An Empirical Investigation of Phishing Detection Through Machine Learning Methods

Pandya Himani[1], Dr. Khyati Zalawadia[2], Dr. Harish Prajapati[3]

[1] *Department of Computer Engineering, Parul Institute of Engineering and Technology; Parul University; Vadodara; Gujarat e-mail:* himanipandya31@gmail.com

[2] *Department of Computer Engineering, Parul Institute of Engineering and Technology; Parul University; Vadodara; Gujarat e-mail: Khyati.Zalawadia29490@paruluniversity.ac.in*

[3] *Department of Computer Engineering, Parul Institute of Engineering and Technology; Parul University; Vadodara; Gujarat e-mail: harish.prajapati35068@paruluniversity.ac.in*

Abstract: Because of the increasing likelihood of deceptive attacks meant to steal sensitive data, phishing threat detection is essential to cybersecurity. The application of machine learning techniques to identify phishing attempts is described in this synopsis. To detect phishing traits, the procedure involves removing aspects from websites, email content, and URLs. To increase accuracy, methods such as OCR, NLP, and behavioural analysis are used.For classification, a variety of machine learning techniques are utilized, such as supervised learning (Logistic Regression, Decision Trees, SVMs), deep learning (CNNs, RNNs, Transformers), and ensemble approaches (GBM, XGBoost). By spotting odd patterns, unsupervised learning techniques like anomaly detection and clustering increase security.During implementation, data must be gathered and pre-processed, features must be engineered, models must be trained, and a real-time detection system that learns from fresh data must be put into place.

Keywords: Machine learning, Cyber-attack Phishing detection, RNN

1. INTRODUCTION

Using two different datasets, this study investigates the efficacy of 24 classifiers from six learning approaches for phishing detection. After evaluating the classifiers using eight different performance measures, the study finds that Random Forest and Filtered Classifier are the best algorithms, and that Info Gain Attribute Eval is the best feature selection method. The results show that Random Forest obtained the best accuracy and True Positive (TP) rate in Dataset 2, while the Filtered Classifier and J48 performed exceptionally well in Dataset 1. The study highlights Random Forest's strong performance in phishing detection and shows how sophisticated machine learning algorithms may be used with efficient feature selection to improve phishing detection systems.[1]

This research integrates crawling, detection, classification, and reporting algorithms to present a comprehensive defense against social engineering (SE) attacks on job-seeking websites. By increasing administrators' and users' knowledge, the model seeks to improve cybersecurity by enhancing the prevention of SE threats. With three bespoke apps, the suggested solution outperforms existing methods with a 72% detection accuracy and successfully distinguishes between dangerous and benign links. The study emphasizes user awareness as a crucial component in bolstering cybersecurity defenses and emphasizes the significance of education and proactive detection in reducing SE risks.[2]

The paper "Machine Learning Based Phishing Detection from URLs" offers a novel real-time anti-phishing system that can detect phishing URLs without the need for outside services. It does this by utilizing machine learning (ML) and natural language processing (NLP). The remarkable 97.98% accuracy of the Random Forest classifier highlights the system's remarkable adaptability. Although the study emphasizes how well NLP-based features, such as URL text semantic analysis, can improve phishing detection, it also points out several drawbacks, like the lack of a diverse dataset and the neglect of user experience issues. To further validate and improve the system's effectiveness, the authors suggest that future study increase the dataset and include user experience evaluations.[3]

In their paper "Business Email Compromise Phishing Detection Based on Machine Learning: A Systematic Literature Review," Hany F. Atlam and Olayonu Oluwatimilehin conduct a thorough review of 38 chosen articles to assess the effectiveness of machine learning techniques in detecting Business Email Compromise (BEC) phishing.It emphasizes the effectiveness of several algorithms—some of which have achieved detection rates of up to 98%—while highlighting the necessity of large, comprehensive datasets and the significance of elements like behavioral patterns, language characteristics, and email metadata in enhancing model accuracy. Although providing a well-organized summary, the paper notes many drawbacks, including the limited scope of the examined articles and the absence of comprehensive implementation details. It therefore advocates for further research to address these problems by creating a wider range of datasets.[4]

This research investigates how different feature selection and data processing methods, such as C-SVC, Random Tree (RndTree), C4.5, and Naïve Bayes, affect the effectiveness of phishing detection classifiers. Together with data transformations like Binary, Logarithmic, Original, and Relative, it assesses feature selection techniques like Original features, Forward Selection, Fisher Filtering, and ReliefF. The results show that although C-SVC performs well when it comes to binary transformation accuracy, it is not as effective when ROC analysis is used to evaluate performance; Random Tree is found to be the best performer based on AUC values. This emphasizes the need of AUC in assessing dependable performance and shows that accuracy by itself is not a complete indicator of classifier efficacy.[5]

The study presents a machine learning method for phishing detection that uses Support Vector Machines (SVM). It shows a high accuracy rate of 95.66% and a low false-positive rate. This approach performs noticeably better than earlier ones, especially when it comes to identifying fresh and fleeting phishing websites. The study recommends that in order to further improve the efficiency and dependability of phishing detection systems, future research should concentrate on examining the effects of feature selection and the use of various classification techniques.[6]

This study evaluates the efficacy of various machine learning algorithms in identifying phishing webpages using a dataset that includes 6,157 genuine and 4,898 phishing webpages described by 30 variables. Sequential Minimal Optimization (SMO), Random Forest (RF), J48, AdaBoost, Decision Table, IB1, and Naïve Bayes are among the techniques evaluated. In order to determine which aspects are most important, the study uses the information gain approach to pick features. According to the results, the best algorithm for phishing detection is Random Forest. Moreover, feature selection improves model performance, as shown by increased phishing detection accuracy and efficiency.[7]

Through the use of ensemble learning and hybrid feature selection, the study "Hybrid Feature Selection and Ensemble Learning Method for Spam Email Classification" improves the accuracy of spam email classification. The authors use Adaptive Boosting (AdaBoost) to enhance model performance and Information Gain to identify the most pertinent characteristics. Regarding precision and additional performance metrics, the hybrid technique outperformed the usual methods. The study's shortcomings, however, include its reliance on a single dataset and its restricted focus on a particular feature selection method, which might not adequately represent the diversity of many email datasets. Subsequent investigations may tackle these constraints by delving into supplementary feature selection methodologies, group methodologies, and conducting tests on diverse datasets to evaluate the applicability of the strategy.[8]

Using the Spambase UCI dataset, the performance of ten machine learning classifiers is assessed in the paper "Ham and Spam E-Mails Classification Using Machine Learning Techniques" by M. Bassiouni, M. Ali, and E. A. El-Dahshan. With an accuracy of 95.45%, the Random Forest classifier is shown to be the most efficient by the study. However, because of possible limits in the dataset's ability to reflect current spam characteristics, the study's conclusions might not be as applicable today. Furthermore, important real-time deployment considerations like classifier scalability and processing requirements are not included in the work.

More recent datasets should be investigated, classifier efficiency for real-time applications should be increased, and natural language processing methods should be integrated into future research.[9]

The efficiency of a 1D convolutional neural network (CNN) in identifying phishing websites is demonstrated in the study "High Accuracy Phishing Detection Based on Convolutional Neural Networks" by Suleiman Y. Yerima and Mohammed K. Alzaylaee. The CNN outperformed conventional classifiers within astounding F1-score of 0.976 and a detection rate of 98.2% after being trained on a dataset containing 6,157 legitimate and 4,898 phishing websites. The study not only demonstrates CNNs' higher phishing detection accuracy, but it also emphasizes the necessity for more investigation into various deep learning architectures and real-time applicability for realistic cybersecurity systems.[10]

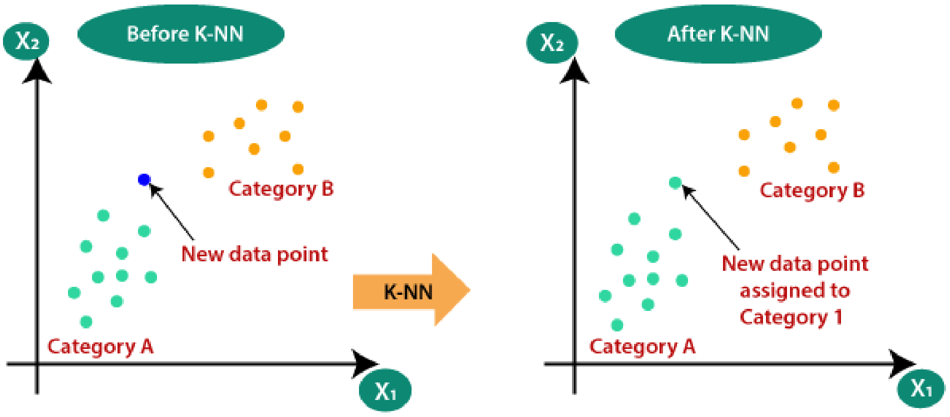
2. LITERATURE SURVEY

2.1 KNN (K-Nearest Neighbors)

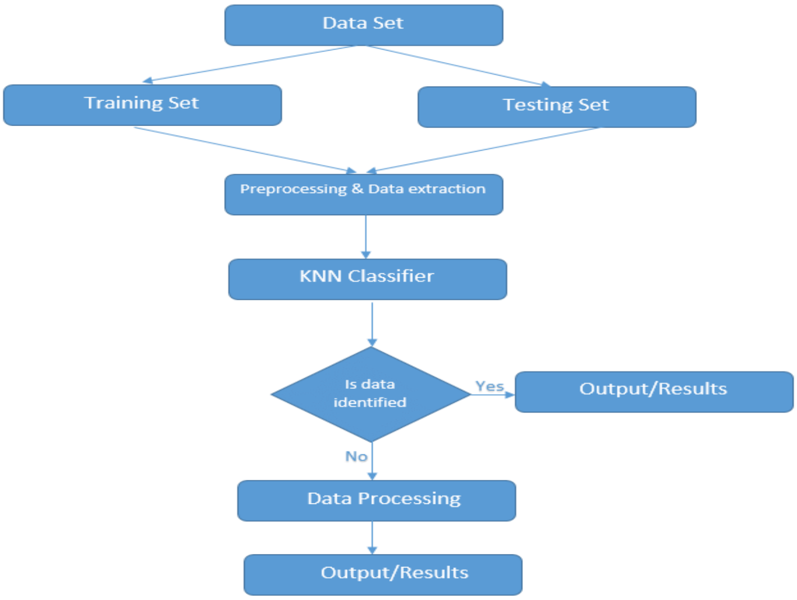
In the study, one of the 24 classifiers used to identify phishing is k-Nearest Neighbors (KNN). The study shows that while KNN is well-known for its ease of use and efficiency in a variety of classification tasks, it was not as successful in phishing detection as some of the other classifiers, like Random Forest and Filtered Classifier. More advanced algorithms are better suited to handle the high dimensionality and intricate patterns included in phishing datasets, which presented challenges for KNN. Although KNN showed respectable accuracy, classifiers that could take advantage of more sophisticated learning approaches and feature selection strategies, such as Info Gain Attribute Eval, fared much better overall. As a result, KNN's involvement in this study emphasizes its shortcomings in handling phishing detection in contrast to other.[1]

The study paper highlights the significance of the k-Nearest Neighbors (KNN) algorithm in identifying and categorizing fake relationships on job-seeking websites. By examining patterns in the dataset, KNN is used to differentiate between linkages that are harmful and those that are not. It is appropriate for this task because of its non-parametric character, which enables it to categorize data points depending on how close they are to pre-labelled instances. The method uses the similarity between data points to identify social engineering threats, which improves the overall detection accuracy of the model. Because KNN can adjust to a variety of input data formats, it can reliably identify and flag abnormal activity, improving cybersecurity resilience.[2]

Among the seven machine learning methods assessed for phishing URL detection in the research paper "Machine Learning Based Phishing Detection from URLs," the K-Nearest Neighbors (KNN) algorithm was featured. KNN is a distance-based algorithm that uses factors like text structure and semantic patterns inside the URLs to classify URLs by comparing them with the most comparable URLs in the training dataset. While KNN is widely acknowledged for its ease of use and efficiency in addressing specific classification tasks, the research revealed that KNN's performance was inferior to that of certain other algorithms, such Random Forest, in this particular scenario. The study showed that although KNN could identify phishing URLs with a respectable degree of accuracy, its efficiency was inferior to that of more sophisticated algorithms, especially when it came to managing subtle patterns.[3]



(Fig.1 KNN classification)
(Source: static. java point)



(Fig.1 KNN flow)
(Source: ResearchGate)

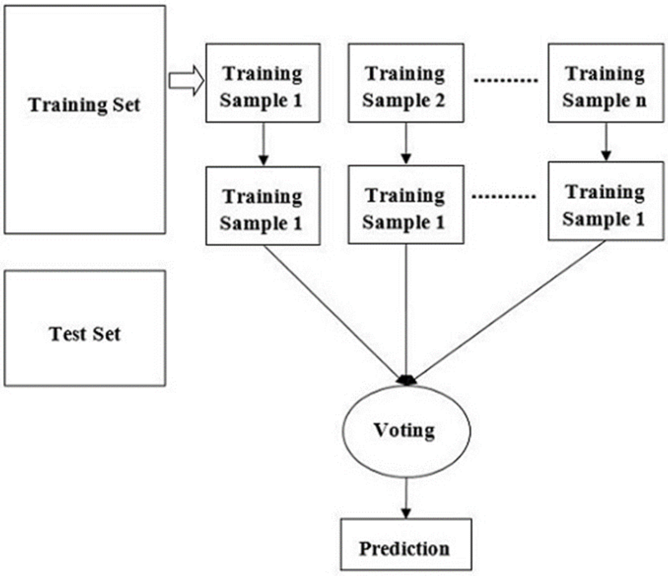
Hany F. Atlam and Olayonu Oluwatimilehin examine the workings of the k-Nearest Neighbors (KNN) algorithm, one of the machine learning algorithms assessed for its efficacy in identifying phishing attempts, in the systematic literature review on BEC phishing detection. It was discovered that phishing emails could be categorized according to how much they resembled actual or well-known phishing email clusters using KNN, which is well-known for being intuitive and efficient at identifying patterns. The research also notes that while KNN can achieve high detection rates, especially when well-trained on different datasets, its performance is heavily dependent on the caliber and diversity of the training data. Furthermore, because of its computational complexity, KNN may become less successful with larger datasets, making it less scalable than other more sophisticated.[4]

2.2 Random Forest:

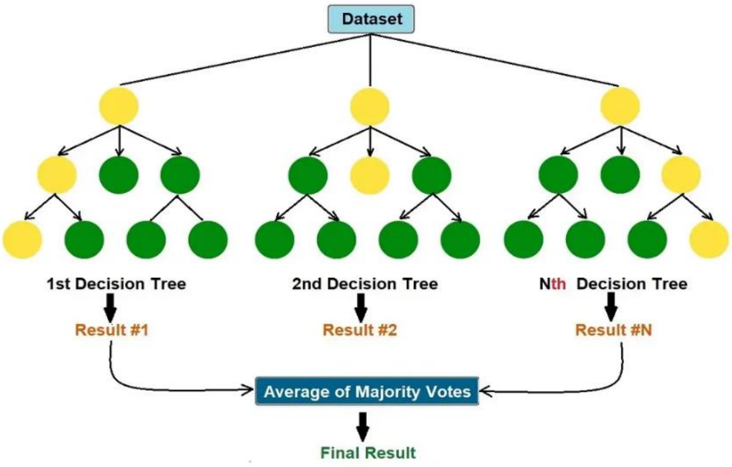
The study finds that when it comes to correctly identifying phishing webpages, the Random Forest (RF) algorithm performs better than other machine learning algorithms. In the study, Random Forest outperforms SMO, J48, and Naïve Bayes algorithms in phishing site detection, using a dataset of 30 variables from 4,898 phishing and 6,157 legitimate websites. The study also shows that when Random Forest is used in conjunction with feature selection through the Information Gain method, it retains its higher detection accuracy and becomes even more effective. The model's build time is shortened by this feature selection process, emphasizing two of Random Forest's main benefits: enhanced computational efficiency with optimized features and high detection accuracy for phishing attacks. [7]

Although random forest (RF) is not the main topic of the study "Hybrid Feature Selection and Ensemble Learning Method for Spam Email Classification," it plays a critical function in setting a baseline for comparison when considering ensemble learning techniques. The study's suggested hybrid method, which combines Information Gain for feature selection with Adaptive Boosting (AdaBoost) for classification accuracy, outperforms conventional methods like Random Forest, despite the fact that Random Forest is a potent ensemble learning technique that builds multiple decision trees to improve classification accuracy. While Random Forest is well-known for its ability to manage huge datasets and minimize overfitting, the research shows that the hybrid technique produces better performance measures, such as increased accuracy. This implies that even while Random Forest is a formidable competitor in the spam.[8]

Among ten assessed machine learning models, the Random Forest (RF) classifier proved to be the most successful in identifying spam emails, according to the study E. A. El-Dahshan, M. Bassiouni, and M. Ali's paper "Ham and Spam E-Mails Classification Using Machine Learning Techniques" With a 95.45% accuracy rate on the UCI Spam base dataset, it demonstrated its exceptional ability to differentiate between spam and legitimate emails. Random Forest's strength is its ensemble learning method, which combines several decision trees to improve classification accuracy, performance, and minimize overfitting. The study does point out a potential drawback to the model's generalizability, though, in that the attributes of contemporary spam emails might not be fully captured in the Spam base dataset.[9]



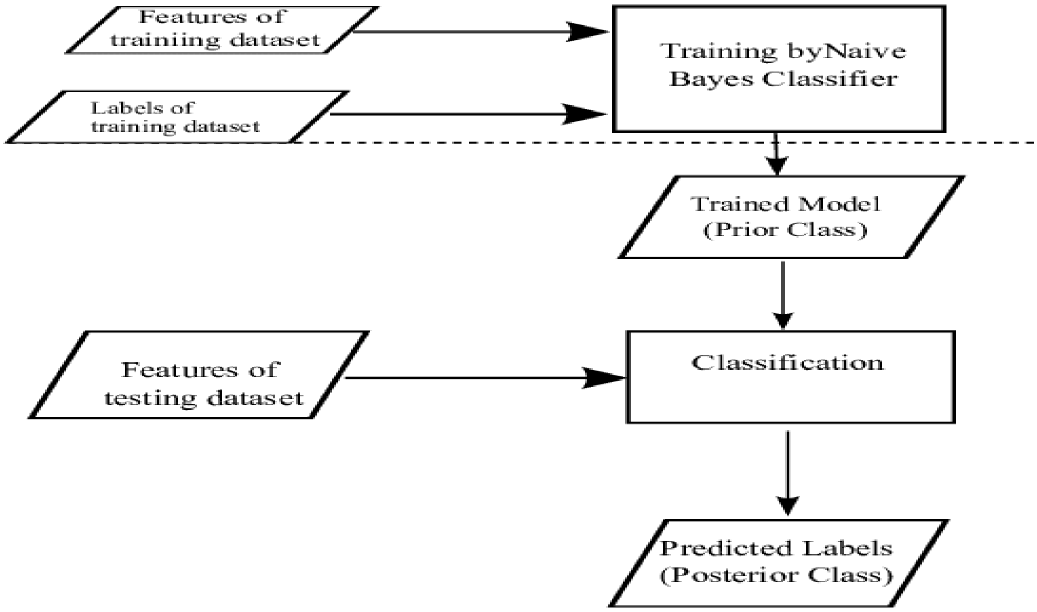
(Fig 3 Random Forest flow)
(Source: tutorials point)



(Fig 4 Random Forest classification)
(Source: ejable 2023)

2.3 Naive Bayes:

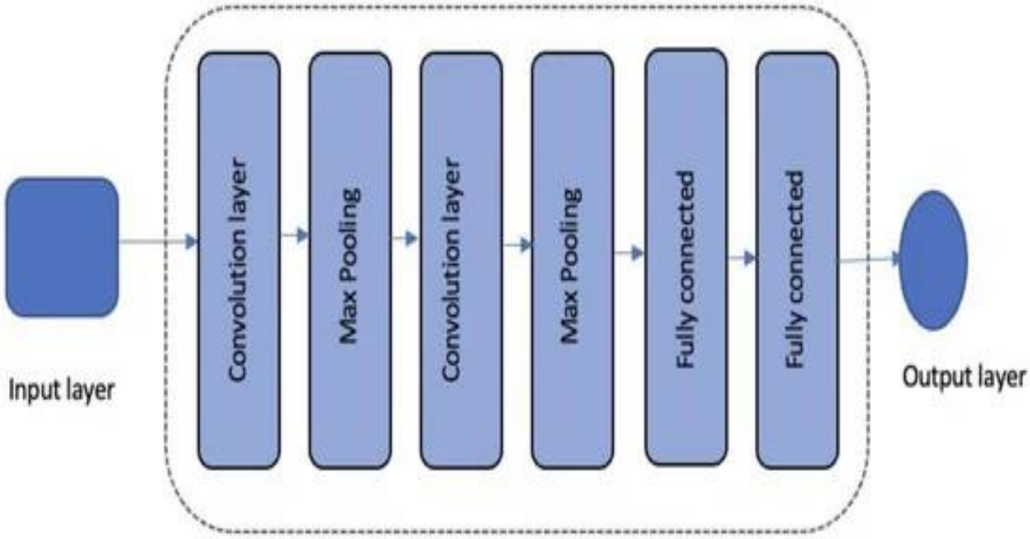
The Naïve Bayes classifier is singled out for praise in the research investigating the effects of feature selection and data processing approaches on phishing detection due to its adaptability to various transformations and feature selection strategies. When binary transformation is used, Naïve Bayes has the lowest performance among the examined classifiers, despite its simplicity and efficiency in many classification applications. This result implies that Naïve Bayes, in contrast to more advanced classifiers like C-SVC and Random Tree, may find it difficult to identify the intricate patterns needed for successful phishing detection. Additionally, the study emphasizes that accuracy on its own isn't a sufficient performance metric; Naïve Bayes continues to fall behind when assessed using ROC analysis and AUC values, highlighting its limits in accurately identifying phishing from legitimate occurrences in this background. According to this research, Naïve Bayes may not be the greatest option for phishing detection, particularly in situations when greater accuracy and dependability are essential, even though it can be helpful for smaller tasks.[5]



(Fig 5 Navie Bayes flow diagram)
(Source: ResearchGate)

2.4 CNN (Cable News Network):

Hany F. Atlam and Olayonu Oluwatimilehin examine the workings of the k-Nearest Neighbors (KNN) algorithm, one of the machine learning algorithms assessed for its efficacy in identifying phishing attempts, in the systematic literature review on BEC phishing detection. KNN, renowned for its ease of use and efficacy in identifying patterns, was observed to have the ability to categorize phishing emails according to their closeness to legitimate or known phishing email clusters. The research also notes that while KNN can achieve high detection rates, especially when well-trained on different datasets, its performance is heavily dependent on the caliber and diversity of the training data. Furthermore, because of its computational complexity, KNN may become less successful with larger datasets, making it less scalable than other more sophisticated.[10]



(Fig 6 CNN basic architecture.)
(Source:[11])

3. CONCLUSION

Through the classification of URLs, emails, and web pages as phishing or legitimate, the project seeks to construct a powerful machine learning-based system for real-time phishing threat detection. The system is intended to achieve high accuracy by utilizing a variety of machine learning algorithms and techniques, including supervised, unsupervised, and deep learning methods. Ensemble approaches, feature selection, and hyperparameter tuning are important techniques for improvement. Future research aims to include continuous learning to adjust to changing phishing threats and real-time detection capabilities into web applications or email clients.

4.REFERENCES

[1] Thakral, I., Kumari, S., Singh, K. K., & Aggarwal, N. (2023, November). An Advanced IoT Based Border Surveillance and Intrusion Detection System. In*2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS)* (pp. 1193-1198). IEEE.

[2] , K., Kodidela, P., & Gurram, P. (2021). IoT based smart intruder detection system for smart homes. *International Journal of Scientific Research in Science and Technology*, *8*(4), 48-53.

[3] Iyer, S., Gaonkar, P., Wadekar, S., Kohmaria, N., & Upadhyay, P. (2020, April). IoT based Intruder Detection System Using GSM. In Proceedings of the 3rd International Conference on Advances in Science & Technology (ICAST).

[4] Golder, A., Gupta, D., Roy, S., Al Ahasan, M. A., & Haque, M. A. (2023, October). GSM Based Home Security Alarm System Using Arduino Using Mobile Call. In 2023 IEEE 14th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON) (pp. 0268-0274). IEEE.

[5] Sahoo, K. C., & Pati, U. C. (2017, May). IoT based intrusion detection system using PIR sensor. In 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT) (pp. 1641-1645). IEEE.

[6] J. Rashid, T. Mahmood, M. W. Nisar and T. Nazir, "Phishing Detection Using Machine Learning Technique," 2020 First International Conference of Smart Systems and Emerging Technologies (SMARTTECH), Riyadh, Saudi Arabia, 2020, pp. 43-46, doi: 10.1109/SMART-TECH49988.2020.00026.

[7] Gandotra, Ekta, and Deepak Gupta. "An efficient approach for phishing detection using machine learning." Multimedia security: algorithm development, analysis and applications (2021): 239-253.

[8] Ablel-Rheem, Doaa Mohammed, et al. "Hybrid feature selection and ensemble learning method for spam email classification." International Journal 9.1.4 (2020): 217-223.

[9] Bassiouni, Mahmoud, M. Ali, and E. A. El-Dahshan. "Ham and spam e-mails classification using machine learning techniques." Journal of Applied Security Research 13.3 (2018): 315-331.

[10] Yerima, Suleiman Y., and Mohammed K. Alzaylaee. "High accuracy phishing detection based on convolutional neural networks." 2020 3rd International Conference on Computer Applications & Information Security (ICCAIS). IEEE, 2020.

[11] Alshingiti, Z., Alaqel, R., Al-Muhtadi, J., Haq, Q. E. U., Saleem, K., & Faheem, M. H. (2023). A deep learning-based phishing detection system using CNN, LSTM, and LSTM-CNN. Electronics, 12(1), 232.

[12] Do, N.Q.; Selamat, A.; Krejcar, O.; Herrera-Viedma, E.; Fujita, H. Deep Learning for Phishing Detection:

Taxonomy, current challenges and Future Directions. IEEE Access 2022, 10, 36429–36463. [Google Scholar] [CrossRef]

[13] Zhang, Q.; Bu, Y.; Chen, B.; Zhang, S.; Lu, X. Research on phishing webpage detection technology based on CNN-BiLSTM algorithm. *J. Phys. Conf. Ser.* **2021**, *1738*, 012131. [Google Scholar] [CrossRef]