Real-Time Email Phishing Detection Using a Custom DistilBERT Model

Edafe Maxwell Damatie
Department of Computing and Mathematics
Manchester Metropolitan University
Manchester, England
edafe.m.damatie@stu.mmu.ac.uk

Amna Eleyan
Department of Computing and Mathematics
Manchester Metropolitan University
Manchester, England
a.eleyan@mmu.ac.uk

Tarek Bejaoui
Computer Engineering Department
University of Carthage
Tunisia
tarek.bejaoui@ieee.org

***Abstract*— This paper presents a real-time email phishing detection system that utilizes a custom DistilBERT model. The custom DistilBERT architecture incorporates dynamic threshold adjustment and an enhanced classifier head, optimized for analyzing email content. With detection response times of under two seconds, the system delivers real-time protection. Experimental results demonstrate 99.29% accuracy in controlled tests and 95.45% in real-world tests, surpassing current state-of-the-art methods. The system maintains high performance at low false positive rates, essential for practical deployment. An adaptive daily retraining mechanism ensures continued effectiveness against evolving phishing tactics. This research advances email security and offers insights into the use of transformer-based models in real-time cybersecurity applications. By addressing the limitations of current systems through distilled transformer models, this work significantly strengthens organizational cybersecurity against advanced phishing threats.**

Keywords—Email phishing detection, DistilBERT, Real-time detection, Adaptive retraining, Transformer models, False positive rate

I. INTRODUCTION

Email phishing continues to be a significant cybersecurity threat, with attacks becoming more sophisticated and frequent. The Anti-Phishing Working Group (APWG) reported 165,772 phishing attacks in the first quarter of 2020, up from 162,155 in the previous quarter, highlighting the persistent and evolving nature of these threats [1, 11, 14].

This paper presents a novel real-time email phishing detection system that utilizes a custom DistilBERT model. DistilBERT, a compressed version of BERT, retains 97% of BERT's language comprehension capabilities while being 40% smaller and 60% faster [4]. This efficiency enables the proposed system to address key limitations in current approaches through several innovations. The system incorporates a custom DistilBERT architecture featuring dynamic threshold adjustment and an enhanced classifier head, specifically optimized for email content analysis. Additionally, it achieves real-time detection with response times under two seconds, a significant improvement over many existing solutions.

Experimental results demonstrate the effectiveness of this approach, achieving 99.29% accuracy in controlled tests and 95.45% in real-world scenarios. These results surpass the performance of current state-of-the-art methods, particularly at very low false positive rates, where phishing detection systems must operate effectively [5].

This research not only enhances email security but also offers valuable insights into deploying transformer-based models in practical, real-time cybersecurity applications. By addressing the shortcomings of current systems and leveraging the efficiency of distilled transformer models, this work contributes to the broader effort of improving organizational cybersecurity.

The rest of the paper is organized as follows: Section II provides a comprehensive review of Related Work, covering recent advancements in phishing detection techniques and identifying existing gaps. Section III details the Methodology of the approach, including the custom DistilBERT model architecture, data preprocessing, Gmail integration, and daily retraining mechanism. Section IV presents the Experimental Setup and Results, demonstrating the system's performance in both controlled and real-world environments. Section V offers a Discussion of the results, interpreting the findings and highlighting the novelty and challenges of the approach. Finally, Section VI concludes the paper and outlines Future Work directions for further enhancing email phishing detection systems.

II. RELATED WORK

Recent research in phishing detection has increasingly focused on advanced machine learning and deep learning techniques. This section provides an overview of key recent works and highlights the gaps that the proposed approach aims to address.

A. Summary of Recent Phishing Detection Techniques

[2] proposed a Deep Neural Network (DNN) approach for phishing URL detection, achieving accuracy rates of 90% for Ham, 92% for Phishing Corpus, and 89% for Phishload.

979-8-3503-6491-0/24/$31.00 ©2024 IEEE

Their work demonstrated the potential of deep learning in handling large data volumes and uncovering latent structures in phishing data.

[3] introduced PhishKiller, a tool that utilizes featureless machine learning techniques. Their approach achieved 98.30% accuracy and could block malicious websites in 81.68 milliseconds, emphasizing the importance of real-time detection capabilities [12].

[6] developed a tool that uses feature selection and neural network algorithms, achieving up to 97.3% accuracy. Their system's ability to perform real-time detection and quarantine suspicious emails highlights the significance of immediate threat prevention.

More recent studies have further advanced phishing detection techniques. [8] presented a deep learning-based framework implemented as a browser plug-in, with their RNN-GRU model achieving 99.18% accuracy. [7] introduced a deep learning-based approach using a hybrid CNN-LSTM model, achieving over 95% accuracy in detecting phishing cyber-attacks.

B. Gap Analysis

Despite recent advancements in phishing detection, several critical gaps remain in current approaches. Many advanced models, particularly deep learning methods, require substantial computational resources for training and inference, posing challenges for real-time detection in resource-constrained environments [2, 3, 8]. While some studies have addressed real-time detection [3, 6], maintaining high accuracy with minimal latency remains a significant challenge, especially in the context of email phishing detection [13]. Additionally, most current systems lack mechanisms for continuous learning and adaptation to new phishing techniques, a crucial limitation given the rapidly evolving nature of phishing attacks.

Achieving high accuracy while maintaining low false positive rates in real-world scenarios is another persistent issue [2, 3, 7]. Furthermore, many studies focus primarily on model performance in controlled environments, leaving a noticeable gap in research that comprehensively compares performance in both controlled and real-world scenarios [2, 5, 6, 8]. This gap limits a full understanding of model effectiveness across different operational contexts.

The research presented here addresses these gaps by introducing a custom DistilBERT-based approach optimized for real-time email phishing detection. Leveraging the efficiency of DistilBERT and incorporating novel architectural modifications, the system is designed to be computationally efficient, highly accurate, and adaptable to real-world challenges in email environments.

III. METHODOLOGY

The proposed approach leverages a custom DistilBERT model for real-time email phishing detection. This section outlines the key components of the methodology.

A. Dataset

A comprehensive email dataset from Kaggle [9] is utilized, consisting of 82,486 entries with a nearly balanced distribution between phishing and non-phishing emails, as shown in Table 1.

Category	**Number of Emails**	**Percentage**
Phishing	43,057	52.20%
Non-Phishing	39,429	47.80%
Total	82,486	100%

Table 1: Email Dataset Composition

B. Data Preprocessing and Automated Feature Extraction

The preprocessing pipeline, made specifically for email data, consists of several key steps, as illustrated in Figure 1. The data is cleaned, relevant columns are identified, and the email content is tokenized using the DistilBERT tokenizer, with a maximum of 128 tokens. A key component is the automated feature extraction, utilizing DistilBERT's architecture to capture complex patterns without the need for manual feature engineering.

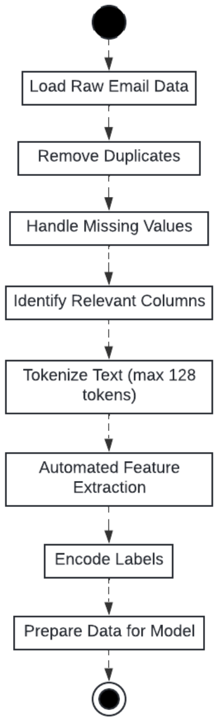

Figure 1:Preprocessing Pipeline Flowchart

C. Custom DistilBERT Model Architecture

The DistilBERT architecture is enhanced with several modifications, as summarized in Table 2.

Component	Standard DistilBERT	Custom Modifications
Classification Threshold	Fixed	Dynamic (learnable alpha parameter)
Classifier Head	Linear classifier	Two-layer feedforward network with ReLU activation
Loss Function	Standard cross-entropy	Combined cross-entropy with FPR penalty term
Regularization	Standard dropout	Dropout rate of 0.3

Table 2: Custom DistilBERT Architecture Modifications

These modifications, summarized in Table 2, enhance the model's adaptability and performance in phishing detection [12].

D. Integration with Gmail for Real-time Detection

A robust Gmail integration is implemented using the Gmail API [10] to enable continuous monitoring and real-time analysis of incoming emails. Figure 2 illustrates the data flow within the system.

The system authenticates via OAuth 2.0, efficiently queries for new emails, processes them through the pipeline, and classifies them using the custom DistilBERT model. Based on the classification, it applies a "PHISHING_WARNING" label and moves suspected phishing emails out of the inbox.

E. Daily Retraining Mechanism

The system incorporates a daily retraining mechanism to maintain effectiveness against evolving phishing techniques (Figure 3).

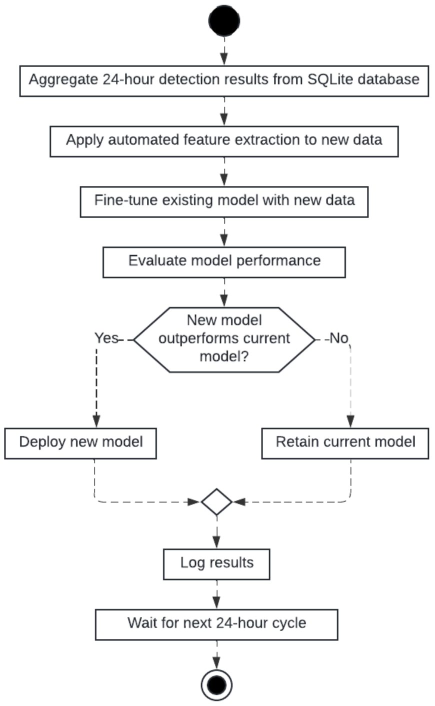
Aggregate 24-hour detection results from SQLite database
Apply automated feature extraction to new data
Fine-tune existing model with new data
Evaluate model performance
New model outperforms current model?
Yes
No
Deploy new model
Retain current model
Log results
Wait for next 24-hour cycle

Figure 3:Daily Retraining Process Flowchart

Every 24 hours, it aggregates new detection results from the previous day, which are then subjected to automated feature extraction to provide fresh data for model updating. The existing DistilBERT model is fine-tuned using this new data, allowing it to learn from the most recent phishing patterns.

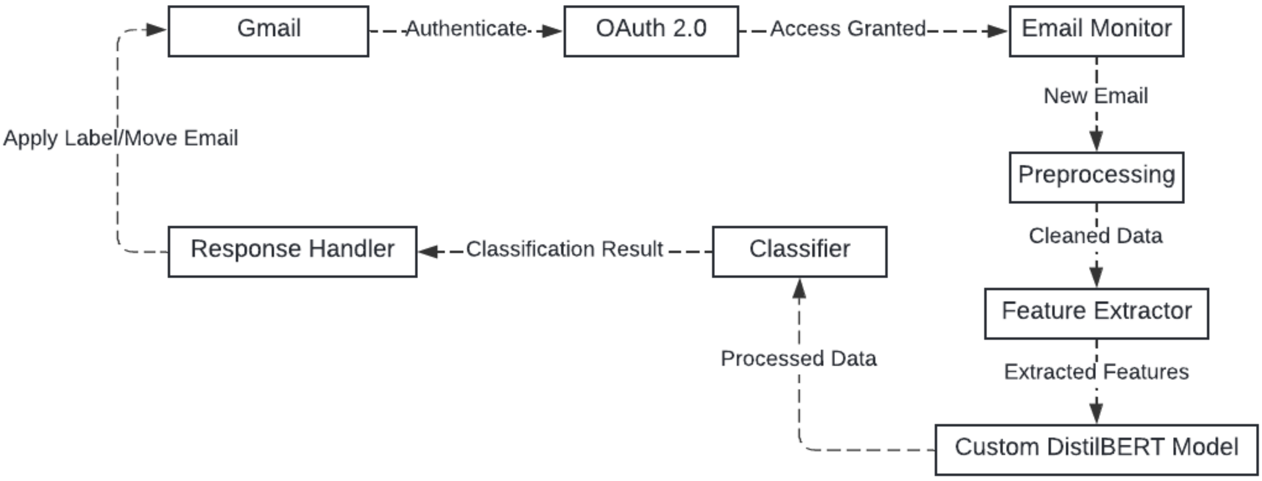

Figure 2:Data Flow Diagram

IV. EXPERIMENTAL SETUP AND RESULTS

The custom DistilBERT-based phishing detection system was evaluated in both controlled and real-world environments, and its performance was compared to existing state-of-the-art methods.

A. Controlled Environment Testing

In the controlled environment, a dataset of 82,486 [9] emails were used, with the data split 70-15-15 for training, validation, and testing. The model's performance metrics in this environment are summarized in Table 3.

Metric	**Value**
Accuracy	99.29%
Precision	99.29%
Recall	99.29%
F1-Score	99.29%
AUC-ROC	0.9994
FPR	0.69%

Table 3:Controlled Environment Performance Metrics

Figure 4 shows the ROC curve for our model, demonstrating its exceptional performance with an AUC of 0.9994.

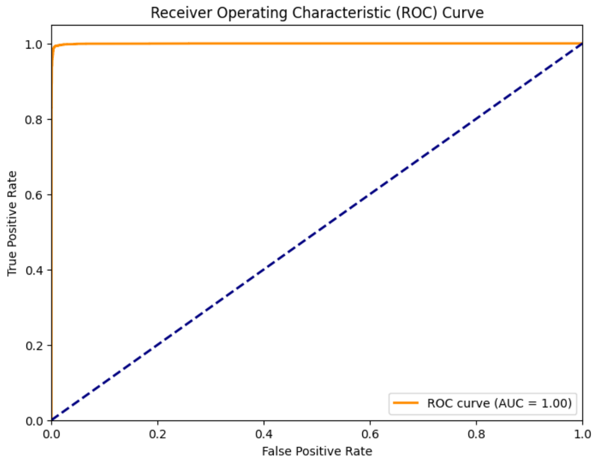

Figure 4: ROC Curve for Email Phishing Detection Model

B. Real-World Performance Analysis

The system was integrated with Gmail for real-time email monitoring via the Gmail API, utilizing OAuth 2.0 for secure authentication. During the testing period, the system processed 22 incoming emails, leveraging Gmail's history feature for efficient querying while ensuring timely

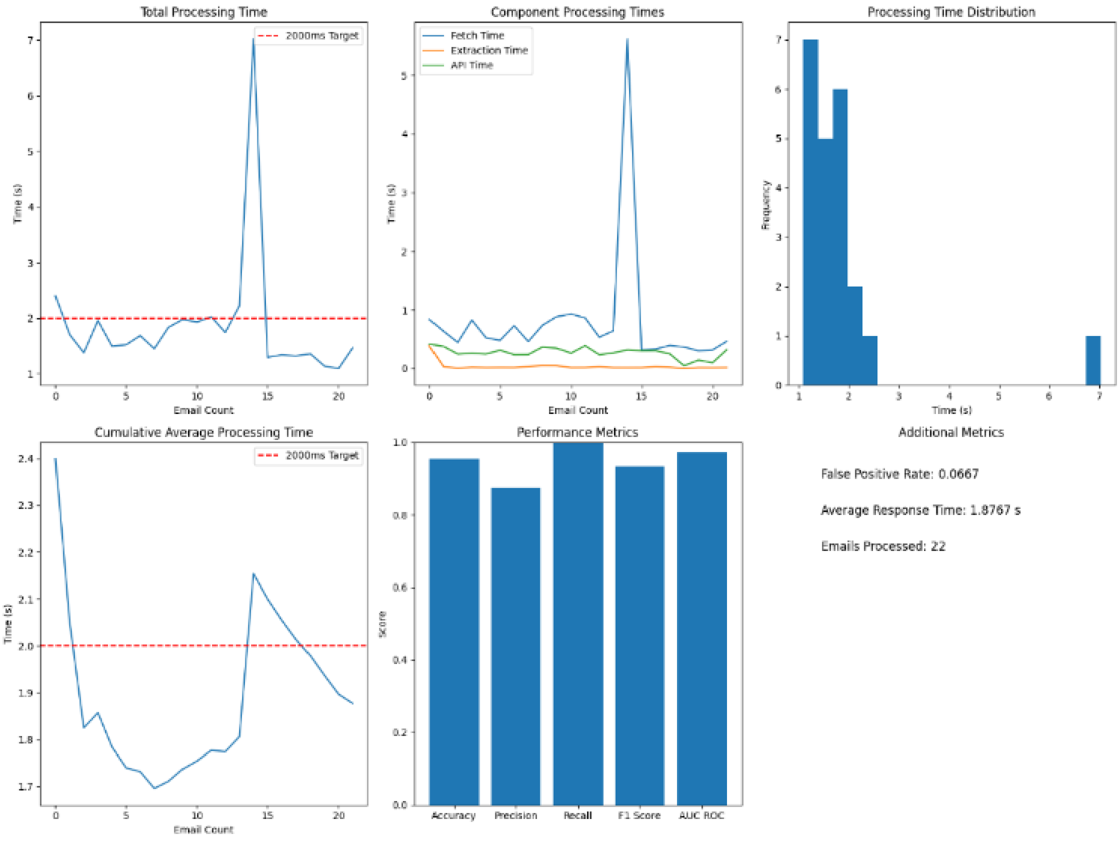

Figure 5: Use Case Performance Matrix for Email Phishing Detection System

processing. Each email followed a workflow involving fetching, feature extraction, preprocessing, and inference with the DistilBERT-based phishing detection model. Key extracted features included sender information, subject lines, email body content, and various metadata. The prediction results were stored in an SQLite database and used for system performance evaluation, revealing promising outcomes, with 21 out of 22 emails correctly classified.

Figure 5 illustrates the system's performance matrix in real-world conditions, showing processing times and their distribution

Table 4 presents the results of the system's performance in real-world testing.

Metric	Value
Accuracy	95.45%
Precision	87.50%
Recall	100%
F1-Score	93.33%
FPR	6.67%
Accuracy	95.45%
Avg. Response Time	1.88s

Table 4: Real-World Performance Metrics

Figure 6 presents the confusion matrix for real-world testing, highlighting the model's high accuracy and perfect recall.

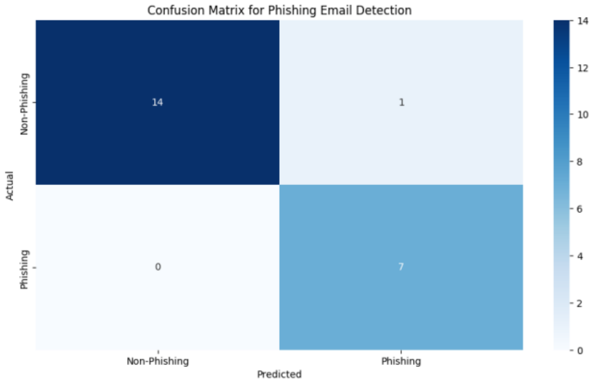

Figure 6: Use Case Confusion Matrix for Email Phishing Detection System

The system maintained high accuracy in real-world conditions, with an average response time of 1.88 seconds, successfully meeting the sub-2-second target for real-time detection.

C. *Comparative Analysis*

The custom DistilBERT model was compared with several state-of-the-art phishing detection approach, as shown in Table 5 and Table 6.

Model	Accuracy	FRP	F1-Score
[2] DNN Approach	92.00%	8.00%	91.96%
[3] PhishKiller	98.30%	1.70%	98.29%
[6] Neural Network	97.30%	2.70%	97.29%
[8] RNN-GRU	99.18%	0.82%	99.17%
Custom DistilBERT Model	99.29%	0.69%	99.29%

Table 5: Comparative Analysis of Phishing Detection Systems

Model	Real-time Capability
[2] DNN Approach	No
[3] PhishKiller	Yes (81.68ms)
[6] Neural Network	Yes (unspecified)
[8] RNN-GRU	Yes (unspecified)
Custom DistilBERT Model	Yes (1.88s)

Table 6: Comparison of Real-time Capabilities Across Phishing Detection Models

The model achieved the highest accuracy among the compared approaches, slightly outperforming even [8] recent framework. Notably, the model maintained this high accuracy while achieving a low false positive rate of 0.69% in controlled environments and 6.67% in real-world testing This combination of high accuracy and a low false positive rate addresses a key challenge in the practical deployment of email phishing detection systems [13].

V. DISCUSSION

A. *Performance Analysis*

The custom DistilBERT model demonstrated strong performance in both controlled and real-world environments. In controlled testing, the model achieved an accuracy of 99.29%, along with a low false positive rate (FPR) of 0.69%, indicating its robust ability to discriminate between phishing and legitimate emails. The high AUC-ROC value of 0.9994 further supports the model's strong discriminatory capability.

In real-world testing, the model maintained high accuracy (95.45%) with perfect recall and an average response time of 1.88 seconds, meeting the real-time detection target and demonstrating suitability for practical applications.

B. *Novelty and Challenges*

The proposed approach introduces several novel aspects to email phishing detection. It demonstrates the feasibility of utilizing DistilBERT for real-time phishing detection, effectively balancing computational efficiency with high accuracy. The model incorporates a dynamic threshold mechanism with learnable parameters, enhancing its adaptability to evolving phishing tactics. Additionally, the system’s seamless integration with Gmail bridges the gap between theoretical models and practical deployment, a step that is often overlooked in academic research.

Despite these innovations, challenges remain. A performance gap between controlled and real-world environments highlights the complexity of real-world email ecosystems. Balancing high detection rates with low false positives remains a significant challenge, particularly in practical settings where false alarms can diminish system effectiveness. Moreover, the computational demands of daily retraining present scalability concerns for large-scale deployment. Addressing these challenges is essential for advancing practical, real-time phishing detection systems.

VI. CONCLUSION AND FUTURE WORK

This paper presents a novel approach to real-time email phishing detection using a custom DistilBERT model. Key contributions include an optimized architecture for email phishing detection, real-time capability, robust performance in both controlled and real-world settings, and an adaptive system with daily retraining. These features collectively ensure continued effectiveness against evolving phishing tactics, demonstrating significant potential for improving email security in real-world environments.

Future work will focus on expanding real-world testing to validate performance across diverse email ecosystems, further refining the model to minimize false alarms without compromising detection rates and extending the system's compatibility with other popular email platforms. Additionally, efforts will aim to optimize the retraining process for more efficient large-scale deployments and investigate the model's resilience against adversarial attacks. By addressing these areas, the system could become a more powerful and versatile tool in the ongoing fight against phishing attacks, contributing to safer digital communications across various platforms and use cases.

REFERENCES

[1] Alkhalil, Z., Hewage, C., Nawaf, L., & Khan, I., 2021. Phishing Attacks: A Recent Comprehensive Study and a New Anatomy. , 3. https://doi.org/10.3389/fcomp.2021.563060.

[2] Sumathi, K., & Sujatha, V., 2019. Deep Learning Based-Phishing Attack Detection. International Journal of Recent Technology and Engineering. https://doi.org/10.35940/ijrte.c6527.098319.

[3] Martins de Souza, C. H., Lemos, M. O., Dantas Silva, F. S., & Souza Alves, R. L. (2019). On detecting and mitigating phishing attacks through featureless machine learning techniques. Internet Technology Letters, 3(1), e135. https://doi.org/10.1002/itl2.135.

[4] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

[5] [5] Maneriker, P., Stokes, J.W., Lazo, E.G., Carutasu, D., Tajaddodianfar, F., & Gururajan, A. (2021). URLTran: Improving Phishing URL Detection Using Transformers. MILCOM 2021 - 2021 IEEE Military Communications Conference (MILCOM), 197-204.

[6] D. Oña et al., "Phishing Attacks: Detecting and Preventing Infected E-mails Using Machine Learning Methods," CSNet, 2019.

[7] Kaushik, P., & Rathore, S., 2023. Deep Learning Multi-Agent Model for Phishing Cyber-attack Detection. International Journal on Recent and Innovation Trends in Computing and Communication. https://doi.org/10.17762/ijritcc.v11i9s.7674.

[8] Tang, L., & Mahmoud, Q., 2022. A Deep Learning-Based Framework for Phishing Website Detection. IEEE Access, 10, pp. 1509-1521. https://doi.org/10.1109/ACCESS.2021.3137636.

[9] Alam, N., 2021. Phishing Email Dataset. *Kaggle*. Available at: https://www.kaggle.com/datasets/naserabdullahalam/phishing-email-dataset [Accessed: 5 August 2024].

[10] Google Developers 2024, *Gmail API overview*, Google Developers. Available at: https://developers.google.com/gmail/api/guides [Accessed: 5 August 2024].

[11] Odeh, N., Eleyan, D. & Eleyan, A., 2021. A survey of social engineering attacks: Detection and prevention tools. Journal of Theoretical and Applied Information Technology, 99(18), pp.4375-4378.

[12] Dawabsheh, A., Jazzar, M., Eleyan, A., Bejaoui, T. & Popoola, S., 2022. An enhanced phishing detection tool using deep learning from URL. In: 2022 International Conference on Smart Applications, Communications and Networking (SmartNets), Palapye, Botswana, pp. 1-6. DOI: 10.1109/SmartNets55823.2022.9993984.

[13] Ammar, N., Eleyan, D., Jazzar, M. & Eleyan, A., 2021. Social engineering attacks: A phishing case simulation. International Journal of Scientific & Technology Research, 10(6), pp.10-16.

[14] Abubaker, A.A., Eleyan, D., Eleyan, A., Bejaoui, T., Katuk, N. & Al-Khalidi, M., 2023. Social engineering in social network: A systematic literature review. In: 2023 International Symposium on Networks, Computers and Communications (ISNCC), Doha, Qatar, pp. 1-7. DOI: 10.1109/ISNCC58260.2023.10323826.