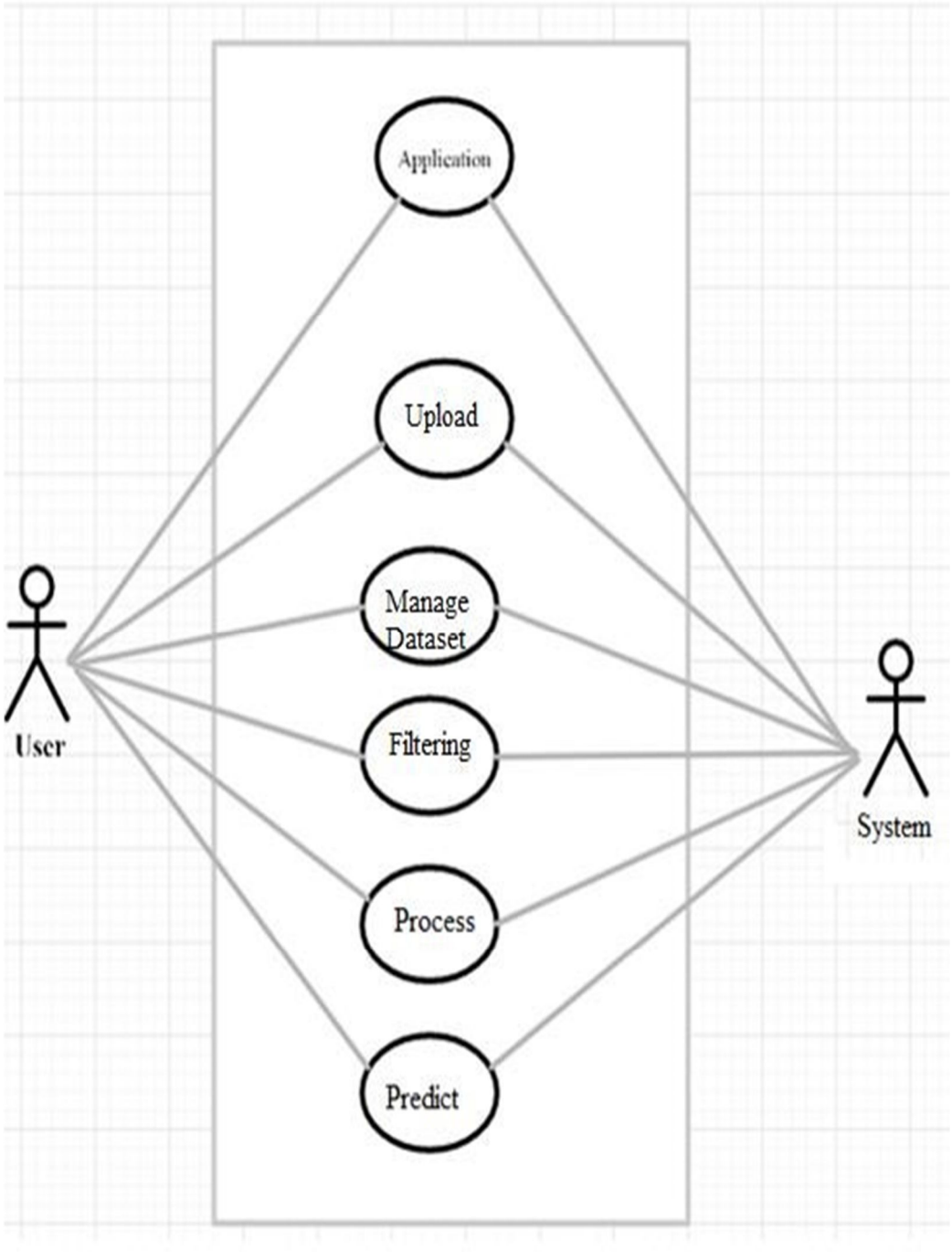


B. Use Case Diagram

At its core, a use case diagram illustrates a user's engagement with the system, displaying the connection between the user and various use cases they participate in. It delineates the diverse users of a system along with their respective use cases and is frequently complemented by additional diagram types. The provided figure illustrates the use case diagram for the system.

The following figure shows the use case diagram:

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue V May 2024- Available at www.ijraset.com



VII. SYSTEM IMPLEMENTATION

A. *System Maintenance*

System implementation is the important stage of project when the theoretical design is tuned into practical system. The main stages in the implementation are as follows:

1) Planning
2) Training
3) System testing and Changeover Planning

©IJRASET: All Rights are Reserved

Planning initiates the system implementation process. During implementation, individuals from various departments and system analysts are involved to address practical challenges in managing activities beyond their respective data processing departments. Line managers are overseen by an implementation coordinating committee.

This committee addresses ideas, issues, and user department complaints, while also considering:

- The system environment implications.
- Self-selection and allocation of implementation tasks.
- Consultation with unions and available resources.
- Standby facilities and communication channels.

B. Training

To realize the objectives and advantages of a computer-based system, it's crucial for involved individuals to have confidence in their roles within the new system. This entails comprehending the system's overall function and its impact on the organization, as well as effectively executing their designated tasks. Therefore, training should commence early on. Training sessions should equip user staff with the necessary skills for their new responsibilities.

C. System Testing

The implementation stage is crucial for ensuring the accurate and effective operation of the system before it goes live. It serves as a validation process to confirm correctness and provides an opportunity to demonstrate to users that the system has been thoroughly tested with test data, ensuring it will perform successfully and deliver expected outcomes under normal conditions. Prior to implementation, the proposed system should undergo testing with raw data to validate the functionality of its modules. Testing with valid data is essential to achieve the system's objectives. System testing aims to detect and rectify errors in the system under consideration. Despite its importance, this phase is often compromised, typically due to project scheduling constraints or user eagerness to proceed directly to conversion. However, testing is indispensable for ensuring that the system's components operate correctly, thereby facilitating the successful attainment of its objectives. This highlights two key issues.

- The time lag between the cause and appearance of the problem
- The effect of system errors on files and records within the system, a small system error can conceivably explode into much larger problem. Effectively early in the process translates directly into long term cost savings from a reduced number of errors.

D. Changeover

Changeover occurs when:

- Phishing websites, a prevalent social engineering tactic, replicate trustworthy Uniform Resource Locators (URLs) and webpages.
- The project's goal is to train machine learning models and deep neural networks using a dataset created to forecast phishing websites.
- A dataset comprising both phishing and benign URLs is compiled, from which essential URL and website content-based features are extracted.
- The performance of each model is assessed and compared.

VIII. CONCLUSION

In this project, we've explored the efficacy of classifying phishing URLs within a dataset containing both benign and malicious URLs. We've covered various aspects including dataset randomization, feature engineering, and extraction techniques such as lexical analysis and host-based features, alongside statistical analysis. Comparative studies using different classifiers have shown consistent findings, with dataset randomization notably enhancing classifier accuracy.

We've employed a straightforward approach to feature extraction using regular expressions, but there's room for experimenting with additional features to potentially improve system accuracy. Given that the dataset used may be slightly outdated, continuous training with updated data would significantly enhance model performance.

While we haven't utilized content-based features due to challenges such as the transient nature of phishing websites and limited data availability, we acknowledge the potential of content analysis for future enhancements.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

We aim to integrate rule-based prediction based on URL content analysis, complementing the classification-based lexical analysis approach. This combined approach would offer a comprehensive solution for detecting phishing URLs effectively.

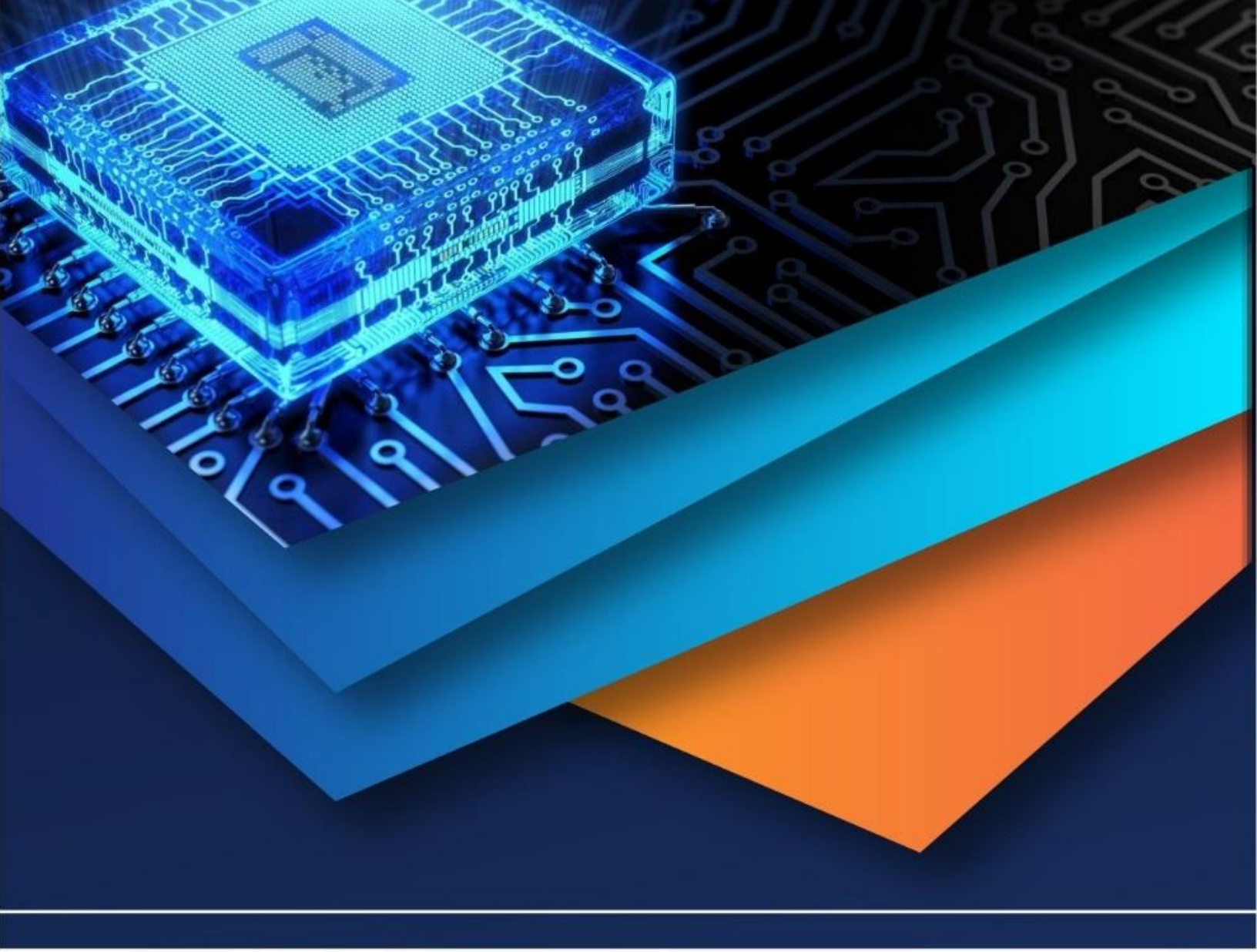
IX. ACKNOWLEDGEMENT

We are very grateful to our guide Professor Moushmi Kuri, MIT School of Computer Science, Pune for providing us with an environment to complete our project successfully. I express my sincere thanks to Dr G.R Pathak, Head of Department of Computer Science and Engineering, along with S.P. Phansalkar for their constant encouragement and support throughout the course of the project period. I am also thankful to all other Faculties for their valuable assistance in carrying out this project work. We appreciate help from other instructors who offered much needed support. Finally, we are grateful to MIT ADT University for giving us this opportunity.

REFERENCES

[1] Li, Y., Cao, Z., & Jin, H. (2017). Detecting phishing websites based on URL entropy and neural network. In 2017 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom) (pp. 182-185). IEEE.
[2] Kumar, A., & Panghal, A. (2020). Phishing Website Detection using URL Classification. In 2020 7th International Conference on Signal Processing and Integrated Networks (SPIN) (pp. 439-444). IEEE.
[3] Singh, S., & Bhatia, S. (2021). Phishing Detection using URL and content based Analysis. In 2021 IEEE 11th International Advance Computing Conference (IACC) (pp. 424-429). IEEE.
[4] Shah, S., & Hussain, M. (2018). Detecting phishing URLs using Machine Learning approach. In 2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON) (pp. 115-119). IEEE.
[5] Singh, H., Goyal, R., & Goyal, A. (2018). An Improved Method for Phishing Website Detection using URL Classification and Machine Learning Techniques. In 2018 International Conference on Recent Innovations in Electrical, Electronics & Communication Engineering (ICRIEECE) (pp. 1-5). IEEE.
[6] Kandhro, A. H., & Deboch, B. T. (2019). Phishing Website Detection Using Machine Learning Techniques. In 2019 International Conference on Computing, Electronics & Communications Engineering (iCCECE) (pp. 1-6). IEEE.
[7] Zhang, Z., & Chen, J. (2018). Detecting phishing websites using URL features and deep learning algorithms. In 2018 International Conference on Identification, Information, and Knowledge in the Internet of Things (IIKI) (pp. 147-150). IEEE.
[8] Singh, A., & Kumar, V. (2020). Phishing URL Detection Using Machine Learning Techniques. In 2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC) (pp. 179-184). IEEE.
[9] Narang, K., & Kumar, A. (2021). Phishing Website Detection using URL Entropy and Machine Learning Techniques. In 2021 International Conference on Intelligent Sustainable Systems (ICISS) (pp. 417-422). IEEE.
[10] Aljawarneh, S. A., & Alhawari, S. (2020). Detection of Phishing Websites Using Machine Learning Techniques. In Proceedings of the Future Technologies Conference (pp. 663-675). Springer, Cham.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)