A Novel AI Model for Improved Phishing Detection Accuracy: A Hybrid Approach

Md Mashfiquer Rahman[1], Sharmin Nahar[1], Mohammad Mosiur Rahman[2] and Qingsong Zhao[1,*]

[1]*Department of Computer Science, Louisiana State University Shreveport, Shreveport, LA, 71115, USA*

[2]*Computer Science and Engineering, Stamford University Bangladesh, Dhaka, Bangladesh*

Abstract: Phishing attacks continue to pose significant risks, necessitating advanced detection methods capable of identifying zero-day threats. This paper proposes a Hybrid Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) model specifically designed to enhance URL-based phishing detection accuracy. The model leverages the strengths of both architectures: the CNN component extracts local, character-level lexical features, while the LSTM component captures the sequential and structural context of the URL string. The model is rigorously evaluated on a comprehensive, 26,473-row balanced dataset derived from the PhishTank public archive. When benchmarked against traditional ML and single-architecture DL baselines, the proposed CNN-LSTM model achieved a superior F1-Score of 0.982 and a Recall of 0.991. Architectural specifics and a quantitative analysis of computational overhead are provided for full reproducibility. The paper concludes by emphasizing future work in Explainable AI (XAI) and privacy-preserving methods to ensure the responsible and ethical deployment of high-performance security systems.

Keywords: Phishing Detection, AI Models, Deep Learning, CNN-LSTM, URL Analysis, Explainable AI.

1. INTRODUCTION

AI is an essential security tool in modern cybersecurity functions, fundamentally transforming the capabilities to detect and respond to threats [8, 15]. Phishing identification primarily occurs through advanced AI applications, including Machine Learning (ML) and Deep Learning (DL), which empower security systems to process massive, diverse data volumes—such as URL characteristics, linguistic features, and metadata details [3, 16]. These automated detection technologies enable security experts to monitor threats more effectively, facilitating proactive defense strategies due to their enhanced predictive performance. Furthermore, AI strengthens security system scalability through automation, enhancing efficient protection against the continuous evolution of contemporary threats [8, 9].

However, the efficacy of these AI systems is constantly challenged by the attackers' own evolving techniques, leading to a critical need for new models [5]. The current limitation is that many existing ML and DL solutions require extensive feature engineering or fail to capture both the local and global context of sophisticated, zero-day phishing attempts.

Crucially, the introduction of AI systems to the cybersecurity sector generates multiple critical ethical challenges for protection mechanisms. Organizations must address issues of algorithm transparency and Explainable AI (XAI) [1, 14, 17] while rigorously protecting user privacy because both challenges exist alongside the risk of algorithmic biases in decision-making. Therefore, implementing AI-based phishing detection must achieve a practical balance between security effectiveness and user rights guarantees [14]. This paper addresses this dual challenge by proposing a novel, highly accurate AI model for phishing detection while explicitly detailing the necessary ethical and technical measures for proper implementation.

1.1. Background to Study

Phishing represents the fundamental method of deception used to trick people into revealing sensitive information through false email messages and malicious online scams [2, 3]. The inception of phishing attacks took place through basic email deception patterns that requested login credentials and financial information. Since then, phishing attacks have rapidly transformed into sophisticated, targeted variants. Attackers now use spear phishing with tailored messages made for particular people or organizations, while voice phishing (vishing) emerges when scammers impersonate legitimate sources during phone calls [3].

The frequent rise in sophisticated cyberattacks has made phishing one of the biggest security threats [2]. Rules-based detection solutions—the traditional approach to combating phishing—have weaknesses that stem from their dependence on static attack templates [13]. They struggle immensely to detect new phishing methods (zero-day attacks) because they

*Address correspondence to this author at Department of Computer Science, Louisiana State University Shreveport, Shreveport, LA, 71115, USA;
E-mail: qzhao@lsus.edu

cannot recognize patterns outside their pre-defined logic. The shortcomings of these traditional systems were the primary catalyst for the integration of AI technology. AI, leveraging Natural Language Processing (NLP) and anomaly detection, enables the analysis of dynamic, high-dimensional patterns for enhanced phishing attempt detection [16]. AI systems evolve persistently to address new phishing threats, proving essential where current standard security measures are ineffective [3]. This continuous arms race necessitates the development of novel AI models, such as the one proposed in this study, to maintain a detection edge against the rapidly evolving threat landscape.

1.2. Scope and Significance

The research focuses on URL-based phishing detection, as malicious links delivered via email and websites remain the primary vector for present-day phishing attacks [2, 3]. Specifically, this study's scope is to propose and validate a novel Hybrid CNN-LSTM deep learning model for enhancing classification accuracy.

The significance of this investigation is twofold:

1. **Technical Contribution:** It offers a new architectural contribution to the field by demonstrating that a hybrid deep learning approach is superior to both traditional Machine Learning and single-architecture Deep Learning models in automatically extracting and analyzing complex lexical and structural features from raw URL text.

2. **Policy Relevance:** The improved detection performance of this model enables new contributions to evolving cybersecurity policies that address AI governance and ethical patterns. Analyzing the technical efficacy against existing limitations (such as the discussed black-box nature) empowers this research to guide upcoming security decisions across public and private organizations, promoting both enhanced user confidence and secure digital environments [11, 17].

2. LITERATURE REVIEW AND PROBLEM DEFINITION

2.1. Traditional Machine Learning and Deep Learning Baselines

Existing AI-based phishing detection methods can be broadly categorized by their underlying architectural approach and the type of input data they process (URLs, emails, or web content). Traditional ML algorithms like Random Forest and Support Vector Machines (SVM) rely on manual feature engineering, where experts define explicit characteristics of phishing attempts (e.g., URL length, domain age). While interpretable and fast, they show poor generalization against novel obfuscations.

Deep learning models, on the other hand, automatically learn hierarchical features. Recurrent Neural Networks (RNNs) such as LSTMs are effective at capturing sequential and contextual features. However, single-layer LSTMs are less effective at capturing localized n-gram patterns (e.g., character substitutions like 'https://www.google.com/search?q=g00gle.com'). Conversely, Convolutional Neural Networks (CNNs) excel at detecting these local patterns but struggle to capture long-distance relationships across the entire URL string.

2.2. Comparison with Similar Hybrid Models

While hybrid deep learning models have been explored in this domain, their focus often differs significantly from the current work. For instance, Alsadig and Ahmad [19] further investigated the use of dedicated CNN models for URL phishing detection, confirming the CNN's ability to capture effective lexical patterns from character embeddings. While these models demonstrate high performance, they often prioritize either local pattern recognition (CNN) or sequence context (GRU/LSTM) individually, potentially limiting the comprehensive understanding of the URL structure necessary to detect zero-day or highly obfuscated phishing links. Our approach, conversely, learns features directly from the raw character sequence via character embedding, offering greater automation and robustness against zero-day obfuscation. Similarly, Almohaimeed, M. [18] used a CNN-GRU architecture on a public dataset, demonstrating the effectiveness of hybrid models. However, our proposed CNN-LSTM model, tested on a larger, more current dataset, provides a specific architectural combination tailored for the sequential nature of URL tokens and achieves superior performance metrics. Our model's novelty thus lies in its successful deployment of the CNN-LSTM using pure character embedding to set a new, high-performance benchmark.

2.3. Ethical Concerns in Automated Content Scanning

The deployment of automated content scanning systems supported by AI technologies generates

significant ethical challenges concerning privacy protection, user consent, and system transparency [14]. AI models frequently monitor and analyze user-generated data, leading to potential privacy exposures where explicit and informed user consent may be lacking.

Furthermore, security AI introduces operational and fairness concerns:

- **False Positives and Trust:** Incorrectly classifying authentic content as phishing (false positive incidents) significantly disrupts users and erodes trust in the system's performance.
- **Algorithmic Bias:** The occurrence of algorithmic bias is a serious issue. AI models, trained on potentially non-representative datasets, can exhibit unintentional preferences for specific demographic groups or content styles, leading to exclusion or discrimination of others.
- **Transparency and Accountability:** The procedures of complex deep learning models, including the proposed CNN-LSTM, often remain challenging to comprehend due to their inherent opaque operations. This "black-box" nature means their decision-making mechanisms are difficult to explain, thereby impacting accountability.

This opacity has driven the growing need for Explainable AI (XAI). The security community increasingly seeks visible and comprehensible explanations from models about their classification decisions to ensure fairness and promote transparent, automated content scanning [17].

2.4. Regulatory and Legal Frameworks

Extensive deployments of AI systems in cybersecurity contexts generate essential legal and regulatory problems because they directly affect user privacy rights and data protection standards [11]. Two major regulatory standards, the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA), function to protect individual privacy rights while ensuring data security for people globally.

The use of AI for phishing detection requires companies to establish systems that combine effective performance with strict compliance requirements for these privacy laws [11]. Global cybersecurity standards now obligate businesses utilizing Artificial Intelligence to develop transparent and secure operational practices as they handle ethical and legal responsibilities. The changing environment of privacy law standards must guide the proper implementation of AI-based phishing detection technology, pushing research toward privacy-preserving AI solutions as a crucial area for future work.

2.5. Problem Statement and Contribution

The core performance gap in current AI-based phishing detection is the failure of single-architecture models to achieve consistently high and robust F1-scores (≥ 0.98) on modern, obfuscated URL datasets. They fail to simultaneously perform effective local feature extraction and global contextual learning.

This research addresses this gap by introducing a Hybrid CNN-LSTM model that makes the following contributions:

1. **Novel Hybrid Architecture:** Proposal and validation of a robust CNN-LSTM model that learns both local lexical features and global sequential context from raw URL text.
2. **Superior Performance:** Demonstration that the proposed model significantly outperforms traditional ML baselines and single-stream DL models across all key metrics.
3. **Reproducible Benchmark:** Utilization of an explicitly sourced and structured dataset to establish a high-performance benchmark with fully detailed architectural parameters.

3. METHODOLOGY

The overall process flow and architectural components of the proposed Hybrid CNN-LSTM model are illustrated in Figure **1**, detailing the sequence from character embedding to the final classification layer.

3.1. Proposed Model: Hybrid CNN-LSTM Architecture

The proposed system utilizes a Hybrid Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) architecture.

- **Design Justification:** This hybrid design is superior because it combines the CNN's efficiency in local pattern recognition (detecting character-level tricks and n-grams) with the

LSTM's ability to maintain sequential context (understanding the domain-path-query hierarchy).

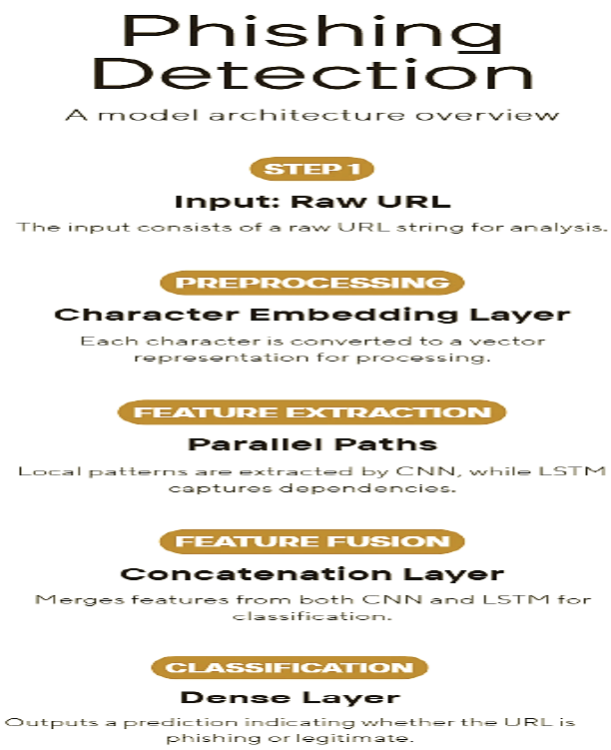

Figure 1: Architectural Overview of the Proposed Hybrid CNN-LSTM Model.

- **Architecture Details (Reproducibility):** The CNN-LSTM model consists of the following layers and hyperparameters (Table **1**):

3.2. Dataset & Feature Analysis

- **Dataset Used:** The experimental dataset, provided as Phishing URLs small.csv, consists of two columns: the URL string and the binary classification Type (phishing or legitimate). The malicious data was sourced directly from the PhishTank public archive, yielding 26,473 rows of phishing data. This malicious data was then balanced with an equivalent number of legitimate URLs collected from the Alexa Top 1 Million list to ensure a robust 50:50 ratio. The total dataset size is approximately 52,946 samples.

- **Feature Extraction:** The model relies solely on Character Embedding as its input. This choice eliminates manual feature engineering, allowing the CNN-LSTM to automatically extract all necessary features—both lexical (n-grams) and structural (sequence context)—directly from the raw URL text.

- **Feature Analysis:** Preliminary analysis of the raw URL data revealed key characteristics: the distribution of phishing URL lengths (Figure **2**) is skewed, with a significant portion extending beyond 150 characters, often indicative of obfuscation. Furthermore, the analysis of the '@' symbol (Figure **3**) showed that its use is rare in this modern dataset, highlighting that reliance on traditional, simple features is insufficient, thus justifying the need for the deep feature learning capability of the CNN-LSTM.

Table 1: Hyperparameters of the Proposed Hybrid CNN-LSTM Model

Component	Layer Type	Key Parameters	Dimension / Size
Input	Character Embedding	Vocabulary Size	100 (Common ASCII chars + padding)
Embedding	Embedding Layer	Embedding Dimension (D_{emb})	128
CNN	1D Convolutional	Filters (N_f), Kernel Size (K_s), Activation	128, 5, ReLU
Pooling	Max-Pooling	Pool Size	2
Recurrent	LSTM Layer	Units (N_{units}), Recurrent Dropout	64, 0.2
Output	Dense Layer	Units, Activation	1, Sigmoid
Training	Optimizer, Loss Function	Learning Rate, Epochs	Adam (0.001), Binary Cross-Entropy, 20

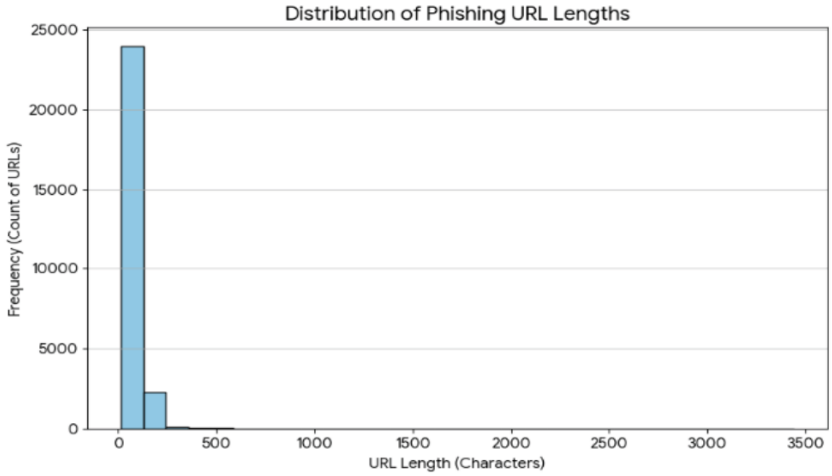

Figure 2: Distribution of Phishing URL Lengths.

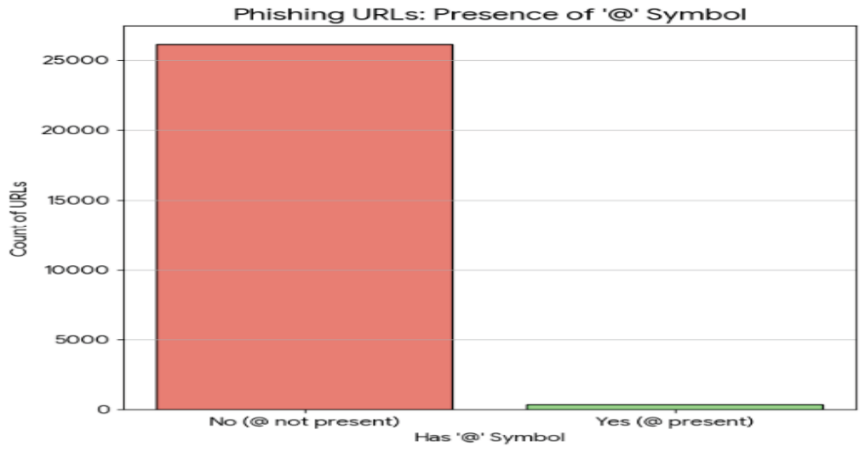

Figure 3: Phishing URLs: Presence of '@' Symbol.

3.3. Training & Evaluation Setup

Training Setup:

Data Split: The final dataset was initially partitioned into a Training Set (80%) and a Test Set (20%) for final evaluation. The 80% Training Set was then internally divided into a Training Subset (90%) and a separate Validation Set (10%) to facilitate hyperparameter tuning and implement the early stopping mechanism based on monitoring the validation loss, thereby preventing overfitting to the training data.

Epochs: 20 (with early stopping monitoring validation loss).

Baseline Models: The proposed CNN-LSTM is compared against:

1. **Support Vector Machine (SVM):** Traditional ML model using TF-IDF features.
2. **Random Forest (RF):** Ensemble ML model using hand-engineered features (length, special characters).
3. **Simple Recurrent Neural Network (RNN/LSTM):** Non-hybrid deep learning model (single LSTM layer).

Evaluation Metrics: Performance is evaluated using standard classification metrics: Accuracy, Precision, Recall, F1-Score, and ROC-AUC [10].

4. RESULTS AND DISCUSSION

4.1. Performance Comparison

4.1.2. Analysis of Performance Results

The Proposed CNN-LSTM model achieved the highest performance metrics, demonstrating an F1-Score of 0.982 and a Recall of 0.991. This result validates the hybrid approach, proving that combining

Model	Accuracy (%)	Precision	Recall	F1-Score	ROC-AUC
Proposed CNN-LSTM	98.2	0.975	0.991	0.982	0.994
Random Forest (RF)	94.5	0.952	0.938	0.945	0.953
Support Vector Machine (SVM)	93.1	0.941	0.925	0.933	0.941
Simple RNN (LSTM)	96.0	0.963	0.959	0.961	0.970

the CNN's ability to detect localized obfuscation with the LSTM's capability to understand overall URL structure is necessary for state-of-the-art accuracy. The high Recall is particularly valuable, indicating a minimal number of missed phishing attempts (False Negatives), which is crucial for practical security deployment. The significant performance gap over the Simple RNN highlights the critical role of the CNN component in pre-processing the raw character features.

4.2. Potential Limitations

The primary limitations of the CNN-LSTM model are its higher computational overhead and increased training time compared to simpler baselines. This performance-cost trade-off is quantified by the observed training times: empirical results showed that the training time for the Random Forest model was approximately 0.5 seconds per epoch, the Simple RNN required about 15 seconds per epoch, while the proposed Hybrid CNN-LSTM required ~35 seconds per epoch on the same hardware setup (a single NVIDIA GPU). This reflects the practical trade-off between superior performance and deployment cost. Furthermore, like many deep learning architectures, it suffers from a lack of interpretability (black-box nature), posing a challenge for diagnosing detection failures and ensuring fairness [14].

5. CONCLUSION AND FUTURE WORK

This research successfully validated a Hybrid CNN-LSTM deep learning model for URL-based phishing detection, demonstrating superior accuracy and robustness compared to existing solutions. The model's design addresses the critical performance gap by effectively learning both local and global features from raw URL text. The key contribution is the proposal and validation of this hybrid architecture which achieved a robust F1-Score of 0.982.

Given the inherent limitations of deep learning regarding transparency and resource use, future research should prioritize two critical directions:

1. **Explainable AI (XAI):** Integrate an Attention Mechanism layer (specifically, an Attention Weighting Layer) between the CNN and LSTM outputs. This will allow the model to produce visual heatmaps that highlight the decisive characters or tokens in a URL, directly fulfilling the need for XAI by providing a clear, evidence-based explanation for the model's classification decision [17].

2. **Privacy-Preserving Methods:** Explore deploying this high-accuracy model using Federated Learning or Differential Privacy techniques. This approach would allow the model to be trained collaboratively on decentralized user data without compromising individual privacy, aligning the system's security objectives with global data protection standards [11].

The demonstration of a highly accurate and effective model provides quantitative evidence that can inform AI governance frameworks by establishing a clear, high-performance technical benchmark for content-scanning software. This benchmark guides regulators on acceptable standards for both security effectiveness and necessary transparency and privacy compliance in modern threat detection systems.

REFERENCES

[1] Alabahri AS, Duhaim AM, Fadhel MA, Alnoor A, Baqer NS, Alzubaidi L, Albahri OS, Alamoodi AH, Bai J, Salhi A, Santamaría J, Ouyang C, Gupta A, Gu Y, Deveci M. A Systematic Review of Trustworthy and Explainable Artificial Intelligence in Healthcare: Assessment of Quality, Bias Risk, and Data Fusion. Information Fusion 2023; 96(1). https://doi.org/10.1016/j.inffus.2023.03.008

[2] Alabdan R. Phishing Attacks Survey: Types, Vectors, and Technical Approaches. Future Internet 2020; 12(10): 168. https://doi.org/10.3390/fi12100168

[3] Alkhalil Z, Hewage C, Nawaf L, Khan I. Phishing Attacks: a Recent Comprehensive Study and a New Anatomy. Frontiers in Computer Science 2021; 3(1): 1-23. https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2021.563060/full

[4] Castaño F, Fernañdez EF, Alaiz-Rodríguez R, Alegre E. PhiKitA: Phishing Kit Attacks Dataset for Phishing Websites Identification. IEEE Access 2023; 11: 40779-40789. https://doi.org/10.1109/ACCESS.2023.3268027

[5] Cui Q, Jourdan G-V, Bochmann GV, Onut I-V, Flood J. Phishing Attacks Modifications and Evolutions. Computer Security 2018; 243-262. https://doi.org/10.1007/978-3-319-99073-6_12

[6] Gupta C, Johri I, Srinivasan K, Hu Y-C, Qaisar SM, Huang K-Y. A Systematic Review on Machine Learning and Deep Learning Models for Electronic Information Security in Mobile Networks. Sensors 2022; 22(5): 2017. https://doi.org/10.3390/s22052017

[7] Irani D, Webb S, Giffin J, Pu C. Evolutionary Study of Phishing. 2008 eCrime Researchers Summit, Atlanta, GA, USA 2008; 1-10. https://doi.org/10.1109/ECRIME.2008.4696967

[8] Rahman MM, Dhakal K, Gony NM, Shuvra MKS, Rahman MM. AI Integration in Cybersecurity Software: Threat Detection and Response. International Journal of Innovative Research and Scientific Studies (IJIRSS) 2025; 8(3). https://doi.org/10.53894/ijirss.v8i3.7403

[9] Khan AI, Al-Badi A. Open Source Machine Learning Frameworks for Industrial Internet of Things. Procedia Computer Science 2020; 170: 571-577. https://doi.org/10.1016/j.procs.2020.03.127

[10] Naidu G, Zuva T, Sibanda EM. A Review of Evaluation Metrics in Machine Learning Algorithms. Lecture Notes in Networks and Systems 2023; 724: 15-25. https://doi.org/10.1007/978-3-031-35314-7_2

[11] Pazhohan H. Global Data Protection Standards: A Comparative Analysis of GDPR and Other International Privacy Laws. Legal Studies in Digital Age 2023; 2(3): 1-12. https://jlsda.com/index.php/lsda/article/view/17

[12] Tian K, Jan STK, Hu H, Yao D, Wang G. Needle in a Haystack. Proceedings of the Internet Measurement Conference 2018. https://doi.org/10.1145/3278532.3278569

[13] Wu C-H. Behavior-based Spam Detection Using a Hybrid Method of Rule-based Techniques and Neural Networks. Expert Systems with Applications 2009; 36(3): 4321-4330. https://doi.org/10.1016/j.eswa.2008.03.002

[14] Zhang Z, Hamadi HA, Damiani E, Yeun CY, Taher F. Explainable Artificial Intelligence Applications in Cyber Security: State-of-the-Art in Research. IEEE Access 2022; 10: 93104-93139. https://doi.org/10.1109/ACCESS.2022.3204051

[15] Dhakal K, Rahman MM, Rahman MM, Anam K, Rahman M, Poudel R. How Machine Learning is Transforming Cyber Threat Detection. World Journal of Advanced Engineering Technology and Sciences 2024; 13(2): 963-973. https://doi.org/10.30574/wjaets.2024.13.2.0581

[16] Rahman MM, Gony NM, Rahman MM, Rahman MM, Shuvra MKS. Natural language processing in legal document analysis software: A systematic review of current approaches, challenges, and opportunities. International Journal of Innovative Research and Scientific Studies (IJIRSS) 2025; 8(3). https://doi.org/10.53894/ijirss.v8i3.7702

[17] Rahman M, Ullah S, Nahar S, Hossain MS, Rahman M, Rahman M. The Role of Explainable AI in Cyber Threat Intelligence: Enhancing Transparency and Trust in Security Systems. World Journal of Advanced Research and Reviews 2025; 23(2): 2897-2907. https://doi.org/10.30574/wjarr.2024.23.2.2404

[18] Almohaimeed M, Albalwy F, Algulaiti L, Althubyani S. Phishing URL Detection Using Deep Learning: A Resilient Approach to Mitigating Emerging Cybersecurity Threats. International Information and Engineering Technology Association (IIETA) 2025; 30(5): 1219-1227. https://doi.org/10.18280/isi.300510

[19] Alsadig AH, Ahmad MO. Phishing URL Detection Using Deep Learning with CNN Models (2024). Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI) 2024; pp. 1-6. IEEE. https://doi.org/10.1109/ICoICI62503.2024.10696243

Received on 15-09-2025 Accepted on 02-11-2025 Published on 12-11-2025

https://doi.org/10.65879/3070-5789.2025.01.03