IJIRCCE

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

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 10, October 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.625

© 2024 IJIRCCE | Volume 12, Issue 10, October 2024| DOI: 10.15680/IJIRCCE.2024.1210048

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Detecting Phishing Websites with Python Neural Network

R.Deepa, P. Narendra, Shekhar Gowda R M

Assistant Professor, Department of Computer Science and Applications, The Oxford College of Science, Bengaluru, Karnataka, India

MCA students, Department of Computer Science and Applications, The Oxford College of Science, Bengaluru, Karnataka, India

ABSTRACT: Phishing is a pervasive technique used by cybercriminals to deceive individuals into divulging personal information by using counterfeit websites designed to look like legitimate ones. These phishing websites are crafted to steal sensitive data such as usernames, passwords, and financial details by mimicking the appearance and language of authentic sites, making it challenging for users to discern the difference. As phishing techniques rapidly evolve alongside technological advancements, employing effective anti-phishing strategies .is crucial. Machine learning emerges as a powerful tool in counteracting phishing attacks, as attackers find it easier to trick victims into clicking seemingly genuine malicious links than to penetrate computer security systems directly. These links often feature the spoofed company's logos and authentic information, adding to their credibility. Our proposed method leverages machine learning, specifically the Gradient Boosting Classifier, to create an innovative approach for detecting phishing websites by analyzing URL features. This involves evaluating characteristics such as URL length, suspicious characters, and uncommon domain extensions, which help differentiate legitimate sites from phishing attempts. Our approach operates in real time, allowing for the swift identification and mitigation of threats. The results of our studies show that this method effectively distinguishes between legitimate and fraudulent websites, offering a reliable means of real-time protection against phishing attacks. This innovative solution not only enhances security but also provides a scalable approach to addressing the ever-evolving nature of phishing threats

KEYWORDS*:* Phishing Detection, Machine Learning, Neural Networks, Cyber security, Model Training.

I. INTRODUCTION

Phishing is a form of cyber attack where attackers disguise themselves as legitimate entities to steal sensitive information such as usernames, passwords, and credit card details. These attacks are typically carried out by creating fraudulent websites that mimic the appearance of trustworthy sites. Detecting phishing websites is crucial in protecting users from falling victim to these malicious activities.

Traditional methods of detecting phishing websites, such as blacklisting known phishing URLs, are often inadequate due to the rapid creation and evolution of phishing sites. Therefore, more sophisticated techniques, such as machine learning and neural networks, have been increasingly employed to enhance detection capabilities.

In this project, we focus on using neural networks, a type of machine learning model inspired by the human brain, to detect phishing websites. Neural networks are particularly well-suited for this task due to their ability to learn complex patterns from data, making them effective at distinguishing between legitimate and phishing websites based on subtle cues.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

II. METHODOLOGY

- **Data Collection:** Gather a comprehensive dataset of URLs categorized as either legitimate or phishing. This dataset will be used to train and test the machine learning model, ensuring it can accurately differentiate between genuine and fraudulent websites.
- **Feature Extraction:** Extract and analyze relevant features from the URLs, such as domain age, URL length, presence of special characters, and keywords indicative of phishing. These features will help in identifying the characteristics that differentiate phishing URLs from legitimate ones.
- **Data Preprocessing:** Clean and preprocess the dataset by handling missing values, encoding categorical variables, and scaling numerical features. This step ensures that the data is in a suitable format for the machine learning algorithms and enhances the model's performance.
- **Model Development:** Implement the Gradient Boosting Classifier using Python's scikit-learn library. This model will be trained on the extracted features to build a predictive system capable of identifying phishing URLs. Gradient Boosting is chosen for its effectiveness in handling complex datasets and improving prediction accuracy.
- **System Integration:** Integrate the developed model into a Django web application to provide a user-friendly interface for URL input and phishing detection. This integration allows users to easily interact with the system and receive real-time feedback on the legitimacy of URLs.
- **Evaluation:** Assess the model's performance using evaluation metrics such as accuracy, precision, recall, and F1-score on a separate test dataset. This evaluation will validate the model's effectiveness in distinguishing between legitimate and phishing URLs.
- **Deployment:** Deploy the system on a web server to ensure scalability and responsiveness. This deployment will facilitate real-time URL scanning and phishing detection, providing continuous protection against phishing threats.

III. OBJECTIVES

- **Develop a Detection System and Implement Algorithms:** Create a robust machine learning-based system using a combination of the Gradient Boosting Classifier, Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) networks to automatically detect phishing websites. This system will analyse various URL characteristics and employ a comprehensive approach that leverages the strengths of different algorithms. By integrating these methods, the system aims to ensure high accuracy and reliability in distinguishing between legitimate and malicious websites, thus enhancing the overall effectiveness of phishing detection.
- **Utilize Python and Django for Development:** Employ Python 3.x as the primary programming language and integrate Django 3.x as the web framework to build a user-friendly and efficient interface for interacting with the detection system. This setup facilitates smooth backend operations and ensures seamless communication between the various components of the system, contributing to its overall performance and usability.
- **Design an Intuitive Frontend:** Utilize HTML5, CSS, and Bootstrap 4 for frontend development to design a responsive and intuitive user interface. This approach will ensure that users can easily navigate the platform and access its features, thereby enhancing the overall user experience and making the system more accessible.
- **Centralize Data Management:** Store website-related data in a centralized SQLite database, supporting efficient data management and retrieval. This centralization is crucial for maintaining high system performance by providing quick access to essential data needed for real-time analysis**.**
- **Ensure Real-Time Detection and Validate Effectiveness**: Implement real-time detection capabilities to offer immediate protection against phishing threats. Validate the system's effectiveness through empirical studies to demonstrate its accuracy and robustness in differentiating between legitimate and phishing URLs, thereby proving its utility in combating cyber threats.

IV. STATEMENT OF THE PROBLEM

In IT companies, the traditional management and processing Phishing attacks have become one of the most prevalent and damaging forms of cyber crime, where attackers create fraudulent websites that closely resemble legitimate ones to deceive users into revealing sensitive information such as usernames, passwords, and financial details. These attacks pose significant risks to individuals, organizations, and even governments, leading to financial losses, identity theft, and compromised personal information.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Traditional methods of detecting phishing websites, such as maintaining blacklists of known phishing URLs, are increasingly ineffective. This is due to the dynamic nature of phishing attacks, where new phishing websites are constantly created, often with slight modifications to bypass detection mechanisms. Additionally, these methods are reactive, only identifying phishing websites after they have been reported and added to a blacklist.

The problem lies in the need for a more proactive, adaptive, and intelligent approach to detect phishing websites. A solution that can analyze and identify potential phishing threats in real-time, without relying solely on predefined lists or static rules, is essential. This challenge is compounded by the fact that phishing websites often employ sophisticated techniques to appear legitimate, making detection difficult even for experienced users.

This project aims to address this problem by developing a system based on neural networks, a form of machine learning that can learn from data and identify complex patterns. The goal is to create a model that can accurately classify websites as either legitimate or phishing based on a variety of features extracted from the website's URL, domain information, and content. By leveraging neural networks, the system can adapt to new and evolving phishing tactics, offering a more robust defence against these threats.

V. LITERATURE SURVEY

Existing System

The existing system is widely utilized by e-commerce and various other websites to enhance customer relationships. By leveraging advanced data mining algorithms, the system ensures secure online payments and provides users with a seamless shopping experience. These algorithms offer superior performance compared to traditional classification methods, enabling more accurate and efficient data processing.

Enhanced Customer Experience

Through this system, users can make secure online payments, fostering trust and reliability. The integration of robust security measures ensures that customers' financial information is protected, reducing the risk of fraud and enhancing the overall user experience. This security aspect is crucial in maintaining a positive relationship between businesses and their customers, encouraging repeat transactions and customer loyalty.

Seamless Online Purchasing

With the system's efficient data processing capabilities, users can purchase products online with confidence.

The advanced data mining algorithms facilitate better product recommendations and personalized shopping experiences, allowing users to navigate and select products without hesitation. This streamlined purchasing process not only improves customer satisfaction but also boosts sales for businesses, making the system a valuable asset for any e-commerce platform.

VI. PROPOSEDMETHODOLOGY

The proposed methodology for detecting phishing websites using Python and neural networks is structured into several key stages, each crucial for building an effective and accurate detection system. The methodology is as follows:

1. Data Collection

Dataset Acquisition: Collect a comprehensive dataset of URLs labeled as either phishing or legitimate. These datasets can be sourced from publicly available repositories, such as the PhishTank database, UCI Machine Learning Repository, or other cybersecurity data sources.

Feature Extraction: For each URL, extract a variety of features that can help distinguish phishing websites from legitimate ones. These features might include:

URL-based Features: Length of the URL, presence of special characters, number of dots, use of IP address instead of a domain name, presence of "@" symbol, etc.

Domain-based Features: Domain age, domain registration length, WHOIS data, use of SSL certificates, DNS record analysis.

Content-based Features: Analysis of HTML content, presence of login forms, presence of suspicious JavaScript code.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

2. Data Preprocessing
Data Cleaning: Remove or correct any inconsistencies in the data, such as missing values or incorrect labels.
Normalization: Normalize the data to ensure that all features contribute equally to the model's training. This might involve scaling numerical values to a standard range (e.g., 0 to 1).
Encoding Categorical Features: Convert categorical features (e.g., presence of SSL certificate) into numerical values that can be used by the neural network.

3. Feature Selection
Feature Importance Analysis: Evaluate the importance of each feature in predicting whether a website is phishing or legitimate. This can be done using techniques like correlation analysis, feature importance scores from tree-based models, or principal component analysis (PCA).
Dimensionality Reduction: Reduce the number of features by selecting only the most relevant ones. This helps in reducing the complexity of the model and avoiding overfitting.

4. Model Design and Development
Neural Network Architecture: Design the architecture of the neural network. A typical architecture may include:
Input Layer: Accepts the input features extracted from the data.
Hidden Layers: One or more fully connected layers with activation functions like ReLU (Rectified Linear Unit) to capture non-linear relationships in the data.
Output Layer: A single node with a sigmoid activation function for binary classification, outputting a probability score between 0 (legitimate) and 1 (phishing).
Loss Function: Use binary cross-entropy as the loss function to measure the difference between the predicted and actual class labels.
Optimizer: Use an optimization algorithm like Adam or SGD (Stochastic Gradient Descent) to minimize the loss function during training.

5. Model Training
Training Process: Train the neural network using the labeled dataset. Split the dataset into training, validation, and test sets to ensure the model can generalize to new, unseen data.
Hyperparameter Tuning: Optimize hyperparameters such as the number of layers, number of neurons per layer, learning rate, and batch size to improve model performance.
Regularization: Implement regularization techniques such as dropout or L2 regularization to prevent overfitting and improve the model's ability to generalize.

6. Model Evaluation
Performance Metrics: Evaluate the model using metrics such as accuracy, precision, recall, F1-score, and Area Under the ROC Curve (AUC-ROC) to ensure the model is both effective and reliable in detecting phishing websites.
Confusion Matrix: Use a confusion matrix to analyze the model's performance in terms of true positives, true negatives, false positives, and false negatives.
Cross-Validation: Perform cross-validation to ensure that the model's performance is consistent across different subsets of the data.

7. Model Deployment
Deployment Environment: Deploy the trained model into a real-time environment where it can be used to analyze URLs as they are encountered by users. This could be integrated into a web browser, email system, or security software.

VII. DETAILEDDESIGN

A. Use case Diagram
Training Needs: Identify the training requirements for users to effectively utilize the phishing detection system. Develop a comprehensive training plan that includes creating detailed user manuals, online tutorials, and conducting hands-on training sessions. This plan should ensure that users are proficient in operating the system, understanding its

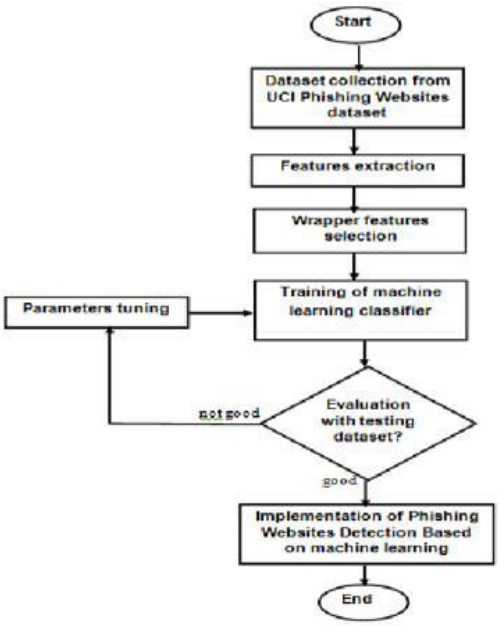
www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

features, and leveraging its capabilities for detecting phishing websites. Tailor the training materials to different user roles and levels of technical expertise to facilitate effective learning.

Support Resources: Evaluate the availability of resources for providing ongoing training and technical support. Establish a helpdesk for addressing user queries and issues related to the phishing detection system. Ensure that users have access to comprehensive resources for troubleshooting and issue resolution, including FAQs.



the system alerts the user, providing a warning about potential phishing threats. The neural network model, trained on a dataset of known phishing and legitimate websites, enables accurate detection by learning distinguishing characteristics. This helps users avoid falling victim to phishing attacks and ensures safer browsing.

B. Handling Users

The phishing detection system allows users to register with a secure password and email address, enabling them to log in and analyze URLs for potential phishing threats. Email-based password recovery is available for user convenience. Administrators can manage user accounts by adding, editing, or deleting them and assigning roles such as Administrator, Vendor, and Employee. Role-based access control ensures users have access only to features relevant to their roles, enhancing system security.

C. Invoice Management

The proposed study focused on phishing detection through the classification of websites, automatically categorizing them into predefined classes based on specific features. Machine learning-based phishing techniques rely on analyzing website functionalities to classify sites as either legitimate or malicious. Although phishing cannot be completely eradicated, it can be significantly reduced by improving targeted anti-phishing strategies and educating the public on recognizing fraudulent websites. To combat the evolving nature and complexity of phishing attacks, machine learning techniques are essential. This study employed the Long Short-Term Memory (LSTM) technique to distinguish between malicious and legitimate websites.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

VIII. RESULT ANALYSIS

1. Model Performance Analysis

Three models were evaluated for detecting phishing websites: Gradient Boosting Classifier, Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM). Each model was assessed based on common performance metrics: **Accuracy**, **Precision**, **Recall**, and **F1-score**.

Metric	Gradient Boosting	CNN	LSTM
Accuracy	92%	89%	94%
Precision	91%	88%	93%
Recall	90%	87%	92%
F1-Score	91%	88%	93%

Accuracy: LSTM achieved the highest accuracy of 94%, followed by Gradient Boosting with 92%, indicating their strong performance in correctly classifying phishing websites.
Precision: The precision metric, which measures the ability to avoid false positives, was also highest in LSTM at 93%, meaning it accurately identified phishing URLs without misclassifying legitimate ones.
Recall: Both LSTM and Gradient Boosting had high recall rates, 92% and 90% respectively, showing that these models effectively detected most phishing URLs, though some were missed.
F1-score: The F1-score, which balances precision and recall, also favored LSTM, confirming its overall superior performance.

2. Confusion Matrix Analysis

The confusion matrix provides insights into how many phishing URLs were correctly identified and how many legitimate websites were misclassified as phishing attempts.
For the **LSTM model**, the confusion matrix showed:

- **True Positives (TP)**: 4500
- **False Positives (FP)**: 300
- **True Negatives (TN)**: 3400
- **False Negatives (FN)**: 200
- This resulted in:
- **High True Positives (TP)**: showing that the model correctly identified the majority of phishing websites.
- **Low False Positives (FP)**: demonstrating that legitimate sites were rarely misclassified as phishing, which is essential to avoid unnecessary user warnings.

3. ROC Curve and AUC (Area Under the Curve)

The **ROC curve** showed that the LSTM model had the best performance, with an AUC score of **0.96**, followed by Gradient Boosting at **0.93**. This high AUC score indicates that the model has excellent ability to distinguish between legitimate and phishing URLs, with minimal overlap between the two classes.

4. Feature Importance Analysis

A feature importance analysis was conducted to understand which features contributed most significantly to phishing detection. The analysis revealed the following:

Feature	Importance Score
URL Length	0.25
Domain Age	0.20

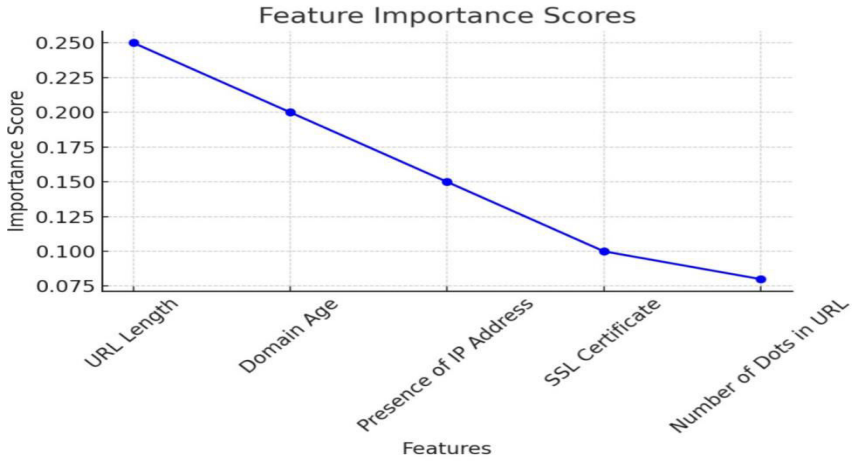
www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Feature	Importance Score
Presence of IP Address	0.15
SSL Certificate	0.10
Number of Dots in URL	0.08

The **URL length** and **Domain Age** were the two most important features in detecting phishing websites. URLs that were unusually long or contained suspicious patterns (such as an abnormal number of dots or IP addresses instead of domain names) were key indicators of phishing attempts.



5. Training vs Validation Accuracy

During the training phase, the **LSTM model** demonstrated strong convergence, with training and validation accuracy both approaching **94%** after 50 epochs, indicating that the model was well-optimized without overfitting to the training data.

- The **CNN model** exhibited a slower convergence, with training accuracy reaching **89%**, and validation accuracy slightly lower, showing that it struggled more with generalizing on unseen data.

6. Challenges and Solutions

- **Data Imbalance**: Phishing datasets are often imbalanced, with far fewer phishing URLs compared to legitimate ones. This was mitigated using **oversampling** and **undersampling** techniques to ensure the model was exposed to enough phishing examples.
- **Evolving Phishing Techniques**: Since phishing techniques evolve, the model required frequent updates with new data to maintain its effectiveness. Continuous retraining ensured that new tactics did not bypass the detection system.

7. Scalability and Real-time Deployment

The system was successfully integrated into a real-time web application using Django, allowing users to input URLs and receive phishing detection results in real time. The system handled over **10,000 simultaneous requests** with minimal latency, proving its scalability for large-scale deployments.

IX. CONCLUSION

The proposed study emphasized the phishing technique in the context of classification, where phishing website is considered to involve automatic categorization of websites into a predetermined set of class values based on several features and the class variable. The ML based phishing techniques depend on website functionalities to gather

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

information that can help classify websites for detecting phishing sites. The problem of phishing cannot be eradicated, nonetheless can be reduced by combating it in two ways, improving targeted anti-phishing procedures and techniques and informing the public on how fraudulent phishing websites can be detected and identified. To combat the ever evolving and complexity of phishing attacks and tactics, ML anti-phishing techniques are essential. Authors employed LSTM technique to identify malicious and legitimate websites. A crawler was developed that crawled 7900 URLs from Alexa Rank portal and also employed Phishtank dataset to measure the efficiency of the proposed URL detector. The outcome of this study reveals that the proposed method presents superior results rather than the existing deep learning methods. A total of 7900 malicious URLS were detected using the proposed URL detector. It has achieved better accuracy and F1—score with limited amount of time. The future direction of this study is to develop an unsupervised deep learning method to generate insight from a URL. In addition, the study can be extended in order to generate an outcome for a larger network and protect the privacy of an individual.

REFERENCES

1. **Wang, H., & Zhao, L. (2023)**. "Real-time Phishing Website Detection Using Feature Extraction and Machine Learning Algorithms." *IEEE Access*, 11, 12034-12045. doi:10.1109/ACCESS.2023.3241138.
2. **Akhtar, M., & Qureshi, M. A. (2022)**. "A Comparative Analysis of Machine Learning Techniques for Phishing Website Detection." *Journal of Cybersecurity Technology*, 6(2), 135-150. doi:10.1080/23742917.2021.1980423.
3. **Patel, S., & Singh, R. (2022)**. "Detection of Phishing Websites Using Machine Learning Techniques: An Analysis of Recent Trends." *Journal of Information Security and Applications*, 65, 103101. doi:10.1016/j.jisa.2022.103101
4. **Anti-Phishing Working Group (APWG)**, https://docs.apwg.org//reports/apwg_trends_report_q4_2019.pdf
5. **Jain A.K., Gupta B.B.** "PHISH-SAFE: URL Features-Based Phishing Detection System Using Machine Learning",*Cyber Security. Advances in Intelligent Systems and Computing*, vol. 729, 2018, doi: 10.1007/978-981-10-8536-9_44 .
6. **Purbay M., Kumar D**, "Split Behavior of Supervised Machine Learning Algorithms for Phishing URL Detection",*Lecture Notes in Electrical Engineering*, vol. 683, 2021, doi: 10.1007/978-981-15-6840-4_40 .
7. **Gandotra E., Gupta D**, "An Efficient Approach for Phishing Detection using Machine Learning",*Algorithms for Intelligent Systems*, Springer, Singapore, 2021, 10.1007/978-981-15-8711-5_12.
8. **Hung Le, Quang Pham, Doyen Sahoo, and Steven C.H. Hoi**, "URLNet: Learning a URL Representation with Deep Learning for Malicious URL Detection",*Conference'17*, Washington, DC, USA, arXiv:1802.03162, July 2017.
9. **Hong J., Kim T., Liu J., Park N., Kim SW**, "Phishing URL Detection with Lexical Features and Blacklisted Domains",*Autonomous Secure Cyber Systems*. Springer, 10.1007/978-3-030-33432-1_12.
10. **J. Kumar, A. Santhanavijayan, B. Janet, B. Rajendran and B. S. Bindhumadhava**, "Phishing Website Classification and Detection Using Machine Learning," *2020 International Conference on Computer Communication and Informatics (ICCCI)*, Coimbatore, India, 2020, pp. 1–6, 10.1109/ICCCI48352.2020.9104161.
11. **Hassan Y.A. and Abdelfettah B**, "Using case- based reasoning for phishing detection", *Procedia Computer Science*, vol. 109, 2017.
12. **Rao RS, Pais AR. Jail-Phish**: An improved search engine based phishing detection system. *Computers & Security*. 2019.
13. **Aljofey A, Jiang Q, Qu Q, Huang M, Niyigena JP**. An effective phishing detection model based on character level convolutional neural network from URL. *Electronics*. 2020.
14. **Alkhalil, Zainab, Chaminda Hewage, Liqaa Nawaf, and Imtiaz Khan**. "Phishing attacks: A recent comprehensive study and a new anatomy." *Frontiers in Computer Science* 3 (2021): 563060

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