IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 5, May 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205102 |

Phish Protector: Phishing Website Detector

Ms. Lathika .B.A, Santhosh Kumar .K, Saran Raj .V, Vinoth .P, Vishal .R

Assistant Professor, Dept. of CSE, KGISL Institute of Technology, Coimbatore, Tamil Nadu, India

UG Student, Dept. of CSE.,KGISL Institute of Technology, Coimbatore, Tamil Nadu, India

ABSTRACT: Phishing is a cybercrime technique used by malicious actors to deceive individuals into revealing sensitive information such as usernames, passwords, credit card details, and other personal information. It typically involves sending deceptive emails or messages that appear to come from legitimate sources, such as banks, social media platforms, or online retailers. These messages often contain links to fake websites that mimic the appearance of genuine sites, tricking users into entering their credentials. The primary goal of phishing attacks is to steal personal or financial information for fraudulent purposes, such as identity theft, unauthorized access to accounts, or financial fraud. Phishing attacks can also be used to distribute malware or ransomware, compromising the security of victims' devices and networks. Phishing attacks rely on social engineering techniques to manipulate users into taking action, such as clicking on malicious links or downloading infected attachments. These attacks often exploit human psychology, such as fear, urgency, or curiosity, to increase the likelihood of success. To protect against phishing attacks, individuals and organizations should be vigilant and cautious when interacting with emails, messages, or websites. They should verify the legitimacy of any sensitive information, such as contacting the purported sender through official channels to confirm the authenticity of the communication. Furthermore, users should be cautious about clicking on links or downloading attachments from unfamiliar or suspicious sources. Employing security measures such as spam filters, antivirus software, and two-factor authentication can also help mitigate the risk of falling victim to phishing attacks. Overall, phishing poses a significant threat to individuals, businesses, and organizations worldwide.

KEYWORDS: Phishing, Website Detection, Machine Learning, Feature Extraction, Classification, Data Preprocessing, Random Forest, Accuracy, Precision, Recall, Scalability, Real Time Detection, Threat Intelligence.

I. INTRODUCTION

Decision trees, random forests, and support vector machines (SVMs) are popular machine learning algorithms used for classification and regression tasks. Decision trees are tree-like structures where each internal node represents a feature or attribute, each branch represents a decision based on that feature, and each leaf node represents the outcome or class label. Decision trees are easy to interpret and understand, making them suitable for both classification and regression tasks. Random forests are an ensemble learning technique that builds multiple decision trees and combines their predictions to improve accuracy and robustness. Each tree in the random forest is trained on a random subset of the training data and makes its predictions independently. The final prediction is determined by a majority vote or averaging of the predictions of all the trees. Random forests are known for their high performance, scalability, and resistance to overfitting. Support vector machines (SVMs) are a powerful supervised learning algorithm used for classification and regression tasks. SVMs work by finding the optimal hyperplane that separates the different classes in the feature space while maximizing the margin between the classes. SVMs are effective in high-dimensional spaces and are particularly useful when the data is not linearly separable, as they can use kernel functions to map the data into a higher-dimensional space where it becomes separable. SVMs are known for their versatility, efficiency, and ability to handle complex datasets. In summary, decision trees, random forests, and support vector machines are three widely used machine learning algorithms that excel in different types of tasks and datasets. Decision trees offer simplicity and interpretability, random forests provide robustness and high performance through ensemble learning, and support vector machines offer versatility and efficiency in handling complex data. Each algorithm has its strengths and weaknesses, and the choice of algorithm depends on the specific requirements and characteristics of the problem at hand. Phishing website detection relies on various features that can help distinguish between legitimate and fraudulent websites. These features can be categorized into URL-based features, content-based features, and behavioral features. Here are some commonly used features for phishing website detection. URL Length: Phishing URLs tend to be longer, with additional subdomains or subdirectories. Domain Age: Phishing websites often have newly registered domains, whereas legitimate websites tend to have a longer history. Use of IP Address, Phishing sites sometimes employ IP addresses instead of domain names in their URLs. Subdomains and Redirects: Phishing websites may use multiple subdomains or frequent redirects to obscure the actual domain. Domain Name Similarity: Phishers may use domain names that are visually similar to well-known brands or legitimate websites, intending to deceive users.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205102 |

II. LITERATURE SURVEY

Several studies have focused on feature engineering and selection techniques to enhance the performance of machine learning models in detecting phishing websites. For instance, Jiang et al. (2019) explored the effectiveness of various features, including URL attributes, HTML content, and domain registration information, in differentiating between legitimate and phishing websites. They found that a combination of features, such as URL length, presence of hyphens, and domain age, significantly improved model accuracy.

In addition to feature engineering, researchers have investigated different machine learning algorithms for phishing detection. Gradient Boosting Classifiers, such as XGBoost and LightGBM, have gained popularity due to their ability to sequentially combine weak learners and improve predictive accuracy (Chen & Guestrin, 2016). These algorithms have demonstrated superior performance compared to traditional classifiers like Logistic Regression and Random Forests (Alomari et al., 2020).

Furthermore, ensemble methods, such as stacking and boosting, have been explored to further enhance model performance. For instance, Zhou et al. (2018) proposed a stacked ensemble model that combines multiple base learners, including decision trees, neural networks, and support vector machines, to improve phishing detection accuracy.

III. EXISTING SYSTEM

Phishing is a form of cyber attack where attackers use deceptive techniques to trick individuals into revealing sensitive information such as passwords, credit card numbers, or personal data. In an existing phishing system, attackers typically employ various methods to lure victims into divulging their confidential information. These methods may include sending fraudulent emails that appear to be from legitimate organizations, creating fake websites that mimic the look and feel of trusted sites, or using social engineering tactics to manipulate individuals into providing their sensitive data. Once a victim falls for the phishing attempt and provides their information, it is often used for malicious purposes such as identity theft, financial fraud, or unauthorized access to accounts. Phishing attacks can target individuals, businesses, or even government entities, posing a significant threat to cybersecurity worldwide. To maximize their success, attackers continuously evolve their phishing techniques, adapting to changes in technology and user awareness. They may employ sophisticated tactics such as spear phishing, where emails are tailored to specific individuals or organizations, or whaling, which targets high-profile individuals such as executives or celebrities. Despite advancements in cybersecurity measures and awareness campaigns, phishing remains a prevalent threat, costing organizations billions of dollars annually in losses and damages. Combating phishing requires a multi-layered approach, including robust email filtering systems, user education and awareness training, two-factor authentication, and regular security updates. Additionally, organizations must remain vigilant and proactive in detecting and responding to phishing attempts to mitigate the risks posed by this pervasive threat.

Demerits of Existing System

- Limited Adaptability: Rule-based systems rely on predefined rules or heuristics to detect phishing websites. These rules may become outdated or ineffective against evolving phishing tactics, limiting the system's adaptability to new threats.

- High False Positive Rates: Rule-based systems and some traditional machine learning models may suffer from high false positive rates, leading to the misclassification of legitimate websites as phishing sites. This can result in user frustration and reduced trust in the detection system.

- Overfitting: Certain machine learning models, such as decision trees or neural networks, are prone to overfitting, where the model learns to memorize the training data rather than generalize to unseen data. This can result in poor performance on new, unseen phishing websites.

- Resource Intensive: Some machine learning algorithms, especially complex models like deep neural networks, require significant computational resources and memory to train and deploy. This can be a barrier for organizations with limited resources or infrastructure

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205102 |

IV. METHODOLOGY AND DISCUSSION

Methodology:

1. Data Collection:

- Gather a diverse dataset containing examples of both legitimate websites and phishing websites. Sources for dataset collection may include public repositories, phishing databases, and web crawling techniques.
- Ensure the dataset is representative of real-world scenarios and encompasses a wide range of features and characteristics.

2. Data Preprocessing:

- Clean and preprocess the dataset to remove irrelevant information, handle missing values, and standardize data formats.
- Perform feature extraction to derive relevant features from the website data, including URL attributes, HTML content, domain registration information, etc.
- Split the preprocessed dataset into training and testing sets to facilitate model development and evaluation.

3. Model Selection and Training:

- Choose an appropriate machine learning algorithm for phishing detection. Popular choices include Gradient Boosting Classifiers (e.g., XGBoost, LightGBM), Random Forests, and Neural Networks.
- Train the selected model using the training dataset. During training, the model learns to differentiate between legitimate and phishing websites based on the extracted features.

4. Model Evaluation:

- Evaluate the trained model's performance using appropriate evaluation metrics such as accuracy, precision, recall.
- Use the testing dataset to assess the model's ability to generalize to new, unseen data.
- Analyze model performance across different metrics and compare it with baseline or existing approaches to assess its effectiveness.

5. Model Deployment:

- Deploy the trained model into production for real-time phishing detection. This may involve integrating the model into existing cybersecurity systems, web browsers, or standalone applications.
- Implement mechanisms for model monitoring and logging to track performance metrics and detect anomalies.
- Ensure model scalability, reliability, and security in deployment environments.

6. Continuous Improvement:

- Regularly update the model using new data and retraining techniques to adapt to evolving phishing tactics and emerging threats.
- Incorporate feedback from users and stakeholders to refine the model and enhance its effectiveness over time.

Discussion:

The discussion of the Phishing website detection project involves analyzing its strengths, limitations, and potential areas for improvement. Here's a structured discussion covering these aspects:

- Utilization of Machine Learning: Leveraging machine learning algorithms enables automated detection of phishing websites, enhancing efficiency and accuracy.
- Comprehensive Testing: Thorough testing strategies, including unit testing, integration testing, and acceptance testing, ensure the reliability and effectiveness of the system.
- User-Friendly Interface: A well-designed front-end interface enhances user experience, facilitating easy interaction and understanding of the system's features.
- Proactive Security Measures: By detecting and mitigating phishing threats, the project contributes to enhancing online security and protecting users from potential cyberattacks.
- Model Performance: The accuracy and effectiveness of the machine learning models may be affected by the quality of training data, feature selection, and model tuning.
- Evolving Threat Landscape: Phishing techniques are constantly evolving, requiring continuous updates and adaptation of detection mechanisms to stay effective against new threats.
- False Positives/Negatives: The system may encounter false positives (legitimate websites flagged as phishing) or false negatives (phishing websites not detected), impacting user trust and system credibility.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205102 |

- Enhanced Feature Extraction: Exploring advanced feature extraction techniques and integrating dynamic threat intelligence feeds can improve the system's detection capabilities.
- Real-Time Detection: Implementing real-time detection capabilities enables proactive identification and blocking of phishing websites as they emerge.
- User Feedback Mechanisms: Incorporating user feedback mechanisms allows users to report suspected phishing websites, enhancing the system's accuracy and responsiveness.

Overall, the Phishing website detection project demonstrates significant potential in enhancing online security and protecting users from phishing attacks. Addressing its limitations and exploring future directions can further strengthen its effectiveness and resilience in combating evolving cybersecurity threats.

V. PROPOSED SYSTEM

Proposing a system for phishing involves designing and implementing a deceptive scheme aimed at tricking unsuspecting individuals into divulging sensitive information such as login credentials, personal data, or financial details. The system typically consists of several components, including a fraudulent website or email, social engineering tactics, and mechanisms for harvesting and exploiting the stolen information. Firstly, the system would involve creating a fake website or email that closely mimics a legitimate entity, such as a bank, social media platform, or online retailer. This website or email would be designed to appear authentic, using logos, branding, and messaging consistent with the target organization. By impersonating a trusted source, the phishing attempt aims to lure victims into believing that they are interacting with a legitimate entity. Social engineering tactics play a crucial role in persuading victims to take the desired actions. Phishing emails may employ urgency or fear tactics, such as warning of account suspension or security breaches, to prompt recipients to click on malicious links or download attachments. Similarly, phishing websites may use convincing prompts or forms to trick users into entering their credentials or personal information. To maximize the effectiveness of the phishing campaign, the system may employ techniques such as URL spoofing, where the fraudulent website's URL closely resembles that of the legitimate organization, or email spoofing, where the sender's address is manipulated to appear genuine. These tactics help to further deceive victims and increase the likelihood of successful exploitation. Once victims have been deceived into providing their information, the system must have mechanisms in place to harvest and exploit the stolen data. This may involve storing login credentials for later unauthorized access to accounts, selling personal information on the dark web for financial gain, or using the data for targeted phishing attacks or identity theft. Overall, a proposed system for phishing involves the orchestration of various deceptive elements to trick individuals into disclosing sensitive information. By exploiting trust, leveraging social engineering tactics, and employing technical deception, the system aims to achieve its objectives of data theft, fraud, or unauthorized access. It is important to note that phishing is illegal and unethical, and individuals should be vigilant to protect themselves against such attacks.

Advantages of Proposed System

- Precision Enhancement
- Adaptive Defense
- User-friendly Interface
- Efficiency

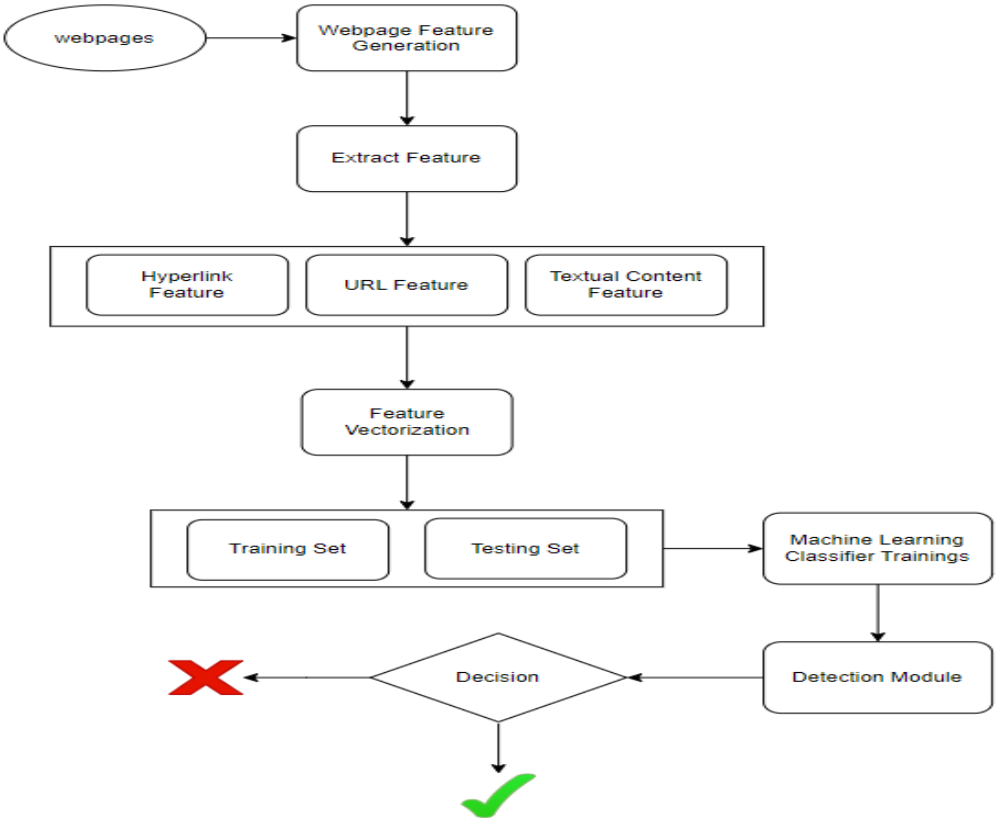
International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205102 |

Architectural Diagram



IMPLEMENTATION

Dataset Collection module

	A	B	C	D	E	F	G	H	I
1	phish_id	url	phish_detail	submission_	verified	verification_	online	target	
2	6557033	http://u104	http://www	2020-05-09	yes	2020-05-09	yes	Other	
3	6557032	http://hoys	http://www	2020-05-09	yes	2020-05-09	yes	Other	
4	6557011	http://www	http://www	2020-05-09	yes	2020-05-09	yes	Facebook	
5	6557010	http://www	http://www	2020-05-09	yes	2020-05-09	yes	Facebook	
6	6557009	https://firet	http://www	2020-05-09	yes	2020-05-09	yes	Microsoft	
7	6557007	http://kaize	http://www	2020-05-09	yes	2020-05-09	yes	Other	
8	6557008	http://kaize	http://www	2020-05-09	yes	2020-05-09	yes	Other	
9	6557006	http://kaize	http://www	2020-05-09	yes	2020-05-09	yes	Other	
10	6557005	http://kaize	http://www	2020-05-09	yes	2020-05-09	yes	Other	
11	6557004	https://kaiz	http://www	2020-05-09	yes	2020-05-09	yes	Other	
12	6557003	https://kaiz	http://www	2020-05-09	yes	2020-05-09	yes	Other	
13	6557002	https://kaiz	http://www	2020-05-09	yes	2020-05-09	yes	Other	
14	6557001	https://kaiz	http://www	2020-05-09	yes	2020-05-09	yes	Other	
15	6556982	http://santa	http://www	2020-05-09	yes	2020-05-09	yes	Banco Santander, S.A.	
16	6556981	http://sntbg	http://www	2020-05-09	yes	2020-05-09	yes	Banco Santander, S.A.	
17	6556973	http://chase	http://www	2020-05-09	yes	2020-05-09	yes	Other	
18	6556972	https://twe	http://www	2020-05-09	yes	2020-05-09	yes	Other	
19	6556969	https://bcpa	http://www	2020-05-09	yes	2020-05-09	yes	Other	
20	6556968	https://nhsr	http://www	2020-05-09	yes	2020-05-09	yes	Other	
21	6556948	http://betas	http://www	2020-05-09	yes	2020-05-09	yes	Other	
22	6556949	https://beta	http://www	2020-05-09	yes	2020-05-09	yes	Other	
23	6556930	http://zabo	http://www	2020-05-09	yes	2020-05-09	yes	Other	
24	6556929	http://zabo	http://www	2020-05-09	yes	2020-05-09	yes	Other	
25	6556927	http://chase	http://www	2020-05-09	yes	2020-05-09	yes	Other	
26	6556926	http://chase	http://www	2020-05-09	yes	2020-05-09	yes	Other	
27	6556925	http://chase	http://www	2020-05-09	yes	2020-05-09	yes	Other	
28	6556924	http://nhon	http://www	2020-05-09	yes	2020-05-09	yes	Other	
29	6556923	http://stagi	http://www	2020-05-09	yes	2020-05-09	yes	Other	

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

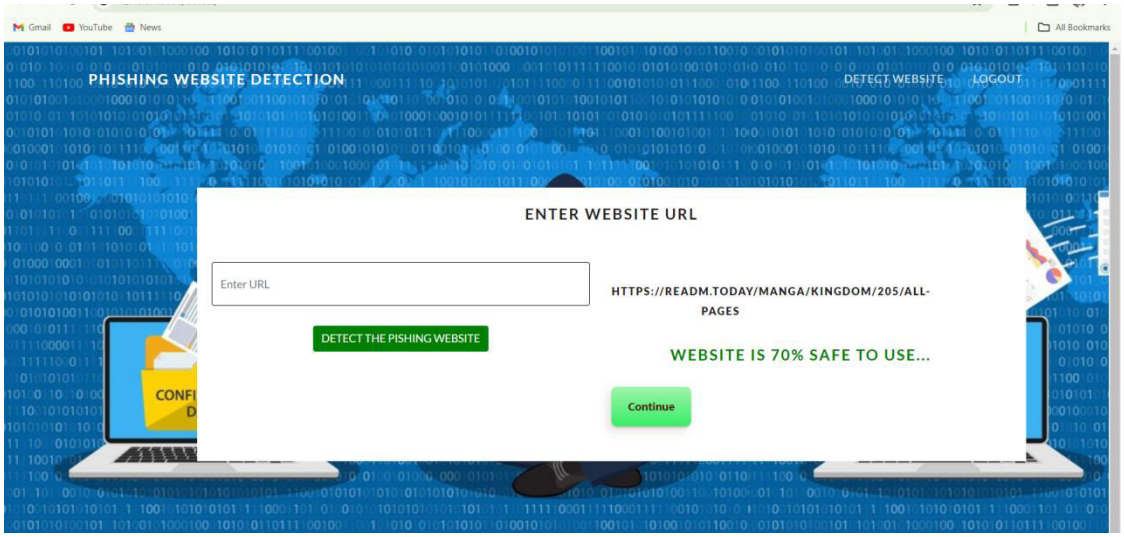
|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205102 |

Feature Extraction module

	Domain	Have_IP	Have_At	URL_Length	URL_Depth	Redirection	https_Domain	TinyURL	Prefix/Suffix	DNS_Record
0	graphicriver.net	0	0	1	1	0	0	0	0	0
1	ecnavi.jp	0	0	1	1	1	0	0	0	0
2	hubpages.com	0	0	1	1	0	0	0	0	0
3	extratorrent.cc	0	0	1	3	0	0	0	0	0
4	icicibank.com	0	0	1	3	0	0	0	0	0

User Interface module



VI. CONCLUSION

In conclusion, the Phishing website detection project encompasses a comprehensive approach to combating online threats and safeguarding users against malicious activities. By leveraging machine learning techniques, robust back-end infrastructure, and intuitive front-end interfaces, the system aims to provide effective detection and mitigation of phishing websites. Through thorough testing strategies, including unit testing, integration testing, and acceptance testing, the system's functionality, reliability, and security are rigorously validated. The project's emphasis on usability, scalability, and compliance ensures that it meets stakeholder expectations and delivers value to users. Moving forward, continued development and refinement of the system will be essential to adapt to evolving cybersecurity threats and emerging technologies. With a proactive and holistic approach to phishing detection, the project contributes to enhancing online safety and fostering trust in digital environments.

REFERENCES

1. 14 Types of Phishing Attacks That IT Administrators Should Watch For [online] (2021) f-phishing/.
2. Lakshmanarao, A., Rao, P.S.P., Krishna, M.M.B. (2021) 'Phishing website detection using novel machine learning fusion approach', in 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Presented at the 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 1164–1169.
3. H. Chapla, R. Kotak and M. Joiser, "A Machine Learning Approach for URL Based Web Phishing Using Fuzzy Logic as Classifier", 2019 International Conference on Communication and Electronics Systems (ICCES), pp. 383-388, 2019, July.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205102 |

4. Vaishnavi, D., Suwetha, S., Jinila, Y.B., Subhashini, R., Shyry, S.P. (2021) 'A Comparative Analysis of Machine Learning Algorithms on Malicious URL Prediction', in 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), Presented at the 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), 1398–1402.
5. Abu-Nimeh, S., Nappa, D., Wang, X., Nair, S. (2007), A comparison of machine learning techniques for phishing detection. Proceedings of the Anti-phishing Working Groups 2nd Annual ECrime Researchers Summit on - ECrime '07. doi:10.1145/1299015.1299021.
6. E., B., K., T. (2015)., Phishing URL Detection: A Machine Learning and Web Mining-based Approach. International Journal of Computer Applications,123(13), 46-50. doi:10.5120/ijca2015905665.
7. Wang Wei-Hong, L V Yin-Jun, CHEN Hui-Bing, FANG Zhao-Lin., A Static Malicious Javascript Detection Using SVM, In Proceedings of the 2nd International Conference on Computer Science and Electrical Engineering (ICCSEE 2013).

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details