IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134145

Detecting Phishing Websites Using Machine Learning

B. Sucharitha[1], B. Chandini[2], D. Satya Kumar[3], M. Surendra[4], Dr. G. Kishor Kumar[5]

Student, Computer Science and Engineering & Business Systems, RGMCET, Nandyala, India[1-4]

Professor and HOD of CSE & BS, RGMCET, Nandyala, India[5]

Abstract: Phishing attacks pose a significant threat to cybersecurity, necessitating effective detection mechanisms. This study explores the application of machine learning algorithms for the automated identification of phishing websites. By collecting a dataset of URLs labelled as phishing or legitimate, relevant features are extracted, pre-processed, and used to train various machine learning models. The performance of these models is evaluated using metrics such as accuracy, precision, recall, and F1-score, highlighting their effectiveness in distinguishing between phishing and legitimate URLs. Continuous monitoring and updates are emphasized to adapt to evolving phishing tactics. This research provides practical insights into the application of machine learning for phishing detection, contributing to the advancement of cybersecurity measures.

Keywords: Phishing detection, Machine learning, Cybersecurity, Feature extraction, Model evaluation

I. INTRODUCTION

Phishing attacks continue to be a pervasive threat in the digital landscape, targeting individuals, businesses, and organizations worldwide. These deceptive tactics involve the use of fraudulent emails, websites, or messages to trick users into divulging sensitive information such as passwords, credit card numbers, or personal details. Despite advancements in cybersecurity measures, phishing remains a lucrative and prevalent form of cybercrime, posing significant financial and reputational risks to victims. Traditional methods of detecting phishing attempts often rely on static blacklists, heuristics, or manual inspection, which are susceptible to evasion tactics employed by sophisticated attackers. As phishing techniques evolve and become increasingly sophisticated, there is a growing need for more robust and adaptive detection mechanisms.

Machine learning, with its ability to analyze vast amounts of data and identify patterns, presents a promising approach to address this challenge. The application of machine learning in phishing detection involves the use of algorithms to analyze features extracted from URLs and other indicators of suspicious activity. By leveraging labelled datasets comprising both phishing and legitimate URLs, machine learning models can learn to distinguish between the two and classify incoming web traffic accordingly. This automated approach offers the potential for real-time detection and mitigation of phishing attacks, enhancing overall cybersecurity posture.

One of the key advantages of using machine learning for phishing detection is its ability to adapt to evolving threats. Unlike static rule-based systems, machine learning models can continuously learn from new data and adjust their classification criteria accordingly. This adaptability is crucial in combating the dynamic nature of phishing attacks, where tactics and techniques are constantly changing to evade detection. However, the effectiveness of machine learning-based phishing detection systems relies heavily on the quality of the data and features used for training. Inadequate feature selection or biased datasets can lead to suboptimal performance and increased false positives or false negatives. Therefore, careful consideration must be given to the selection and preprocessing of features to ensure the robustness and generalization of the model.

In this study, we aim to explore the feasibility and effectiveness of using machine learning algorithms for phishing detection. We will investigate various feature extraction techniques, machine learning algorithms, and evaluation metrics to assess the performance of different models. By analyzing the strengths and limitations of these approaches, we seek to provide insights into the practical implementation of machine learning in combating phishing threats and advancing cybersecurity defences. Ultimately, the development of accurate and scalable machine learning-based phishing detection systems has the potential to significantly reduce the impact of phishing attacks, safeguarding individuals, businesses, and organizations against financial loss, data breaches, and reputational damage. Through ongoing research and innovation in this field, we can continue to stay ahead of cyber threats and build a more resilient digital ecosystem.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024
DOI: 10.17148/IJARCCE.2024.134145

II. LITERATURE SURVEY

R. Yetis and O. K. Sahingoz, "Blockchain Based Secure Communication for IoT Devices in Smart Cities,"
In smart city technologies we have witnessed advanced technological improvements in small computing devices, which can be connected to the Internet and named as Internet of Thing (IoT) devices, and cooperatively working complex systems. With this increased use of new technologies, the security problem is becoming more and more important because complex systems lead to unpredictable security vulnerabilities, which result in financial and private information losses. As a recently emerged technology, Blockchain was emerged as an alternative solution to security breaches of a different application environment. In contrast to the central structure used by most systems, it is preferred especially in the area of security by its distributed structure and the cryptographic hash algorithm it uses. Today, structures such as Smart Home, Smart City, Smart Environment and Smart Agriculture, which are created by using IoT are seen as active research areas with more security shortages. The reason for the security weakness in these areas arises from the hardware restriction on the IoT devices used. In the proposed system, an authorization system for IoT devices has been tried to be set up by using the distributed node structure of Blockchain system and blocks kept in these nodes. UDP (User Datagram Protocol), which uses a simple communication model without establishing a connection to the minimum protocol mechanism for communication of nodes in the system, was preferred. The communication between the nodes has been encrypted using encryption methods, thus creating a secure environment.

Awasthi and N. Goel, "Generating Rules to Detect Phishing Websites Using URL Features,"
Phishing, a well-known type of the cybercrime, is a fraudulent activity created and executed by cyber criminals all over the cyberspace worldwide. Several techniques are used by these cyber criminals or scammers to carry out these attacks on the user to deceive a user. There are also some dangerous types of phishing attacks such as: by using Email messages and the Multimedia Internet Mail Extensions (MIME) attachments with the email messages that are delivered to the targeted users that include newbies as well as old and experienced users. Other types include botnet-based attack that is performed using malwares sent either through emails or somehow forcing users to click on the links planted wherein these malicious messages are attached by the scammers in the body of email text or their webpages. Spear attacks are also dangerous kind of phishing attacks. In this kind of attack an individual or a company is targeted by these people to steal the sensitive data over web. Keeping in mind this very purpose, emails and malicious codes are used as their primary weapons. In present paper, Apriori algorithm is used to generate the rules so that these rules can play a vital role in detecting & predicting the phishing and non-phishing website URLs. In order to give the web users a basic idea against this threat, an effective tool is used to spread the word among the users, so that users may get the required knowledge about these and developing a tool to make users able to identify the phishing websites/content while working online

M. Korkmaz, O. K. Sahingoz and B. Diri, "Detection of Phishing Websites by Using Machine Learning-Based URL Analysis,"
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 analyze 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.

III. PROPOSED SYSTEM

The proposed system is designed as a user-friendly website aimed at detecting phishing websites with high accuracy. Built using HTML, CSS, JavaScript, and Flask framework in Python, the website offers an interactive and responsive platform for users. With a focus on simplicity and ease of use, the website ensures that all users can navigate it effortlessly. The system is trained on a dataset comprising various features crucial for determining the legitimacy of a website. Notably, the dataset does not include actual website URLs but encompasses essential attributes indicative of phishing behaviour. Leveraging the Gradient Boosting Classifier, the system analyzes these features to classify URLs as either legitimate or phishing. Upon analysis, if a website is identified as phishing, the system alerts the user accordingly, thereby enhancing online security.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134145

The proposed system boasts several advantages, including a user-friendly interface that simplifies the detection process. By incorporating a diverse range of features into the training data, the model achieves a high level of accuracy in distinguishing between legitimate and phishing websites.

Furthermore, the system demonstrates superior performance compared to other models, especially when handling large datasets. With support for categorical features and native handling of missing values, the proposed system offers robust capabilities for detecting phishing threats effectively.

IV. METHODOLOGY

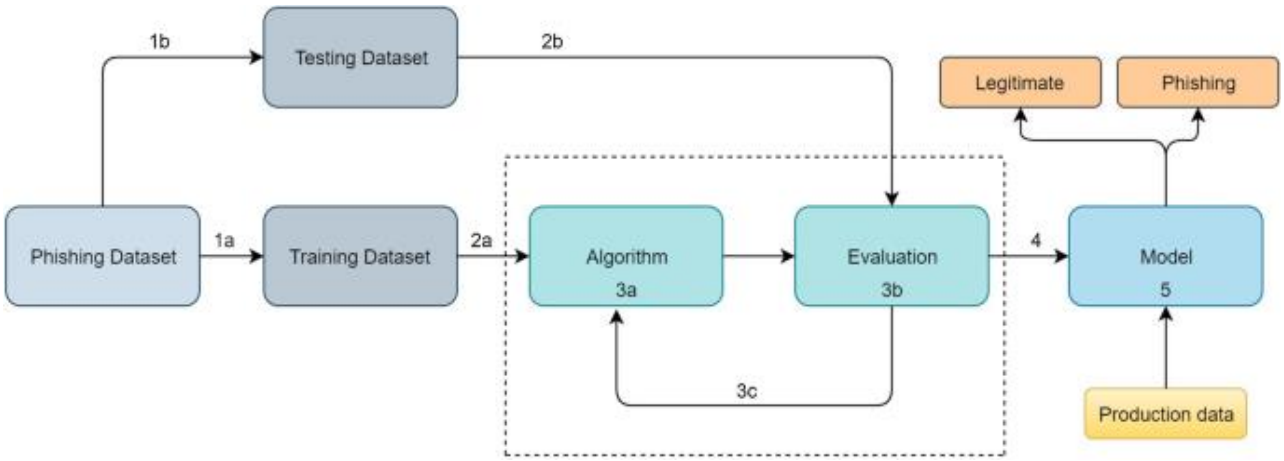

Fig. 1. Architecture for Proposed System

Phishing Dataset: This represents the raw data used to train and evaluate the machine learning model. It consists of features extracted from websites.

Data Splitting: The dataset is divided into two main parts:

Training Dataset (1a & 2a): This larger portion of the data is used to train the machine learning model. The model learns to identify patterns within the data that differentiate phishing websites from legitimate ones.

Evaluation Dataset (1b & 2b): This is used to assess the performance of the trained model on unseen data. It helps to identify how well the model generalizes to real-world scenarios and avoids overfitting to the training data.

Training the Model (2a): The training dataset is fed into a machine learning algorithm. The algorithm learns from the data and builds a model that can classify new websites as phishing or legitimate.

Evaluation (3a, 3b & 3c): The evaluation dataset is used to test the trained model. The model's predictions are compared to the actual labels of the websites in the evaluation dataset to calculate metrics like accuracy, precision, and recall.

Production Model (5): If the model performs well on the evaluation dataset, it can be deployed into production. This means the model can now be used to classify real-time website traffic and flag potential phishing attempts.

A. Dataset:

The dataset utilized in this project comprises 11,055 URLs collected from Kaggle, encompassing a total of 32 features. Each URL in the dataset is labeled as either legitimate (coded as 1) or phishing (coded as -1). To ensure data quality, any instances with missing values were removed, resulting in a clean dataset ready for analysis. Subsequently, the dataset was split into training and testing sets in a 80:20 ratio, facilitating the training and evaluation of the models.

B. Classifiers:

Delving into the classifiers employed in this project, three prominent algorithms were selected: Decision Tree, Random Forest, and Gradient Boosting.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024
DOI: 10.17148/IJARCCE.2024.134145

Beginning with Decision Tree, this algorithm constructs a hierarchical tree structure where each internal node represents a feature, and each leaf node represents a class label. By recursively partitioning the feature space based on the values of different attributes, Decision Tree creates a set of decision rules that collectively enable classification. However, Decision Trees are susceptible to overfitting, particularly in complex datasets with high variance.

To address the limitations of Decision Trees, Random Forest was employed as an ensemble learning technique. Random Forest builds multiple decision trees using bootstrapped samples of the dataset and randomly selected subsets of features. By aggregating the predictions of these individual trees, Random Forest mitigates the overfitting issue while improving predictive accuracy. Moreover, Random Forest's ability to handle large datasets efficiently makes it a popular choice for classification tasks in various domains.

Lastly, Gradient Boosting was utilized as another ensemble learning method renowned for its exceptional predictive performance. Unlike Random Forest, which constructs trees independently, Gradient Boosting builds a sequence of decision trees iteratively, with each subsequent tree focusing on correcting the errors made by its predecessors. By fitting new trees to the residual errors of the previous ones, Gradient Boosting gradually improves the model's predictive power, making it particularly effective in capturing complex patterns and relationships in the data.

C. Performance Metrics:
Through rigorous evaluation and comparison, these algorithms were assessed based on their performance metrics such as accuracy, precision, recall, and F1 score. By selecting the most suitable model based on these criteria, we aim to deploy a robust phishing detection system capable of accurately identifying and classifying URLs as legitimate or malicious, thereby enhancing cybersecurity measures and safeguarding users against online threats.

1. Accuracy:
Accuracy measures the proportion of correctly predicted instances to the total instances. It offers a general overview of the model's performance but may not be suitable for imbalanced datasets.

- **Formula:**

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

2. Precision:
Precision focuses on the proportion of correctly predicted positive cases out of all predicted positives. It is useful when the cost of false positives is high, and minimizing false positives is crucial.

- **Formula:**

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

3. Recall:
Recall measures the proportion of actual positives that were correctly identified by the model. It is beneficial when the cost of false negatives is high, and minimizing false negatives is essential.

- **Formula:**

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

4. F1-Score:
The F1-score is the harmonic mean of precision and recall, providing a balance between the two metrics. It is particularly useful when there is an uneven class distribution or when false positives and false negatives have different costs.

- **Formula:**

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134145

V. RESULTS AND DISCUSSION

TABLE-I Comparison of Performance Metrics

Model	Accuracy	Precision	Recall	F1- Score
Decision Tree	0.960	0.967	0.961	0.964
Random Forest	0.969	0.967	0.977	0.972
Gradient Boosting	0.989	0.990	0.994	0.986

Based on the results obtained from our model evaluation, it's evident that all three models, Decision Tree, Random Forest, and Gradient Boosting, exhibit high accuracy rates. However, when considering other performance metrics such as precision, recall, and F1-Score, Gradient Boosting outperforms the other models consistently across all measures.

Precision indicates the proportion of correctly predicted positive cases out of all predicted positive cases, while recall represents the proportion of correctly predicted positive cases out of all actual positive cases. F1-Score is the harmonic mean of precision and recall, providing a balanced measure of the model's performance.

In our case, Gradient Boosting demonstrates superior precision, recall, and F1-Score compared to Decision Tree and Random Forest. This suggests that Gradient Boosting not only correctly identifies phishing websites with high precision but also effectively captures a significant portion of actual phishing cases.

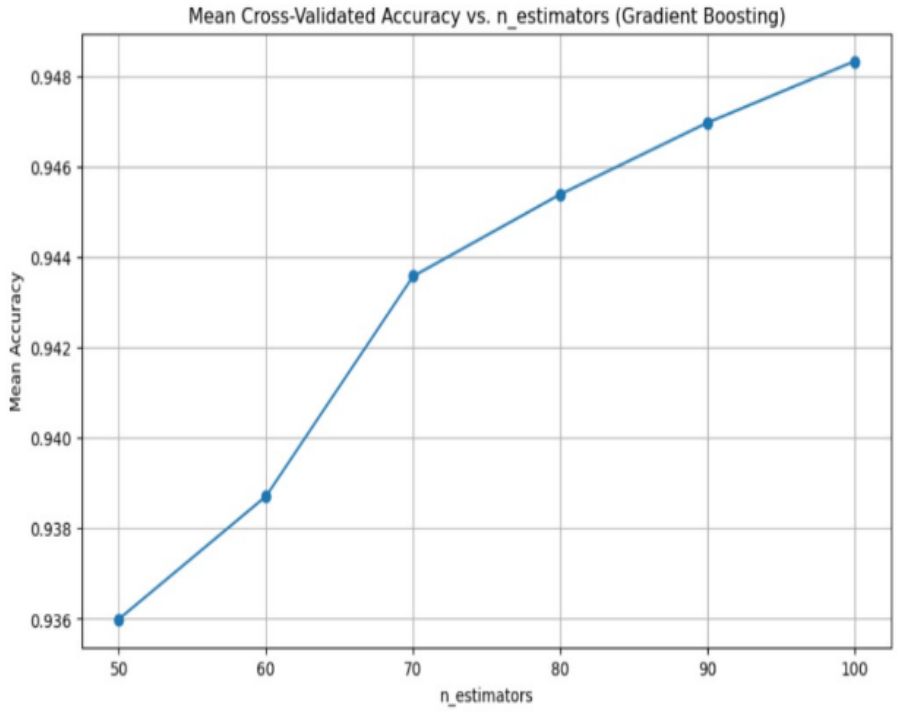

Fig. 2. Accuracy for n_estimators

In Gradient Boosting, each new tree is trained to correct the errors made by the previous trees. As a result, each subsequent tree focuses more on the mistakes of the previous ones.

Therefore, based on these results, we have made the decision to proceed with Gradient Boosting as our final model for detecting phishing websites. Its exceptional performance across multiple evaluation metrics makes it the most suitable choice for our project, ensuring robust and reliable detection of phishing attempts.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024
DOI: 10.17148/IJARCCE.2024.134145

Deploying the model onto a web server with a user-friendly interface crafted using HTML, CSS, and JavaScript, integrated with the Flask framework, marks a significant milestone in our project. This deployment enables users to conveniently input the URL of any website into the interface.

Upon submission, the model processes the input and swiftly provides the results, indicating whether the given URL is deemed safe or unsafe. This seamless interaction between the user and the model via the web interface enhances accessibility and usability, empowering users to make informed decisions about website safety in real-time.

VI. CONCLUSION

In this project, we employed machine learning techniques to identify phishing websites and reduce their success rates. For training and testing, we chose an 11,055-row dataset with 32 features from Kaggle. Our dataset was trained using three machine learning algorithms.

Gradient Boosting, Decision Tree, and Random Forest Tree were the algorithms employed. Gradient Boosting demonstrated the highest test accuracy (0.989), precision (0.990), recall (0.994), and F1-score (0.986) among the employed methods.

Gradient Boosting is the most effective technique for identifying phishing websites, according to our findings. Our techniques for identifying phishing websites are so accurate that they set our project apart from others.

Future Scope:
Looking forward, it's clear that just looking at the words in a web address might not be enough. Cybercriminals are good at changing web addresses to trick security systems. So, a better way is to also look at other things like where the website is hosted.

In our plans for the future, we want to make our phishing detection system even better. We're thinking about turning it into a service that lots of people can use on the internet. This upgraded system could learn from new phishing tricks as they come up, so it's always getting smarter at spotting them. We're also considering making a special tool that you can add to your web browser. This tool would be really good at spotting dangerous websites that try to trick you. If we make it happen, this tool would fit right into your web browser and help keep you safe while you're surfing the web.

REFERENCES

[1] Leon Reznik," Computer Security with Artificial Intelligence, Machine Learning, and Data Science Combination," in Intelligent Security Systems: How Artificial Intelligence, Machine Learning and Data Science Work For and Against Computer Security, IEEE, 2022, pp.1- 56, doi: 10.1002/9781119771579.ch1.
[2] O. K. Sahingoz, U. Cekmez and A. Buldu, "Internet of Things (IoTs) Security: Intrusion Detection using Deep Learning" 2021, Journal of Web Engineering, 2021, pp. 1721–1760, vol. 20, iss. 6, doi: 10.13052/jwe1540-9589.2062.
[3] R. Yetis and O. K. Sahingoz, "Blockchain Based Secure Communication for IoT Devices in Smart Cities," 2019 7th International Istanbul Smart Grids and Cities Congress and Fair (ICSG), 2019, pp. 134-138, doi: 10.1109/SGCF.2019.8782285.
[4] A. Awasthi and N. Goel, "Generating Rules to Detect Phishing Websites Using URL Features," 2021 1st Odisha International Conference on Electrical Power Engineering, Communication and Computing Technology(ODICON), 2021, pp. 1-9, doi: 10.1109/ODICON50556.2021.9429003.
[5] M. Korkmaz, O. K. Sahingoz and B. Diri, "Detection of Phishing Websites by Using Machine Learning-Based URL Analysis," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020, pp. 1-7, doi: 10.1109/ICCCNT49239.2020.9225561.
[6] L. Tang and Q. H. Mahmoud, "A Deep Learning-Based Framework for Phishing Website Detection," in IEEE Access, vol. 10, pp. 1509-1521, 2022, doi: 10.1109/ACCESS.2021.3137636.
[7] M. Korkmaz, O. K. Sahingoz and B. Diri, "Feature Selections for the Classification of Webpages to Detect Phishing Attacks: A Survey," 2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), 2020, pp. 1-9, doi: 10.1109/HORA49412.2020.9152934.
[8] E. Kocyigit, M. Korkmaz, O.K. Sahingoz, B. Diri, "Real-Time ContentBased Cyber Threat Detection with Machine Learning". In: Abraham, A., Piuri, V., Gandhi, N., Siarry, P., Kaklauskas, A., Madureira, A. (eds) Intelligent Systems Design and Application, 2021, ISDA 2020. Advances in Intelligent Systems and Computing, vol 1351. Springer, Cham. https://doi.org/10.1007/978-3-030-71187-0129.

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134145

[9] U. Ozker and O. K. Sahingoz, "Content Based Phishing Detection with Machine Learning," 2020 International Conference on Electrical Engineering (ICEE), 2020, pp. 1-6, doi: 10.1109/ICEE49691.2020.9249892.

[10] Y. A. Alsariera, V. E. Adeyemo, A. O. Balogun and A. K. Alazzawi, "AI Meta-Learners and Extra-Trees Algorithm for the Detection of Phishing Websites," in IEEE Access, vol. 8, pp. 142532-142542, 2020, doi: 10.1109/ACCESS.2020.3013699.

[11] L. Tang and Q. H. Mahmoud, "A Deep Learning-Based Framework for Phishing Website Detection," in IEEE Access, vol. 10, pp. 1509-1521, 2022, doi: 10.1109/ACCESS.2021.3137636.

[12] P. Yang, G. Zhao and P. Zeng, "Phishing Website Detection Based on Multidimensional Features Driven by Deep Learning," in IEEE Access, vol. 7, pp. 15196-15209, 2019, doi: 10.1109/ACCESS.2019.2892066.

[13] W. Ali and S. Malebary, "Particle Swarm Optimization-Based Feature Weighting for Improving Intelligent Phishing Website Detection," in IEEE Access, vol. 8, pp. 116766-116780, 2020, doi: 10.1109/ACCESS.2020.3003569.

[14] C. Pham, L. A. T. Nguyen, N. H. Tran, E. -N. Huh and C. S. Hong, "Phishing-Aware: A Neuro-Fuzzy Approach for Anti-Phishing on Fog Networks," in IEEE Transactions on Network and Service Management, vol. 15, no. 3, pp. 1076-1089, Sept. 2018, doi: 10.1109/TNSM.2018.2831197.

[15] Abdullateef O. et al., "Improving the phishing website detection using empirical analysis of Function Tree and its variants", Heliyon, vol 7, Issue 7, 2021, e07437,

[16] Dong-Jie Liu, Guang-Gang Geng, Xiao-Bo Jin, Wei Wang, An efficient multistage phishing website detection model based on the CASE feature framework: Aiming at the real web environment, Computers Security, vol 110, 2021, 102421, ISSN 0167-4048,

[17] Xi Xiao, Wentao Xiao, Dianyan Zhang, Bin Zhang, Guangwu Hu, Qing Li, Shutao Xia, Phishing websites detection via CNN and multihead self-attention on imbalanced datasets, Computers Security, Vol 108,2021,102372,ISSN 0167-4048.

[18] A.V. Ramana, Rao, K.L. Rao, R.S. Stop-Phish: an intelligent phishing detection method using feature selection ensemble. Soc. Netw. Anal. Min. 11, 110 (2021).

[19] SatheeshKumar, M., Srinivasagan, K.G. UnniKrishnan, G. A lightweight and proactive rule-based incremental construction approach to detect phishing scam. Inf Technol Manag (2022).