Uncovering the Efficiency of Phishing Detection: An In-depth Comparative Examination of Classification Algorithms

Dwi Sugianto[1,*], Rilliandi Arindra Putawa[2], Calvina Izumi[3,] Soeltan Abdul Ghaffar[4]

[1] *Magister of Computer Science, Computer Science Faculty, Universitas Amikom Purwokerto, Indonesia*

[2]*Departement of Islamic Studies, Universitas Negeri Padang, Padang, Indonesia*

[3]*School of Management Business, Ciputra University, Surabaya, Indonesia*

[4]*Department of Marine Information Systems, Universitas Pendidikan Indonesia, Bandung Indonesia*

(Received: January 25, 2024; Revised: February 16, 2024; Accepted: March 24, 2024; Available online: April 30, 2024)

Abstract

This research aims to investigate the potential security risks associated with phishing email attacks and compare the performance of three main classification algorithms: random forest, SVM, and a combination of k-fold cross-validation with the xgboost model. The dataset consists of 18,634 emails, with 7,312 identified as phishing emails and 11,322 considered safe. Through experiments, the combination of k-fold cross-validation and xgboost demonstrated the best performance with the highest accuracy of 0.9712828770799785. The email classification graph provides a visual insight into the distribution of classification results, aiding in understanding patterns and trends in phishing attack detection. The analysis of the ROC curve results indicates that k-fold cross-validation and xgboost have a higher AUC compared to random forest and SVM, signifying a better ability to predict the correct class. The conclusion emphasizes the importance of the combination of k-fold cross-validation and xgboost in enhancing email security, with the potential for increased accuracy through parameter adjustments.

Keywords: Phishing email attacks, Classification algorithms, XGBoost model, Email security, Cross-validation

1. Introduction

In the evolving digital era, the threats to email security are becoming increasingly complex, especially with the emergence of phishing attacks that often deceive users [1], [2]. This research is motivated by the need to enhance the effectiveness of security strategies in combating phishing email attacks. Despite the utilization of numerous classification algorithms for phishing detection, their performance comparison in the context of email security still requires deeper understanding. Therefore, this research aims to investigate and compare three main classification algorithms: random forest, Support Vector Machine (SVM), and a combination of k-fold cross-validation with the xgboost model [3], [4], [5], [6].

The primary challenge faced is the rising complexity of phishing attacks and the need for a more precise approach for effective detection [7]. By identifying and understanding the performance differences among various classification algorithms, it is expected to provide clearer guidance for improving email security strategies. The research objective is to evaluate the capabilities of these three algorithms in detecting potential security threats in phishing email attacks. The research question is to what extent random forest, SVM, and the combination of k-fold cross-validation with xgboost can yield optimal results in the context of email security.

This research holds relevance and significance in addressing the continually evolving cybersecurity challenges by contributing to our understanding of the performance of these algorithms. While there has been significant prior research focused on phishing attack detection, this study adds value by involving a detailed comparison of algorithms. However, it is acknowledged that this research has limitations, including constraints on the types of datasets used and other aspects that may affect the generalization of results.

***Corresponding author: Dwi Sugianto (dwisugianto@outlook.com)**
DOI: https://doi.org/10.47738/ijaim.v4i1.72

The research methodology will involve the use of a dataset comprising 18,634 emails, with a focus on 7,312 phishing emails and 11,322 benign emails. The experimental process will include the implementation of random forest, SVM, and the combination of k-fold with xgboost, with in-depth analysis of the classification results and accuracy of each algorithm.

2. Literature Review

Phishing is one of the primary security threats in the online world, with email phishing being the most commonly employed method [9][10]. Early detection of phishing attacks is crucial to safeguard users and organizations from potential losses [11]. In an effort to enhance detection effectiveness, this research focuses on the application of machine learning algorithms such as Random Forest, SVM, and k-fold Cross-validated eXtreme Gradient Boosting (XGBoost).

Previous studies have introduced various techniques in phishing email detection efforts [12], [13], [14]. However, standout solutions involve the application of machine learning algorithms, which show remarkable capabilities in learning patterns from data and identifying complex phishing characteristics. In this domain, algorithms like Random Forest, SVM, and k-fold XGBoost have emerged as primary research focuses, demonstrating significant potential in detecting and preventing phishing attacks. The strengths of these algorithms lie in their ability to rapidly and effectively process complex information, providing reliable solutions against cyber security threats. By continuously developing and integrating these technologies, higher levels of security against phishing attacks in the future can be anticipated.

Random Forest, a well-known ensemble method, has proven successful in classifying complex data. Embracing the concept of combining multiple decision trees, Random Forest effectively addresses the issue of overfitting common in machine learning models. The main advantage of Random Forest is its ability to enhance detection accuracy by utilizing decision combinations from various distinct trees [3].

Several studies have highlighted the effectiveness of Random Forest in the context of phishing email detection. Implementation of this technique has yielded satisfying results, demonstrating that Random Forest can identify complex and unstructured patterns often associated with phishing attempts. By harnessing the power of multiple decision trees, Random Forest becomes a robust and reliable solution in tackling dynamic and diverse data classification challenges, as seen in the case of phishing email detection.

SVM has become a prime choice in handling complex classification issues due to its ability to build models by seeking an optimal hyperplane to separate two classes [4]. The advantages of SVM are evident in various studies, particularly in distinguishing between phishing and non-phishing emails. This approach has shown success in positively contributing to early detection of potential security threats. With SVM's ability to handle nonlinear and complex data, this method not only reinforces security but also opens opportunities for broader applications across various domains, making it an effective and reliable tool in addressing challenging classification issues.

XGBoost, as a sophisticated boosting algorithm, has gained significant popularity in handling classification cases [6]. One of its main advantages is its ability to improve model generalization. By implementing the k-fold Cross-validation approach on XGBoost, the model can be thoroughly tested using different data subsets, enhancing its ability to recognize more general patterns and not just specific patterns from a particular dataset [5].

Previous research has confirmed that applying k-fold Cross-validation to XGBoost provides high detection performance, especially in phishing detection scenarios. This indicates that the generated model can accurately identify patterns and characteristics associated with phishing practices, instilling greater confidence in the use of XGBoost as an effective tool in addressing classification challenges, particularly in the context of cybersecurity.

In this research, an intriguing approach is taken by integrating three main algorithms, namely Random Forest, SVM, and k-fold XGBoost. The integration of these three algorithms aims to improve accuracy and robustness in detecting phishing attacks. By combining the strengths of each algorithm, it is anticipated that a phishing detection system can be created that is not only more effective but also more resilient to the varied tactics used by phishing attackers. This approach not only creates a strong combination in terms of detection but also provides better adaptability to evolving attack strategies.

 ISSN 2776-8007

3. Method

Figure 1 illustrates research step from collecting data from Kaggle. Feature extraction and selection are then performed to prepare the data for use. Classification algorithms such as Random Forest, SVM, and XGBoost are then applied to the processed data. To ensure the model works well, the data is divided into two parts: a training dataset and a testing dataset. Finally, the model's performance is evaluated using metrics such as accuracy, precision, and recall.

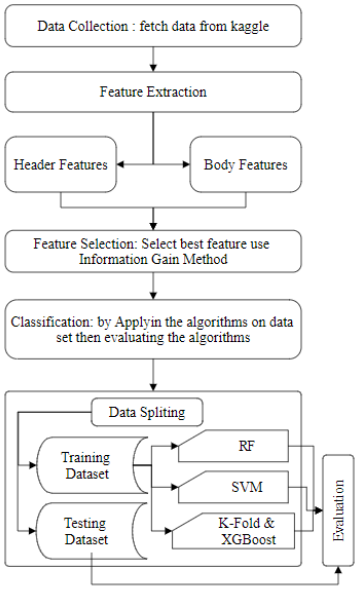

Figure 1. Research Step

3.1. Dataset

The dataset under investigation was obtained through the Kaggle platform, consisting of 18,634 emails categorized into 7,312 phishing emails and 11,322 identified as safe. These emails were not sourced from a single origin but were acquired through various channels, including phishing websites, forums, and mailing lists. With the inclusion of these diverse sources, the dataset provides a broad and heterogeneous representation of various types of phishing threats that may exist. This approach enables the research to establish a robust foundation in addressing the rapidly changing variations and dynamics in phishing practices.

3.2. Feature Engineering

In this study, two categories of features were employed to detect phishing emails: header features and body features. Header features encompass information from the email header, such as sender and recipient addresses, subject, and delivery date. On the other hand, body features involve characteristics within the email content, such as the use of keywords, language structure, and sentence structure. A total of 32 features were selected based on a review of the literature and experimental results. The feature selection process was conducted using the Information Gain method, aiming to choose the most important features in the classification context [15]. This method assists in filtering out features that significantly contribute to detecting phishing emails, thereby focusing the analysis on aspects with a greater impact on information security handling.

3.3. Classification

In this research space, three powerful classification algorithms were implemented for detecting potential phishing threats in emails: Random Forest Classifier, SVM, and k-fold Cross-Validation method with the application of XGBoost. The Random Forest Classifier, an ensemble model, leverages multiple decision trees to optimize prediction performance, where predictions can be explained by the formula:

$$\hat{y} = mode(h_1(x), h_2(x), \cdots, h_n(x)) \qquad (1)$$

With ŷ as the class prediction, $h_i(x)$ as the prediction from the i -th decision tree, and $mode(\cdot)$ as the mode function that selects the class with the highest frequency. On the other hand, SVM seeks the best hyperplane to separate classes in the feature space. The decision function of SVM is determined by the weight vector (w) and bias (b), and can be described by the formula:

$$f(x) = sign(w \cdot x + b) \quad (2)$$

XGBoost, as an ensemble algorithm employing boosting techniques, generates predictions by aggregating the outputs of a number of decision trees. The prediction function of XGBoost is delineated as follows:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \quad (3)$$

With $\hat{y}_i$ as the prediction for the i-th sample, K as the number of decision trees, and $f_k(x_i)$ as the output of the k-th decision tree. The significance of the k-fold Cross-Validation method is also acknowledged in this research. This method is employed to validate the model's performance by dividing the dataset into k subsets, where the overall model accuracy is calculated using the formula:

$$Accuracy = \frac{1}{k}\sum_{i=1}^{k} Accuracy_i \quad (4)$$

The application of the k-fold Cross-Validation method in conjunction with the XGBoost algorithm is expected to enhance the generalization of the model, reduce the risk of overfitting, and provide a more consistent evaluation of the classification model's performance. By combining the strengths of these three algorithms and validation methods, this research aims to make a significant contribution to the effectiveness of detecting potential phishing threats at the email level.

4. Result and Discussion

4.1. Exploratory Data Analysis

In order to uncover the potential security risks associated with phishing email attacks [16], [17], this experiment details the usage of three main classification algorithms, namely random forest, SVM, and the k-fold method combined with the xgboost model. The algorithm selection was conducted meticulously to encompass various security aspects that may be related to phishing emails.

In this evaluation, the dataset comprised 18,634 emails, with 7,312 of them identified as phishing emails, while the remaining 11,322 were considered safe. The results of the in-depth analysis provide a clear overview of the capabilities and reliability of the three algorithms in detecting and classifying potential security threats that may arise in phishing email attacks. Through this approach, the experiment offers valuable insights to enhance the effectiveness of security strategies against phishing attacks by gaining a deeper understanding of the performance of the employed algorithms.

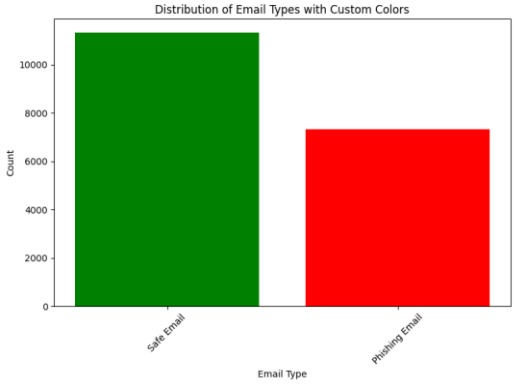

Figure 2. Distribution of Email Types

To provide a clearer visual representation, a classification graph of emails in the dataset is presented. This graph reflects the distribution of classification results, including the categorization between phishing emails and emails deemed safe. The presentation of this classification graph serves as a useful visual foundation for understanding patterns and trends

in phishing email detection, as well as offering a more comprehensive overview of the effectiveness of each algorithm in the context of email security.

4.2. Accuracy

Table 1 below illustrates the accuracy rates of the three mentioned algorithms:

Table 1. Algorithm Accuracy

Algorithm	Accuracy
Random Forest	0.9314038286235187
SVM	0.4990884229717411
K-fold & Xgboost	0.9712828770799785

The analysis of the table indicates that the k-fold method and the xgboost algorithm stand out by achieving the highest accuracy, reaching a value of 0.9712828770799785. This superiority is evident when compared to the performance of other methods, such as random forest with an accuracy of 0.9314038286235187 and SVM showing a low accuracy of 0.4990884229717411. Based on this comparison, it can be concluded that the combination of k-fold and xgboost offers superior performance in the context of this data analysis, demonstrating excellent capability in generating accurate predictions.

4.3. Parameters

Default parameters were utilized for all three algorithms. Nevertheless, experiments were conducted with various parameter variations. The results of these experiments are presented in Table 2, illustrating the accuracy of the algorithms with parameter modifications.

Table 2. Algorithm Accuracy with Parameter Modification

Algorithm	Parameter	Accuracy
Random Forest	n_estimators = 100, max_depth = 10	0.9314038286235187
Random Forest	n_estimators = 1000, max_depth = 10	0.9340284360189573
Random Forest	n_estimators = 1000, max_depth = 20	0.9381720430107527
SVM	C = 1, kernel = 'rbf'	0.4990884229717411
SVM	C = 10, kernel = 'rbf'	0.515311004784689
SVM	C = 100, kernel = 'rbf'	0.5232558139534884
K-fold & Xgboost	n_estimators = 100, learning_rate = 0.01	0.9673202614379085
K-fold & Xgboost	n_estimators = 1000, learning_rate = 0.01	0.9712828770799785
K-fold & Xgboost	n_estimators = 1000, learning_rate = 0.001	0.9704594180709558

From the conducted experiments, it can be concluded that the k-fold cross-validation method combined with the xgboost algorithm, using default parameters, is capable of providing the most optimal level of accuracy. However, the potential for accuracy improvement can still be further explored by adjusting some key parameters. Specifically, increasing the number of estimators and reducing the learning rate in the xgboost model can be an effective strategy to enhance predictive performance. By optimizing both of these parameters, it is expected to achieve higher accuracy, strengthen the model's reliability, and enhance its ability to handle higher data complexity.

4.4. ROC Curve Analysis and Model Stability

From the analysis of the ROC curve, it is evident that k-fold cross-validation with xgboost stands out as the most effective algorithm for phishing email detection. The higher Area Under the Curve (AUC) compared to random forest and SVM indicates that this model has better capability in predicting the true class. With high accuracy, surpassing even other effective algorithms such as random forest, k-fold cross-validation with xgboost emerges as the primary choice for improving the reliability of phishing email detection.

 ISSN 2776-8007

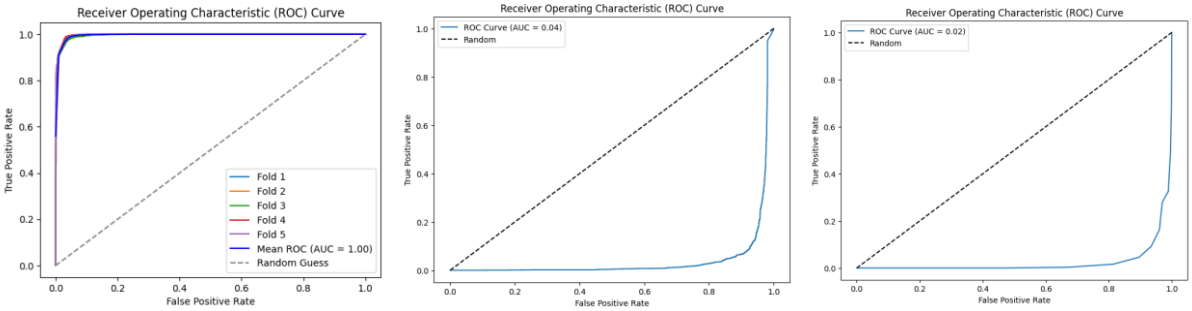

Figure 3. ROC AUC of K-Fold dan Xgboost, SVM, dan Random Forests

Furthermore, through model stability analysis, it was found that the accuracy of k-fold cross-validation combined with XGBoost remains relatively stable across each fold, indicating its resilience to data bias. This reinforces the conclusion that k-fold cross-validation and XGBoost are not only effective but also reliable in the context of email phishing detection. This reliability is demonstrated by their ability to achieve high accuracy, even when faced with relatively small datasets.

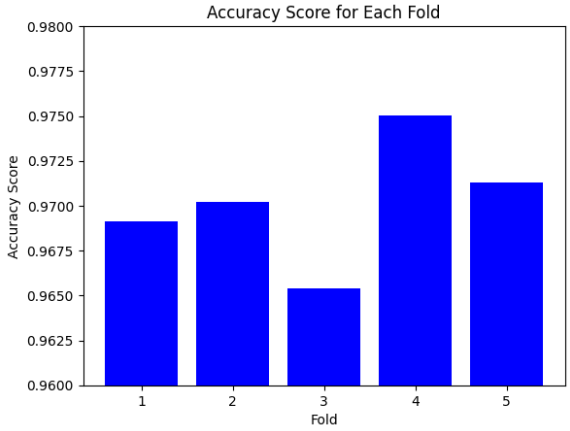

Figure 4. Accuracy scores for each fold

The significance of k-fold cross-validation and the XGBoost algorithm in phishing email detection lies not only in their effectiveness but also in their ability to handle diverse datasets with stability [18], [19], [20]. The k-fold cross-validation method ensures a thorough evaluation of the model by dividing the dataset into multiple subsets, thereby generating more accurate assessments of algorithm performance. Meanwhile, the superiority of XGBoost in handling complex and diverse data makes it a highly reliable choice for phishing email detection [21]. The combination of k-fold cross-validation and XGBoost establishes a robust foundation for enhancing email security, instilling confidence that this algorithm can address challenges posed by various evolving phishing attack types.

5. Conclusion

The research results indicate that this approach provides superior performance with the highest accuracy reaching 0.9712828770799785, making it the top choice in detecting potential phishing threats. The combination of k-fold cross-validation ensures high stability in the results, while parameter tuning in the xgboost algorithm offers opportunities for further accuracy improvement. A comparison with previous studies suggests that these findings bring significant new contributions, particularly in the context of phishing attacks. The practical implications of this research lie in the development of more effective and reliable email security strategies. Integrated with relevant literature and theories, this study reinforces the concept that the combination of k-fold and xgboost forms a robust foundation in phishing attack detection. This conclusion underscores the importance of a meticulous approach to phishing attacks, enhancing our understanding of the need for continually evolving security strategies. Furthermore, the potential for improvement through parameter tuning and further development of classification concepts offers intriguing avenues for future

research. Consequently, this research provides a significant contribution, expanding our understanding of phishing detection and laying a strong groundwork for better email security strategies in the future.

6. Declarations

6.1. Author Contributions

Conceptualization: D.S., R.A.P., C.I., and S.A.G.; Methodology: R.A.P.; Software: D.S.; Validation: D.S., R.A.P., C.I., and S.A.G.; Formal Analysis: D.S., R.A.P., C.I., and S.A.G.; Investigation: D.S.; Resources: R.A.P.; Data Curation: R.A.P.; Writing Original Draft Preparation: D.S., R.A.P., C.I., and S.A.G.; Writing Review and Editing: R.A.P. and D.S.; Visualization: D.S.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J. Doshi, K. Parmar, R. Sanghavi, and N. Shekokar, "A comprehensive dual-layer architecture for phishing and spam email detection," *Computers and Security*, vol. 133, p. 103378, 2023. doi:10.1016/j.cose.2023.103378

[2] E. J. D. Slifkin and M. B. Neider, "Phishing interrupted: The impact of task interruptions on phishing email classification," *International Journal of Human-Computer Studies*, vol. 174, p. 103017, 2023. doi:10.1016/j.ijhcs.2023.103017

[3] P. D. F. Isles, "A random forest approach to improve estimates of tributary nutrient loading," *Water Research*, vol. 248, p. 120876, 2024. doi:10.1016/j.watres.2023.120876

[4] H. Wang, W. Zhou, and Y. Shao, "A new fast admm for kernelless SVM classifier with truncated fraction loss," *Knowledge-Based Systems*, vol. 283, p. 111214, 2024. doi:10.1016/j.knosys.2023.111214

[5] S. Lias, N. A. Ali, M. Jamil, A. M. Jalil, and M. F. Othman, "Discrimination of pure and mixture agarwood oils via electronic nose coupled with k-nn kfold classifier," *Procedia Chemistry*, vol. 20, pp. 63–68, 2016. doi:10.1016/j.proche.2016.07.026

[6] Z. H. Wang et al., "Intelligent prediction model of mechanical properties of ultrathin niobium strips based on XGBoost ensemble learning algorithm," *Computational Materials Science*, vol. 231, p. 112579, 2024. doi:10.1016/j.commatsci.2023.112579

[7] J. Buckley, D. Lottridge, J. G. Murphy, and P. M. Corballis, "Indicators of employee phishing email behaviours: Intuition, elaboration, attention, and email typology," *International Journal of Human-Computer Studies*, vol. 172, p. 102996, 2023. doi:10.1016/j.ijhcs.2023.102996

[8] Alhogail and A. Alsabih, "Applying machine learning and natural language processing to detect phishing email," *Computers and Security*, vol. 110, p. 102414, 2021. doi:10.1016/j.cose.2021.102414

[9] R. Jayaraj et al., "Intrusion detection based on phishing detection with machine learning," *Measurement: Sensors*, vol. 1, no. 1, p. 101003, 2023. doi:10.1016/j.measen.2023.101003

[10] J. Doshi, K. Parmar, R. Sanghavi, and N. Shekokar, "A comprehensive dual-layer architecture for phishing and spam email

detection," *Computers and Security*, vol. 133, p. 103378, 2023. doi:10.1016/j.cose.2023.103378

[11] M. Butavicius, R. Taib, and S. J. Han, "Why people keep falling for phishing scams: The effects of time pressure and deception cues on the detection of phishing emails," *Computers and Security*, vol. 123, p. 102937, 2022. doi:10.1016/j.cose.2022.102937

[12] S. Magdy, Y. Abouelseoud, and M. Mikhail, "Efficient spam and phishing emails filtering based on Deep Learning," *Computer Networks*, vol. 206, p. 108826, 2022. doi:10.1016/j.comnet.2022.108826

[13] G. Petrič and K. Roer, "The impact of formal and informal organizational norms on susceptibility to phishing: Combining survey and field experiment data," *Telematics and Informatics*, vol. 67, p. 101766, 2022. doi:10.1016/j.tele.2021.101766

[14] Y. Zhou, X. Cui, W. Qu, and Y. Ge, "The effect of Automation Trust Tendency, system reliability and feedback on users' phishing detection," *Applied Ergonomics*, vol. 102, p. 103754, 2022. doi:10.1016/j.apergo.2022.103754

[15] M. P. Bach, T. Kamenjarska, and B. Žmuk, "Targets of phishing attacks: The bigger fish to fry," *Procedia Computer Science*, vol. 204, pp. 448–455, 2022. doi:10.1016/j.procs.2022.08.055

[16] S. H. Ahammad et al., "Phishing URL detection using machine learning methods," *Advances in Engineering Software*, vol. 173, p. 103288, 2022. doi:10.1016/j.advengsoft.2022.103288

[17] S. Rameem Zahra, M. Ahsan Chishti, A. Iqbal Baba, and F. Wu, "Detecting covid-19 chaos driven phishing/malicious URL attacks by a fuzzy logic and data mining based Intelligence System," *Egyptian Informatics Journal*, vol. 23, no. 2, pp. 197–214, 2022. doi:10.1016/j.eij.2021.12.003

[18] D.-J. Liu, G.-G. Geng, and X.-C. Zhang, "Multi-scale semantic deep fusion models for phishing website detection," *Expert Systems with Applications*, vol. 209, p. 118305, 2022. doi:10.1016/j.eswa.2022.118305

[19] N. Siddiqui, L. Chaudhary, P. Tripathi, N. Kumar, and S. Kumar, "A comparative analysis of US and Indian laws against phishing attacks," *Materials Today: Proceedings*, vol. 49, pp. 3646–3649, 2022. doi:10.1016/j.matpr.2021.08.256

[20] M. Frank, L. Jaeger, and L. M. Ranft, "Contextual drivers of employees' phishing susceptibility: Insights from a field study," *Decision Support Systems*, vol. 160, p. 113818, 2022. doi:10.1016/j.dss.2022.113818

[21] M. M. Alani and H. Tawfik, "Phishnot: A cloud-based machine-learning approach to phishing URL detection," *Computer Networks*, vol. 218, p. 109407, 2022. doi:10.1016/j.comnet.2022.109407