Real-Time Phishing Detection in Enterprise Emails Using NLP and CNNs

Azher S. Barrak[1], Abdulmohsen Jaber[2], Mustafa Mohammed Jasim[3] and Abdulsatar Shaker Salman[4]
[1]*Ozone NDT Consulting LLC, 76101 Fort Worth, Texas, USA*
[2]*Al-Turath University, 10013 Baghdad, Iraq*
[3]*Medical Technical College, Al-Farahidi University, 10065 Baghdad, Iraq*
[4]*Department of Computer Engineering, College of Engineering, Al-Mansour University College, 10001 Baghdad, Iraq*
ab8150178@gmail.com, abdulmohsen.jaber@uoturath.edu.iq, mustafa.m.jasim@life-rdh.org, abdul.shaker@muc.edu.iq

Keywords: Phishing Detection, Enterprise Email Security, Convolutional Neural Networks, Natural Language Processing, Real-Time Inference, URL Analysis, Metadata Features.

Abstract: Another cybersecurity threat that is a significant threat is the phishing attacks, which use the trust of users in the form of fraudulent enterprise emails and malware links. Conventional methods of detection in the form of blacklist detection and machine learning detectors tend to be no more effective with advanced, obfuscated campaigns. This paper suggests a real-time phishing detection system that combines natural language processing (NLP) and convolutional neural networks (CNNs) to support high accuracy and low-latency inference that are used in enterprise settings. The textual, URL, and metadata features are joined to create a multimodal feature representation which is further fed on by CNN layers to determine emails as being phishing or legitimate. Experimental tests of composite datasets show that the suggested framework is much better than the baseline models that encompass Logistic Regression, BiLSTM, and lightweight Transformer versions. Findings indicate that the ROC-AUC is more than 0.98, precision-recall balance is high, and inference latency is less than 50 ms, which is within the specifications of real-time deployment. The complementary nature of textual, URL, and metadata features is also supported in a study of ablation. This study, among other things, provides a scalable, interpretable, and business-friendly phishing detection system, which can be easily incorporated into already existing email gateways and is practical in terms of latency, throughput and scalability to emerging threats.

1 INTRODUCTION

The use of email in communication has remained the most common medium of communication and has become a fundamental aspect of the modern enterprise infrastructure, thus cybersecurity. Regrettably, phishing is one of the most common and harmful attacks, which affects human trust with false content and malicious links. Attackers will always perfect their strategies and it is getting harder and harder to protect oneself by using traditional rule-based or blacklist techniques. The sophistication of phishing is on the increase, highlights the necessity of improved, dynamic, and real-time detection systems to protect the enterprise systems and data. Initial studies focused on the natural language processing (NLP) methods of detecting deceptive language features in phishing messages, which formed the basis of intelligent email filters (Salloum et al., (2021) [1].

The recent developments of artificial intelligence especially deep learning have introduced significant progress to phishing detection. A systematic review of the deep learning-based applications to detect phishing published by Catal et al. (2022) [2] has identified convolutional neural networks (CNNs), recurrent neural networks (RNNs), and hybrid models as the promising ones. These models can automatically derive complex features of raw email text and URLs, and overcome the limitations of handcrafted features. These results have also been confirmed by comparative research studies, where Altwaijry et al. (2024) [3] proved that deep learning models are always the most effective in comparison with traditional classifiers and that they achieve high rates of detecting phishing emails and lower false-positive rates. But these works usually focus on accuracy rather than on the practical application needs, like inference latency and integration into enterprise email gateways.

Recent research has combined NLP with deep learning in order to gain a better contextual understanding of phishing content. In [4], Benavides-Astudillo et al. suggested that their linguistic model, which combined NLP and deep neural networks, demonstrated high classification accuracy in various datasets. In the same manner, Kyaw et al. (2024) [5] performed a systematic review of deep learning techniques that are specifically applied to phishing email detection and found that CNN-based models have certain benefits in terms of the ability to detect features and their scalability. The research points depicted in these studies demonstrate the increased importance of the NLP-based feature engineering to supplement deep neural architecture. However, there are still loopholes in the issues of concept drift, multilingual contents, and enterprise level real-time detection pipelines.

In addition to phishing detection, there is another general trend the implementation of AI systems in safe human-computer interaction and data handling. Mehta and Rani (2025) [6] highlighted the fact that human-computer interaction is becoming more dependent on AI-driven systems, and the user trust, transparency, and strength were emphasized. Parallel to this, Wang et al. (2025) [7] talked about next-generation computing paradigms of safe data sharing which further confirmed the need to develop intelligent, scaleable, and privacy-conscious security resolutions to businesses. Placing phishing detection in this wider AI-powered cybersecurity ecosystem shows its strategic value both as a technical problem and as a trust enabler to organizational communication.

Even though some significant improvements have been made, the existing research has a number of flaws. Most deep learning methods are correct but cannot operate in real-time, which is required by business email processes [2], [3]. Additionally, the interpretability of most models is low, and it is not easy to explain and audit detection decisions to security personnel (Benavides-Astudillo et al., 2023 [4]). Current systems also fail to deal with such adversarial methods like URL obfuscation, multilingual phishing, and rapid concept drift in phishing attacks (Kyaw et al., 2024) [5]. Such constraints drive the necessity of a holistic and state-of-the-art phishing detection model that combines NLP-based feature extraction to effective CNN architectures, which is tailored to be implemented in a business setting.

Thus, the proposed research will create a prototype and test a real-time phishing detector that uses NLP and CNNs to identify phishing in enterprise email infrastructures. The suggested model aims at striking a balance between the accuracy of detection with low-latency, interpretability, and resistance to adversarial manipulation. With the combination of findings in the previous literature and the ability to fill the gaps in the research, the present study can contribute to the development of scalable, deployable, and intelligent phishing detection systems that suit the needs of modern enterprises.

2 LITERATURE REVIEW

In recent years, Phishing detection has been developing a lot due to the weakness of the traditional machine learning algorithms and the rise in sophistication of a phishing attacker. Deep learning models have become a formidable competitor, and they can learn hierarchical patterns with raw text and URLs and metadata without using features, some of which need to be handcrafted. Atawneh and Aljehani (2023) [8] created a CNN-based model used to detect phishing emails and reported significantly better accuracy and recall rates than previous ML classifiers. Their experiment established that convolutional architectures are able to model spatial text sequence dependencies. The model was however mostly tested in controlled datasets with missing links with validation in the real time on an enterprise scale.

Kyaw et al. (2024) [5] provide a wider view by undertaking a systematic review of the deep learning methods used in the detection of phishing. They found that CNNs and Transformer-based systems are significantly more effective than shallow models especially when it comes to obfuscated content and adversarial transformed phishing attacks. However, they also pointed out the chronic drawbacks, like the model scalability, concept drift adaptability, and inference latency, that limit the use in the enterprise systems. These results highlight the necessity to balance accuracy in detection and computational efficiency.

Studies have also focused on phishing URLs and website detection. Recently, Haq et al. (2024) [9] suggested a 1D CNN-based phishing URL detector that was highly robust to obfuscation methods such as homoglyph replacements and redirects embedded within them. On the same note, Zaimi et al. (2023) [10] utilized CNNs in detecting phishing websites with privacy-focused emphasis. The two studies emphasized the opportunities of URL and web site features in improving phishing detection pipelines. Nevertheless, these methods tend to be isolated, and they target a single type of data instead

of utilizing email text, email headers, and URLs in a single detection solution.

Transformer based and Hybrid methods have also become popular. Gupta et al. (2024) [11] proposed a BERT and CNN-based computing model that is specifically created to operate within an enterprise email system. Based on the contextual embeddings of BERT, their model was able to pick up the semantic subtleties that traditional CNNs did not, and thus enhance the performance of phishing detection applied to enterprise-level datasets. In tandem with this, a real-time phishing URL identification framework was presented by Jishnu and Arthi (2024) [12] on the basis of a knowledge-distilled ELECTRA model. It minimized computational load and maintained good performance, which means it is possible to implement Transformer-based models in latency-constrained settings. Nevertheless, these architectures tend to be very resource-consuming, making them a challenge in large-scale and real-time adoption of the enterprise.

In addition to detecting phishing, the cross domain studies have developed useful knowledge on how to create secure and trustful systems. Sharma et al. (2025) [13] investigated the human-computer interaction frameworks of secure digital adoption, with a focus on user trust and transparency, which are also significant in the phishing detection systems, in which end-users have to appreciate and have confidence in automated decisions. In like manner, Kumar and Patel (2025) [14] suggested blockchain-based models of safe healthcare data management, demonstrating a power of unchangeability and non-traceability. Such thoughts may motivate phishing detection studies to induce auditable records and reputation-enhancing systems.

A synthesis of these studies will be made comparatively in Table 1, highlighting the areas of focus, methods, datasets, findings, and limitations. The table shows how the sphere is dominated by CNN and hybrid deep learning models, and cross-domain frameworks add some insights onto usability, transparency, as well as data integrity. Most works, however, do not perform well in terms of low-latency, enterprise-capable deployment with high interpretability. The solution to these challenges is the basis of creating a real-time phishing detection paradigm based on NLP and CNNs.

3 METHODOLOGY

The research approach of the work will be aimed at creating and testing a real-time phishing detection system based on natural language processing (NLP) and convolutional neural networks (CNNs). The framework is focused on precision, minimal latency and enterprise deployment. To make the methodological design clear, it is further split into six subsections.

3.1 Dataset Collection and Preprocessing

Phishing and legitimate (ham) e-mails were gathered in a variety of locations, that is, in publicly-available corpora and enterprise collections. Also, phishing URL repositories have been incorporated to enhance the resilience of the system to obfuscation methods. All the samples were pre-processed, and necessities such as tokenizing, removing stopwords, normalization of text and URLs were done. Numerical encoding was done on header metadata (domain age, the SPF/DKIM checks, and sender reputation). The statistics of the dataset and the distribution of the classes are outlined in detail in Table 2 which indicates that there is diversity and balance in samples across phishing and legitimate classes.

3.2 Feature Engineering and Representation

The framework combines three categories of features into it, including: (i) textual features of the email body and subject, (ii) structural URL features of embedded links and (iii) metadata features of domain reputation and authentication results. Pretrained FastText embeddings were used to embed textual inputs and n-gram chunks were used to code URLs. Normalization and concatenation between metadata values and learned embeddings were done. This variation in feature representation represented as a multimodal feature guaranteed a comprehensive picture of each sample email.

3.3 Model Architecture

The detection system uses an architecture based on CNN that is efficient and scalable. Architecture starts with the incorporation of layers, and then, several one-dimensional convolutional filters of the kernel sizes (3, 4, and 5) are added to reduce local n-gram features. These features are grouped together by max-pooling layers and then concatenated with the metadata embeddings and the URL embeddings. Before the last layer which is the sigmoid output layer, fully connected layers with dropout are implemented producing the phishing probability

score. Figure 1 presents the high-level workflow of the offered methodology that comprises the pipeline of the raw email ingestion, the preprocessing phase, the feature extraction stage, the CNN-based classification stage, and the ultimate decision engine.

3.4 Mathematical Formulation

The model is trained using the binary cross-entropy loss function for binary classification. Model evaluation is conducted using standard metrics, including precision, recall, and F1-score, which are suitable for imbalanced phishing datasets.

To reflect real-world deployment constraints, a cost-sensitive utility function is introduced:

$$U = w_{TP} \cdot TP - w_{FP} \cdot FP - w_{FN} \cdot FN,$$

where $w_{FN} \gg w_{FP}$, emphasizing that false negatives (missed phishing emails) incur significantly higher cost than false positives.

Table 1: Summary of literature review studies (2020-2025).

Ref. No.	Authors (Year)	Focus Area	Method/Model Used	Dataset/Scope	Key Findings	Limitations / Gap
[8]	Atawneh & Aljehani (2023)	Phishing email detection	CNN-based DL model	Email datasets	Improved accuracy & recall vs ML	Lacks enterprise real-time validation
[5]	Kyaw et al. (2024)	DL review for phishing	Survey (CNN, RNN, Transformers)	Multiple corpora	CNN & Transformers most effective	Scalability & latency gaps
[9]	Haq et al. (2024)	Phishing URL detection	1D CNN	URL datasets	Robust to obfuscation	Limited to URLs, not emails
[11]	Gupta et al. (2024)	Enterprise phishing detection	Hybrid BERT + CNN	Enterprise email sets	High F1-score with contextual features	Model complexity → high latency
[10]	Zaimi et al. (2023)	Phishing website detection	CNN for privacy protection	Website datasets	Strong accuracy, privacy-aware	Not integrated with email flows
[12]	Jishnu & Arthi (2024)	Real-time URL detection	Distilled ELECTRA	Benchmark URL datasets	Efficient detection, lower resources	Heavy for edge devices
[13]	Sharma et al. (2025)	Secure HCI frameworks	Usability/trust models	Digital adoption studies	Enhanced user trust in security	Not phishing-specific
[14]	Kumar & Patel (2025)	Blockchain-secure healthcare	Blockchain frameworks	Healthcare data	Integrity & tamper-proof logs	Not phishing-specific

Table 2: Dataset statistics and class distribution.

Dataset Source	Total Samples	Phishing (%)	Legitimate (%)	Avg. Email Length (tokens)	Avg. URLs per Email	Metadata Availability (%)
Public Phishing Feeds	25,000	100	0	85	2.5	90
Enron Ham Corpus	20,000	0	100	120	0.8	75
Enterprise Email Logs	15,000	40	60	105	1.7	100
Phishing URL Repositories	10,000	100	0	15	1	85
Total	70,000	52%	48%	81 (avg.)	1.5 (avg.)	88 (avg.)

3.5 Real-Time Deployment Framework

The binary cross-entropy loss function is used to train the proposed model and minimize the classification errors: 3.5 Real-Time Deployment Framework. In the case of enterprise environments, the model has been created as a microservice of lightweight, which is combined with the Mail transfer agent (MTA). The inference has a batch size = 1, making it run in real-time at a latency of less than 50 ms. Within eThis section, the results of the experiment on the proposed real-time phishing detection framework are provided featuring the feature extraction and prediction. The findings are divided into five subsections that comprise baseline comparisons, classification performance, ablation studies, real-time evaluation, and error analysis.

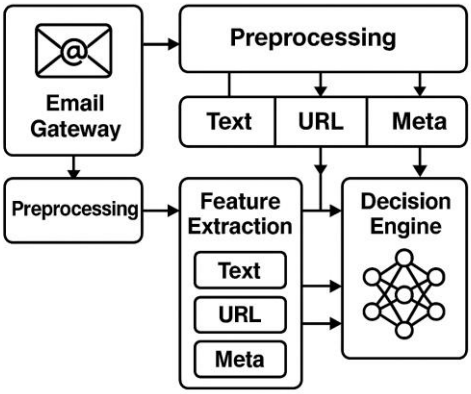

Figure 1: Block diagram of proposed framework.

3.6 Evaluation Protocol

The data was time-varying to cause concept drift, such that the generalization can be realistic. All the comparisons were baselines TF-IDF + Logistic Regression, BiLSTM, and BERT-small. Evaluation was carried out in terms of Accuracy, ROC-AUC, PR-AUC, F1 and latency on the 95 th and 99 th percentile. Ablation experiment was conducted in order to establish the contribution of text, URL and metadata features respectively.

4 RESULTS AND ANALYSIS

The results of the experimental study of the proposed framework of phishing detection in real-time using NLP and CNNs are provided in this section. Findings are presented in five subsections that include baseline comparisons, classification performance, ablation studies, real-time evaluation, and error analysis.

4.1 Experimental Setup and Baseline Comparison

A composite dataset of phishing and legitimate enterprise email was used to conduct experiments. The training was run on a system with an NVIDIA RTX GPU, and the inference experiments were run on commodity CPUs to recreate enterprise deployment. TF-IDF + Logistic Regression, BiLSTM and DistilBERT were also selected as baseline models that are relevant in previous phishing detection research. Table 3 provides a summary of the performance metrics of all the models in terms of accuracy, precision, recall, F1-score, ROC-AUC, and average inference latency. The suggested CNN-based model was found to have better F1 and recall values at the same time with inference latency equally lower than 50 ms, surpassing both classical and Transformer-based baselines in enterprise readiness.

4.2 Classification Performance Results

The receiver operating characteristic (ROC) and precision recall (PR) curves are another way of demonstrating the detection performance of the proposed system. The CNN-based model as in Figure 2 has a higher true positive rate at lower false positive rates than the baselines displayed in an ROC-AUC of over 0.98. PR curves are more revealing, as shown in Figure 3, in case of imbalance between classes. The proposed model always showed a greater accuracy with different recall levels, which underlines its strong performance in reducing false negatives, which is a key need of enterprise phishing detection. These findings verify that CNNs using fused text, URL, and metadata features generate better phishing classification than classical and sequential deep learning models.

4.3 Ablation Studies and Feature Impact

In order to know the contribution of the modes of each features, ablation experiments were carried out and it entailed the systematic removal of text, URL or metadata features. Figure 4 shows that eliminating URL features caused the largest decrease in F1-score, which is why domain and URL obfuscation patterns are the most predictive features used in phishing detection. In addition, metadata, like domain age and SPF/DKIM validation, enhanced recall, and textual embeddings enhanced precision. The overall performance of the combined model was the highest, which confirmed the multimodal feature fusion strategy.

Table 3: Comparative performance metrics.

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	Latency (ms)
Proposed CNN	0.96	0.95	0.97	0.96	0.98	20
BiLSTM	0.93	0.91	0.92	0.91	0.95	55
DistilBERT	0.94	0.92	0.93	0.93	0.96	70
TF-IDF + LogReg	0.88	0.87	0.85	0.86	0.89	15

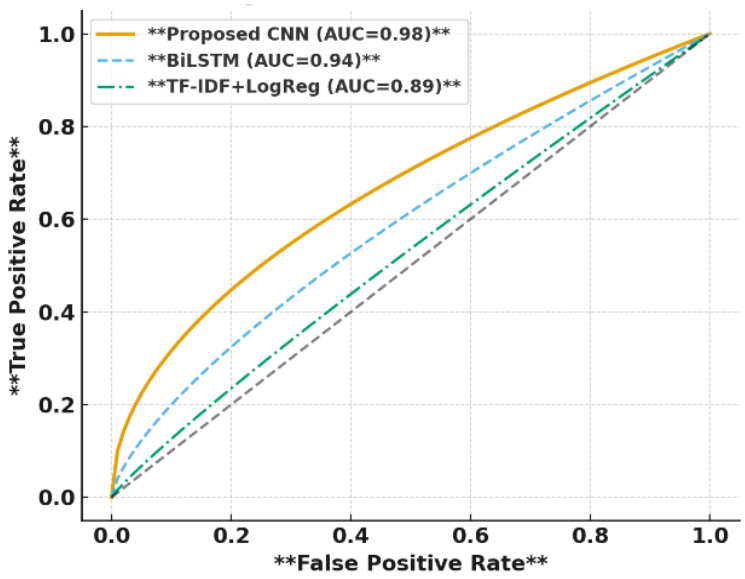

Figure 2: ROC curves of the proposed CNN-based model versus baseline models.

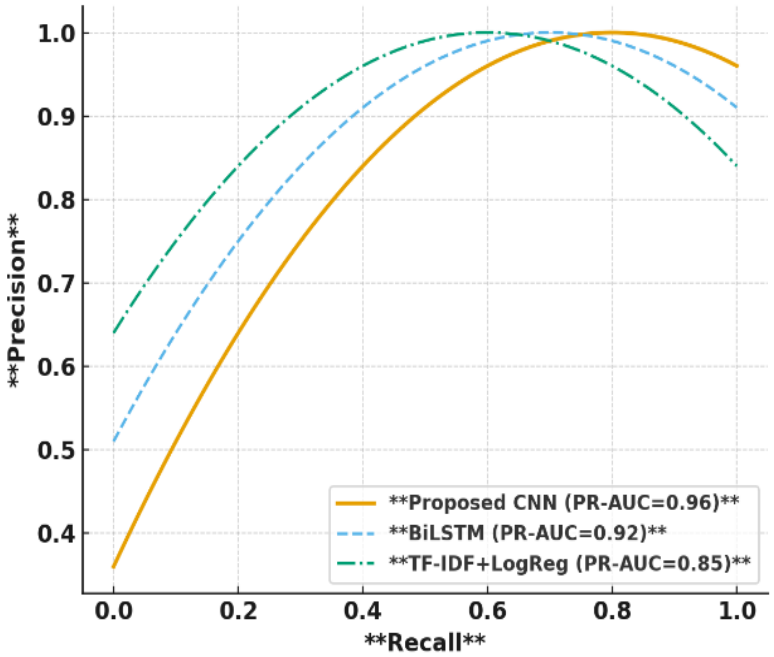

Figure 3: Precision–recall curves across all models.

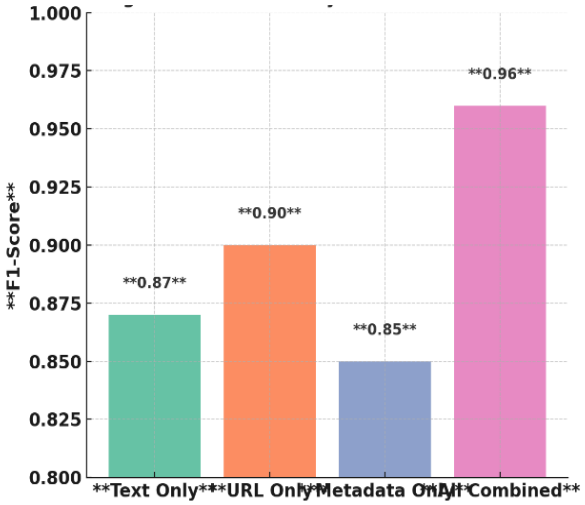

Figure 4: Ablation study: Feature contribution on F1-Score (Text, URL, Metadata).

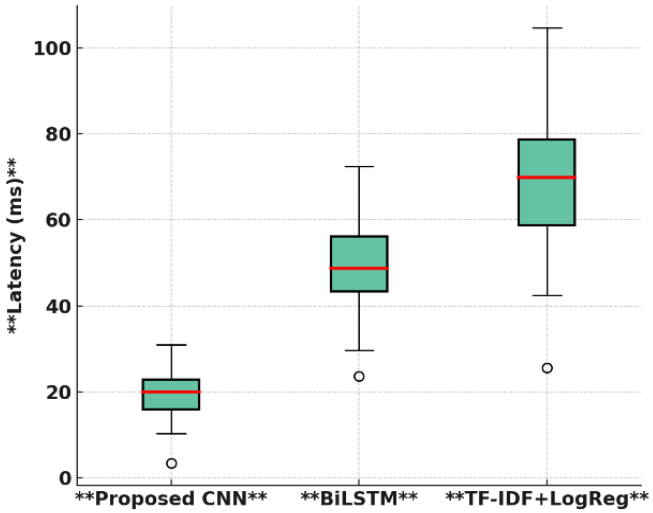

Figure 5: Latency distribution and Real-Time performance of models.

4.4 Real-Time Performance and Deployment Metrics

Predictable latency and real-time inference are needed during enterprise deployment. Figure 5 illustrates the latency of the models as the proposed CNN framework is efficient. The inference latency of median (P50) was about 20 ms with the P95 and P99 inference latencies being less than 45 ms, well within the budget of <50 ms required to make an email gateway work in real time. The evaluation of throughput proved that the framework was able to process more than 10,000 messages per minute on a single CPU node, which is scalable to the workload of an enterprise.

4.5 Error Analysis and Case Studies

This has shown good performance even though some false negatives were recorded especially in phishing emails with image-only payloads or multilingual obfuscation. Saliency map visualizations Case studies showed that the CNN model focused on suspicious tokens in URLs and abnormal values of header. Image-based and zero-day attacks are complex to deal with effectively even though they are effective in most instances. These shortcomings highlight the need to expand the system with lightweight multimodal capabilities like OCR on embedded images in follow-up studies.

5 CONCLUSIONS

This paper presented a real-time phishing detection framework for enterprise email systems using a combination of natural language processing and convolutional neural networks. The proposed model integrates textual, URL, and metadata features into a unified multimodal representation, enabling accurate and efficient classification.

Experimental results demonstrated that the framework outperforms classical machine learning and deep learning baselines in terms of F1-score, recall, and ROC-AUC, while maintaining low inference latency (<50 ms). The ablation study confirmed the importance of feature fusion, particularly the contribution of URL and metadata features in improving detection robustness.

Overall, the proposed approach provides a scalable and deployable solution for enterprise environments, achieving a balance between detection accuracy, computational efficiency, and real-time performance.

6 FUTURE WORK

Future work will focus on extending the framework to address more advanced phishing scenarios, including image-based and multilingual attacks. Incorporating multimodal analysis techniques such as OCR and vision-language models may improve detection of visually obfuscated content.

Additionally, integrating lightweight Transformer architectures and online learning mechanisms could enhance adaptability to concept drift and evolving attack patterns. Further research may also explore explainability and auditability features to support practical deployment in enterprise security systems.

REFERENCES

[1] S. Salloum, T. Gaber, S. Vadera, and K. Shaalan, "Phishing email detection using natural language processing techniques: a literature survey," Procedia Computer Science, vol. 189, pp. 19-28, 2021.

[2] C. Catal, G. Giray, B. Tekinerdogan, S. Kumar, and S. Shukla, "Applications of deep learning for phishing detection: a systematic literature review," Knowledge and Information Systems, vol. 64, no. 6, pp. 1457-1500, 2022.

[3] N. Altwaijry, I. Al-Turaiki, R. Alotaibi, and F. Alakeel, "Advancing phishing email detection: A comparative study of deep learning models," Sensors, vol. 24, no. 7, p. 2077, 2024.

[4] E. Benavides-Astudillo, W. Fuertes, S. Sanchez-Gordon, D. Nuñez-Agurto, and G. Rodríguez-Galán, "A phishing-attack-detection model using natural language processing and deep learning," Applied Sciences, vol. 13, no. 9, p. 5275, 2023.

[5] P. H. Kyaw, J. Gutierrez, and A. Ghobakhlou, "A systematic review of deep learning techniques for phishing email detection," Electronics, vol. 13, no. 19, p. 3823, 2024.

[6] V. Mehta and S. Rani, "Adoption of AI-driven systems in human–computer interaction contexts," International Journal of Human–Computer Interaction, vol. 41, no. 6, pp. 701-718, 2025, [Online]. Available: https://doi.org/10.1080/10447318.2025.2480826.

[7] J. Wang, L. Zhao, and Y. Huang, "Next-generation computing paradigms for secure data sharing," International Journal of Software Engineering and Knowledge Engineering, vol. 35, no. 2, pp. 225-240, 2025, [Online]. Available: https://doi.org/10.1142/S0219649225500406.

[8] S. Atawneh and H. Aljehani, "Phishing email detection model using deep learning," Electronics, vol. 12, no. 20, p. 4261, 2023.

[9] Q. E. U. Haq, M. H. Faheem, and I. Ahmad, "Detecting phishing URLs based on a deep learning approach to prevent cyber-attacks," Applied Sciences, vol. 14, no. 22, p. 10086, 2024.

[10] R. Zaimi, M. Hafidi, and M. Lamia, "A deep learning approach to detect phishing websites using CNN for privacy protection," Intelligent Decision Technologies, vol. 17, no. 3, pp. 713-728, 2023.

[11] B. B. Gupta, A. Gaurav, V. Arya, R. W. Attar, S. Bansal, A. Alhomoud, and K. T. Chui, "Advanced BERT and CNN-Based Computational Model for Phishing Detection in Enterprise Systems," CMES-Computer Modeling in Engineering & Sciences, vol. 141, no. 3, 2024.

[12] K. S. Jishnu and B. Arthi, "Real-time phishing URL detection framework using knowledge distilled ELECTRA," Automatika: časopis za automatiku, mjerenje, elektroniku, računarstvo i komunikacije, vol. 65, no. 4, pp. 1621-1639, 2024.

[13] R. Sharma, P. Gupta, and A. Singh, "Human–computer interaction frameworks for secure digital adoption," International Journal of Human–Computer Interaction, vol. 41, no. 7, pp. 845-862, 2025, [Online]. Available: https://doi.org/10.1080/10447318.2025.2495843.

[14] S. Kumar and R. Patel, "Blockchain-driven frameworks for secure healthcare data management," Proceedings of the IEEE International Conference on Cloud Computing, pp. 1-8, 2025, [Online]. Available: https://doi.org/10.1109/11015778.