JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 14(1), MARCH, 2026
E-ISSN: 3093-0898, PRINT ISSN: 2277-0011; Journal homepage: www.atbufstejoste.com

An Explainable Hybrid LSTM-CNN Model for Phishing URL Detection

[1]Khadija Bala Gidado, [2]Nurudeen Mahmud Ibrahim, [3]Ridwan Kolapo, [4]Prema Kirubakaran, [5]Mansir Muhammad, [6]Ahmad Salkida, [7]Faruku Umar Ambursa
[1,3&4]Department of Information Technology and Information Systems
[2]Department of Cyber Security
Faculty of Computing, Nile University of Nigeria, Abuja
[5&6]HumAngle Media Limited, Abuja
[7]Department of Information Technology, Bayero University Kano

ABSTRACT
Phishing attacks are one of the most common forms of cybercrime that exists, it uses social engineering techniques that are advanced and character-level obfuscation in order to avoid the traditional detection techniques. As much as deep learning approaches have boosted phishing detection, its application remains limited due to two main challenges: vulnerability of models to evolving evasion tactics and their lack of interpretability in model decisions. Addressing these limitations is crucial for developing reliable phishing detection system suitable for real-world cybersecurity operations. This paper proposes an explainable hybrid LSTM-CNN model for phishing URL detection. The model was designed to learn both local and sequential patterns in URLs, with the SHAP (Shapley Additive Explanations) framework integrated to ensure explanations for classification decisions were interpretable. The model displayed an excellent performance having overall accuracy of 98.09%, and low false-positive rate of 0.72%. The model used a large dataset of 549,346 URLs with an Accuracy of 98.09%, Precision of 98.14%, Recall of 95.10%, F1-Score of 96.59% and ROC-AUC of 99.72%. The SHAP aspect showed how the model could identify phishing indicators like random character sequences, suspicious top-level domains that are unusual.

ARTICLE INFO
Article History
Received: August, 2025
Received in revised form: December, 2025
Accepted: January, 2026
Published online: March, 2026

KEYWORDS
Deep Learning, Explainable AI, LSTM-CNN, Phishing URL Detection, SHAP

INTRODUCTION

Phishing attacks are one of the most common types of cybercrime. The Anti-Phishing Working Group (APWG, 2024), documented that phishing attacks got to a very high level, with almost 5 million incidents reported in 2023. The attack works by tricking users to disclose sensitive information (login credentials and financial data, etc) (Aljofey et al., 2020) . The complexity of these attacks has grown to complex schemes.

Systems for phishing detection depend on rule-based approaches and traditional machine learning methods that use features like URL length (Sahingoz et al., 2020). As effective as they are against familiar attack patterns, these methods have shown some down sides when approached with adaptable phishing techniques. Although traditional machine learning approaches are more adaptive, they rely on feature engineering, which may not be able to capture subtle patterns in phishing attacks (Faizal et al., 2022).

The emergence of Deep Learning has changed some particular areas of cybersecurity by giving the chance for automated feature extraction abilities. Convolutional Neural Networks (CNN) have shown impressive performance in their ability to recognize images by learning spatial patterns, the Long Short-Term Memory (LSTM) network works well at recognizing sequential dependencies in temporal data. Deep learning models can recognize subtle changes like homoglyphs and typo squatting which traditional models often miss (Li et al., 2025).

Corresponding author: Khadija Bala Gidado
gidadokhadija59@gmail.com
Department of Information Technology and Information Systems, Faculty of Computing, Nile University of Nigeria, Abuja.

Despite this improvement, issues still remain in the deployment of deep learning-based phishing detection systems in operational environments. Its drawback is the black-box nature of complex neural networks, which limits interpretability. Some works have looked at the problems that are tied to phishing detection, but in seclusion. (Faizal et al., 2024) and (Li et al., 2025) showed how effective character-level deep learning models are. (Shendkar et al., 2024) used Explainable AI (XAI) to help improve model transparency for security applications. Even with this progress, there is a void in research with regard to having an integrated approach that encompasses resilient evasion model design with explainability. Addressing this void will give a chance to create an extensive framework that not only identifies advanced phishing attempts but will provide actionable insights for security personnel.

RELATED WORKS

This section reviews the applicable literature on usage of deep learning models for phishing detection, explainable AI, and adversarial robustness, placing the current work as an integration of these areas.

Alsabri & Al-Hadi, (2025) proposed a CNN-BLSTM model, the model achieved 81% accuracy on 50,000 URLs by using features that are structural and lexical, which demonstrated the strength of hybrid models in the learning of sequential patterns. Also, ensemble approaches have been explored for high-accuracy classification. Zara et al., (2024) did a comparative analysis of machine learning, deep learning, and ensemble models, and found that Random Forest achieved 99% accuracy on a dataset of 11,055 websites. As much as this is powerful, the models usually function as black boxes, which offer limited insights and makes it a critical shortcoming for security applications.

Deep learning models for analysis of phishing URL have grown rapidly, with comparative studies highlighting architectural strengths and limitations. Atanda et al., (2025) did a comparative analysis of RNN and CNN models for URL-based phishing detection, RNN (LSTM) achieved 91.56% accuracy, which outperformed CNN (84.56%), specifically in recall. This shows RNN's strength in processing sequential dependencies in URLs, which is a primary insight for model selection. In another approach, (Kulkarni, 2023) transformed URL feature vectors into grayscale images and applied a simplified DCNN, which achieved 85.47% accuracy. Although it is creative, the method depended on manual feature engineering and used a small dataset, which limited scalability and application in real-world.

In order to address model ambiguity, the field of explainable AI (XAI) has developed techniques which are necessary for auditing, and analysis. Recent studies have shown the useful application of SHAP (Shapley Additive exPlanations) to improve explainability in phishing detection. (Gharkan, 2025) applied SHAP to compare some ML models, K-NN achieved 99.1% accuracy, and SHAP was used to distinguish top features like redirect patterns that are suspicious and also external links, which provide comprehensible global explanations. Similarly, Shaurya & Vaghela, (2023) also applied SHAP to compare ANN model and Random Forest for classification of URL, which they achieved 95.93% and 96.59% accuracy.

Warnecke et al., (2020) evaluated XAI methodologies by using descriptive accuracy and robustness, and concluded that Integrated Gradients and Layer-wise Relevance Propagation are the most capable for security tasks, but also illustrated effectiveness is highly task-dependent, which shows that there is a need for a tailored explainable AI (XAI) in phishing URL detection.

At the same time, the increase of evasion techniques has made hostile robustness a crucial concern. Research has shown vulnerability of deep models to crafted inputs. In the aspect of network security, (Li et al., 2025) assessed multi-view deep learning models for intrusion detection against FGSM attacks, and found that feature diversity enhanced robustness, which pointed out the relevance of architectural choices and adversarial training. Though, this research has not really been applied to phishing URL detection, and almost not in connection with explainable models.

Corresponding author: Khadija Bala Gidado
gidadokhadija59@gmail.com
Department of Information Technology and Information Systems, Faculty of Computing, Nile University of Nigeria, Abuja.

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 14(1), MARCH, 2026
E-ISSN: 3093-0898, PRINT ISSN: 2277-0011; Journal homepage: www.atbufstejoste.com

Omar et al., (2023) show that machine learning models like LightGBM can attain a high accuracy of (99.7%) in detecting malicious URLs. However, they underscore the necessity of retraining with current data to sustain performance, analogous to the human requirement for continuous training and updates. In a systematic review conducted by Jampen et al., (2020) . They found that psychological traits like impulsivity and inattention, as well as email habits, are better predictors of susceptibility than demographics. The study supports the idea that psychological factors are important. They support a mixed training method that is interesting, happens again and again, and uses psychological principles to help people who are weak in their thinking.

METHODOLOGY

The dataset was acquired from Kaggle repository (phishing_site_urls.csv). The dataset had a total of 549,346 URL samples. 392,924 Legitimate URL samples (71.5%) and 156,422 Phishing URL samples (28.5%).

Data Preprocessing

1. Label Encoding: Translated from text to numeric with the use of binary mapping. Legitimate (0), Phishing (1)
2. Data Splitting: 80% for training, 20% for testing. 439,476 training samples and 109,870 test samples
3. Character-Level tokenization: Phishing campaigns usually utilise character-level manipulations, hence a character-level tokenization was selected instead of word-level based methods. A tokenizer was generated for URL corpus so as to create an index that links to each unique character.

Model Architecture

1. Hybrid LSTM-CNN architecture
2. Components:

Embedding layer: Character indices were transformed into dense vector representations of 128 dimensions. The approach let the model to learn the relationships between characters during the training, which resulted in semantically similar characters developing similar vector representations.

CNN layer: Two consecutive CNN blocks were designed to extract localised n-gram-like features in URLs. The first block utilised 128 filters with a kernel size of 3, succeeded by batch normalisation and max pooling with a pool size of 2. In the second block, the number of filters was increased to 256 while keeping the same size of kernel, followed by batch normalisation and pooling. The layers impressively recognised fishy tokens, duplicate path segments, and domain patterns.

LSTM layer: Processed token sequences in both forward and reverse directions using 64 units in each direction (128 total). This captured contextual dependencies across URL segments, allowing the model to understand relationships between domain names and path structures regardless of position.

Dense layer: Two fully connected layers (64 and 32 units respectively) with ReLU activation transformed the learned representations into a classification space through progressive dimensionality reduction.

Explainability with SHAP

1. Kernel SHAP was selected because it provides model-independent interpretability and its compatibility with black-box neural architectures.
2. Background samples of 100 URLs of which 50 were phishing and 50 were legitimate. This was selected in order to indicate a reference classification for the Shapley computation.
3. Explanation samples of 10 URLs of which 5 were phishing and 5 were legitimate. The 10 URLs were selected to put together a comprehensive per-character feature score.

Corresponding author: Khadija Bala Gidado
gidadokhadija59@gmail.com
Department of Information Technology and Information Systems, Faculty of Computing, Nile University of Nigeria, Abuja.

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 14(1), MARCH, 2026
E-ISSN: 3093-0898, PRINT ISSN: 2277-0011; Journal homepage: www.atbufstejoste.com

Model Training and Evaluation

1. **Model Training:** The model was trained using supervised learning concentrating on binary cross-entropy as the objective function. Adam optimizer was opted for to handle gradient-based optimization because of its adaptive learning rate behavior and offers efficient convergence properties. Training was conducted over multiple epochs with mini-batch processing. A hold-out validation subset comprising of 10% of training data (approximately 43,948 samples) was established from the training partition to track the model's generalisation throughout the training process.
2. **Model evaluation:** Accuracy, precision, recall, F1-score, ROC-AUC, confusion matrix.

System Requirements and Libraries
Development Components:

1. Languages/Frameworks: Python, TensorFlow 2.x with Keras API, SHAP 0.50.0.
2. Environment: Google Colaboratory

Key Libraries:

1. Pandas
2. Numpy
3. Scikit-learn
4. Matplotlib
5. Seaborn

RESULTS AND DISCUSSION

For the hybrid LSTM-CNN model developed in this study, key metrics were used to quantify the performance.

Accuracy: Accuracy is the measure of URLs that were classified correctly out of all the predictions made by the model.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Precision: Precision is the measure of phishing URLs correctly identified out of the URLs that were classified as phishing.

$$Precision = \frac{\text{TP}}{\text{TP+FP}} \quad (2)$$

Recall (Sensitivity): Recall evaluates how a model can effectively identify actual phishing URLs. $Recall = \frac{\text{TP}}{\text{TP+FN}}$ (3)

F1-Score: The F1-score is harmonic mean of precision and recall.

$$F1 - Score: 2 \text{ X } \frac{\text{Precision X Recall}}{\text{Precision+Recall}} \quad (4)$$

ROC-AUC: The area under the ROC curve indicates the ability to distinguish between phishing URLs and legitimate URLs.

ANALYSIS

The evaluation results shown indicate that the hybrid LSTM-CNN model performs very well in identifying phishing and legitimate URLs. The model showed very good performance on key metrics, 98.09% overall accuracy, a high ROC-AUC of 0.9972 and a low false-positive rate of 0.72%. The integration of the SHAP framework (using KernelExplainer) allowed post-hoc explainability, and it also showed how the model demonstrates patterns in phishing and legitimate URLs. The SHAP analysis pointed out the model's ability to recognize character-level irregularities like random hexadecimal sequences. This indicates that the model can efficiently identify typosquatting, via its CNN-based feature extraction.

Character-level tokenization was used for gathering of URLs, this resulted in the following distribution attributes based on their lengths:

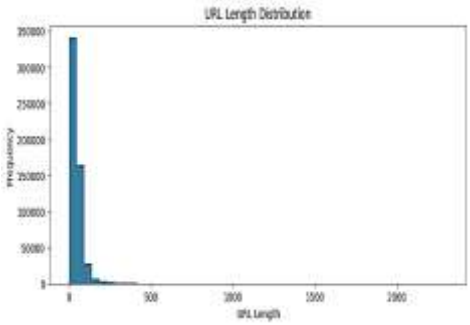

Figure 1: URL Length Statistics

Based on this analysis, the maximum sequence length was set to 200 characters, which 95% of URLs were captured without any

Corresponding author: Khadija Bala Gidado
gidadokhadija59@gmail.com
Department of Information Technology and Information Systems, Faculty of Computing, Nile University of Nigeria, Abuja.

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 14(1), MARCH, 2026
E-ISSN: 3093-0898, PRINT ISSN: 2277-0011; Journal homepage: www.atbufstejoste.com

truncation all while keeping the computational efficiency in check.

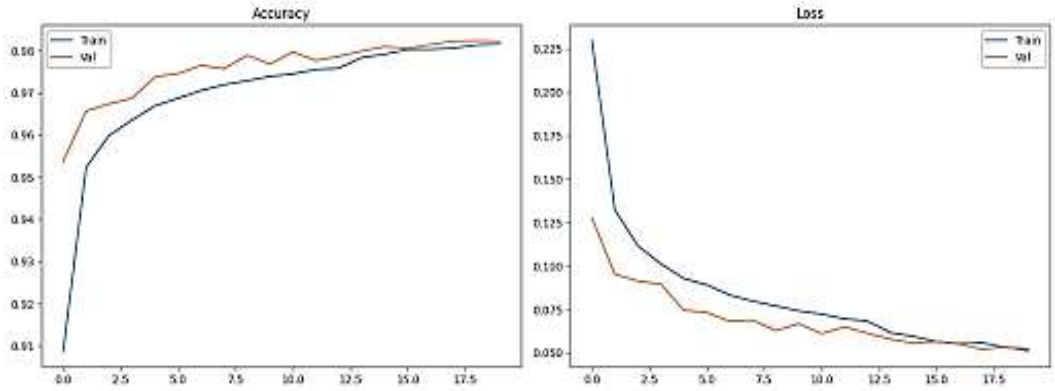

Figure 2: Graph showing a 0.3 normalisation in models learning dynamics

There was a 0.3 difference in accuracy between the training and validation sets which was minimal. This proves that using dropout layers and batch normalization was efficient in avoiding overfitting.

Table 1: Performance Metrics Result

Hybrid LSTM-CNN Model	
Accuracy	98.09%
Precision	98.14%
Recall	95.10%
F1-Score	96.59%
ROC-AUC	99.72%

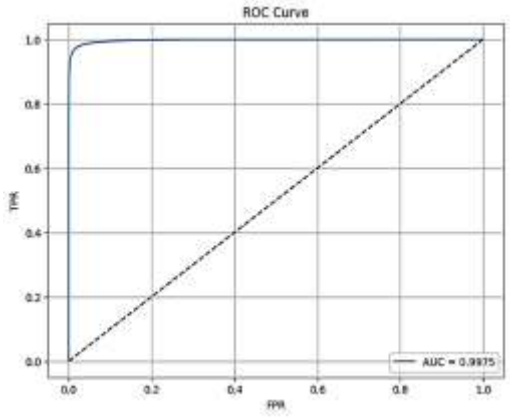

Figure 3: ROC Curve

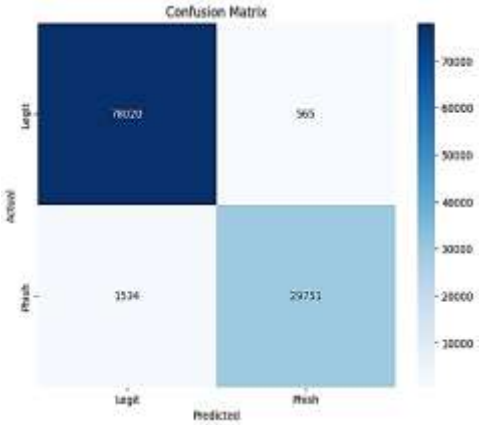

Figure 4: Confusion Matrix

Confusion Matrix Analysis:

1. True Negatives (TN) = 78,020 illustrates Legitimate URLs that were correctly identified.
2. True Positives (TP) = 29,751 illustrates Phishing URLs that were correctly identified.
3. False Positives (FP) = 565 illustrates Legitimate URLs incorrectly flagged as phishing
4. False Negatives (FN) = 1,534 illustrates Phishing URLs incorrectly classified as legitimate.

Corresponding author: Khadija Bala Gidado
gidadokhadija59@gmail.com
Department of Information Technology and Information Systems, Faculty of Computing, Nile University of Nigeria, Abuja.

Table 2: Phishing URL Explanations: SHAP Analysis for Phishing URL samples

Sample	URL Pattern	Prediction
1	https://9d345009-a-62cb3a1a-s-sites.googlegroups.com/.../login.html	100%
2	ocbc.org.au/wp-content/uploads/.../hotmail.html	99.97%

Key Findings for Phishing Indicators:

1. **Random Character Sequences:** In the model high positive SHAP values were fixed in the positions that contained random hexadecimal or alphanumeric strings. Sample 1 shows this sequence `9d345009-a-62cb3a1a` where a strong phishing alert was generated, characters like '3' which is situated at position 15 with a SHAP value of +0.064 and '0' which is situated at position 6 with a SHAP value of +0.058, specifically contributed positively in classifying it as phishing.
2. **Compromised Site Patterns:** The path in sample 2 `/wp-content/uploads/` and the file `hotmail.html` increased phishing signals, showing the model's capacity to recognize mismatches linking the domain and the content it hosts.

Table 3: Phishing URL Explanations: SHAP Analysis for Legitimate URL samples

Sample	URL Pattern	Prediction
1	https://gigaom.com/cleantech/brightsource-energys-s-1-by-the-numbers/	0.04%
2	https://manta.com/ic/mtqqz1v/ca/ecole-nationale-d-administration-publique	0.04%

Key Findings for Legitimate Indicators

1. **Trusted Domain Patterns:** Legitimate URLs indicated strong negative SHAP values at the positions where domain separators are found (.), this shows a major point for classifying as legitimate. The "." in the URL gigaom.com has a SHAP value of -0.171, this is why it was classified as a legitimate site.
2. In Sample 2, there is a clear domain separator (/) and also a readable path structure with a .com TLD.

Table 5: Comparison with Existing Works

Model	Accuracy	Precision	Recall	F1-Score
LSTM-CNN Model (This study)	98.09%	98.14%	95.59%	96.59%
(Alsabri & Al-Hadi, 2025) CNN-BLSTM	80%	81%	80%	80%
(Atanda et al., 2025) CNN model	84.56%	95.43%	72.19%	82.20%
(Atanda et al., 2025) RNN model	91.56%	93.17%	89.46%	91.28%

Table 6: Character-level vs Word-level

Metric	Character-Level	Word-Level
Accuracy	98.09%	96.77%
Precision	98.14%	95.85%
Recall	95.10%	92.65%
F1-Score	96.59%	94.22%
ROC-AUC	99.72%	99.35%
Latency (ms/URL)	0.14	79.88

Corresponding author: Khadija Bala Gidado
gidadokhadija59@gmail.com
Department of Information Technology and Information Systems, Faculty of Computing, Nile University of Nigeria, Abuja.

The above table shows the tokenization comparison where the character-level outperforms the word-level on classification metrics. The character-level had an inference latency of 0.14 ms/URL which supports real-time deployment while the word-level's latency was ≈79.88 ms/URL and is prohibitive for high-throughput scenarios. Character-level tokenization is suggested where priorities are sensitivity to character-level obfuscation (typosquatting and homoglyphs) and detection speed.

CONCLUSION AND FUTURE WORK

This work successfully developed a hybrid LSTM-CNN model integrated with SHAP framework. This research concludes that by integrating explainability with high-performance deep learning techniques in a phishing detection model gives excellent defensive outcomes. The hybrid LSTM–CNN model demonstrated high predictive performance with (98.09% accuracy and 99.72% ROC-AUC) it also maintained a real-time latency of (0.14 ms per URL).

The model also provided an interpretable explanation through SHAP. Another aspect of this study is the character-level perspective, which showed a potential for detecting fine-grained character phishing, as evidenced by SHAP attribution to specific character positions in phishing URLs. Overall, these findings address a gap in research and also paves way for a more reliable and transparent deep learning systems in cybersecurity applications. Future endeavor may explore: Analysing the system in real life corporate context, data that is accurate can be collected under real traffic conditions. Long-term retraining of model and analysing how the techniques change with time.

REFERENCES

APWG. (2024, April). *APWG Q4 Report Finds 2023 Was Record Year for Phishing*. APWG. Retrieved January 20, 2026, from https://s29837.pcdn.co/apwg-q4-report-finds-2023-was-record-year-for-phishing/

Aljofey, A., Jiang, Q., Qiang Qu, Huang, M., & Niyigena, J.-P. (2020, September 15). AnEffective Phishing Detection Model Based on Character Level Convolutional Neural Network from URL. *9*(9), 24. https://doi.org/10.3390/electronics9091514

Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2020, March). Machine learning based phishing detection from URLs. *Expert Systems with Applications*, *117*, 345 - 357. https://doi.org/10.1016/j.eswa.2018.09.029

Faizal, D. (2024). *Enhancing Phishing Threat Detection and Resilience: Leveraging Machine Learning, AI, and User Education in Cybersecurity*. ResearchGate. https://www.researchgate.net/publication/384367000

Li, M., Qiao, Y., & Lee, B. (2025, October 7th). Adversarial Robustness Evaluation for Multi-View Deep Learning Cybersecurity Anomaly Detection. *17*(10), 22. https://doi.org/10.3390/fi17100459

Shendkar, B. D., Chandre, P. R., Madachane, S. S., Kulkarni, N., & Deshmukh, S. (2024). Enhancing Phishing Attack Detection Using Explainable AI: Trends and Innovations. *ASEAN Journal on Science and Technology for Development (AJSTD)*, *42*(1). https://doi.org/10.61931/2224-9028.1604

Alsabri, A. A., & Al-Hadi, M. A. (2025). A Hybrid CNN-BLSTM Model for Phishing Attack Detection Using Deep Learning to Strengthen Internet Security. *Sana'a Univeristy Journal of Applied Sciences and Technology*, *3*(4), 964 - 972. https://doi.org/10.59628/jast.v3i4.1822

Zara, U., Ayyub, K., Khan, H. U., Daud, A., Alsahfi, T., & Ahmad, S. G. (2024, October 25th). *IEEE Access*, *12*.

Corresponding author: Khadija Bala Gidado
gidadokhadija59@gmail.com
Department of Information Technology and Information Systems, Faculty of Computing, Nile University of Nigeria, Abuja.

https://ieeexplore.ieee.org/abstract/document/10735206

Atanda, O. G., Amoyedo, F. E., Olowe, O. T., & Jimoh, E. R. (2025). Deep Learning for Phishing URL Detection: A Comparative Analysis of CNN and RNN Models for Enhanced Cybersecurity. *NIPES - Journal of Science and Technology Research*, *7*, 1121 - 1129.

Kulkarni, A. D. (2023). Convolution Neural Networks for Phishing Detection. *Computer Science Faculty Publications and Presentations*, *23*. http://hdl.handle.net/10950/4224?utm_source=scholarworks.uttyler.edu%2Fcompsci_fac%2F23&utm_medium=PDF&utm_campaign=PDFCoverPages

Gharkan, D. K. (2025, December 30). Performance and Explainability of Machine Learning Models in Phishing Detection Using SHAP. *Al-Mustansiriyah Journal of Science*, *36*(4). https://doi.org/10.23851/mjs.v36i4.1707

Shaurya, & Vaghela, R. S. (2023, December). Exploring feature importance in phishing URL detection models. *NFSU - Journal of Cyber Security and Digital Forensics*, *2*(2). https://jcsdf.nfsu.ac.in/

Warnecke, A., Arp, D., Wressnegger, C., & Rieck, K. (2020). *Evaluating Explanation Methods for Deep Learning in Security*. 2020 IEEE European Symposium on Security and Privacy (EuroS&P). https://ieeexplore.ieee.org/abstract/document/9230374

Omar, A., Tale, S., & Shaheen, M. (2023). From Phishing Behavior Analysis and Feature Selection to Enhance Prediction Rate in Phishing Detection. *(IJACSA) International Journal of Advanced Computer Science and Applications*, *14*(5). https://pdfs.semanticscholar.org/9414/b6ceba2d159b7e58b90ab7c428fb8796af19.pdf

Jampen, D., Gur, G., Sutter, T., & Tellenbach, B. (2020, December 1). Don't click: towards an effective anti-phishing training. A comparative literature review. *Human-centric Computing and Information Sciences*, *10*(1), 41. 10.1186/s13673-020-00237-7.

Corresponding author: Khadija Bala Gidado
gidadokhadija59@gmail.com
Department of Information Technology and Information Systems, Faculty of Computing, Nile University of Nigeria, Abuja.