ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13, Issue 1, January 2026

Interpretable Deep Learning-Based Phishing Detection Using Convolutional and LSTM Networks

[1] Dewashish Kumar, [2] Amanpreet Kaur, [3] Mohit Dewangan, [4] Mahipal Singh Papola

[1] [2] [3] [4] Department of Computer Science & Engineering, Lovely Professional University, Phagwara, Punjab, India
Corresponding Author Email: [1] kalim.mulani.cs@ghrcem.raisoni.net, [2] sahil.sapate.cs@ghrcem.raisoni.net, [3] naila.tamboli.cs@ghrcem.raisoni.net, [4] pradeep.taware@raisoni.net

Abstract*— Phishing continues to be a major cybersecurity concern, as attackers increasingly rely on deceptive websites and social engineering techniques to obtain sensitive user information. Accurately identifying phishing websites remains difficult because attack patterns evolve rapidly and web-related features are often complex and interdependent. In this work, we develop an explainable deep learning approach for phishing website detection based on a hybrid Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) architecture. The model is trained and evaluated using the UCI Phishing Websites Dataset, which contains a diverse set of URL and HTML-related features. CNN layers are used to capture local structural patterns in the input features, while LSTM units model sequential relationships that are difficult to learn using conventional classifiers. Experimental evaluation shows that the proposed model achieves an accuracy of approximately 95–96%, outperforming traditional machine learning approaches such as Random Forest. To improve transparency, SHapley Additive exPlanations (SHAP) are employed to analyze model decisions and highlight the contribution of individual features. The explainability analysis indicates that factors such as URL length, redirection behavior, and the use of special characters play a key role in phishing classification. By combining strong detection performance with meaningful explanations, the proposed framework offers a practical and trustworthy solution for real-world phishing detection systems.

Index Terms*— Phishing website detection, explainable artificial intelligence, hybrid CNN–LSTM architecture, convolutional neural networks, long short-term memory networks, SHAP-based interpretability, deep learning methods, UCI phishing websites dataset, cybersecurity analytics, web security.

I. INTRODUCTION

The rapid growth of digital services, online transactions, and e-commerce platforms has transformed the way users interact with the internet. At the same time, this expansion has increased exposure to sophisticated cyber threats, among which phishing remains one of the most widespread and damaging. Phishing attacks rely on deceptive websites and impersonation techniques to mislead users into revealing sensitive information such as login credentials, financial details, or personal data. Unlike traditional cyberattacks that exploit system vulnerabilities, phishing primarily targets human behavior, making it particularly difficult to prevent using conventional security mechanisms.

Recent security reports indicate that attackers continuously adapt their phishing strategies by employing techniques such as domain impersonation, URL obfuscation, and redirection abuse. As a result, rule-based systems and signature-driven detection tools struggle to keep pace with these evolving attack patterns. To address these limitations, researchers and practitioners have increasingly turned to machine learning (ML) and deep learning (DL) approaches for phishing website detection. Traditional ML models, including Decision Trees, Support Vector Machines (SVMs), and Random Forests, have been widely used to classify websites based on handcrafted features extracted from URLs, HTML elements, and scripting behavior. While these methods are computationally efficient and interpretable, their performance often depends heavily on manual feature engineering and domain expertise, which limits their ability to generalize to newly emerging phishing techniques.

Deep learning models offer an alternative by learning feature representations directly from data in an end-to-end manner. In particular, hybrid architectures have shown promise by combining complementary learning mechanisms. Convolutional Neural Networks (CNNs) are effective at identifying localized patterns and structural relationships within input features, while Long Short-Term Memory (LSTM) networks are well suited for modeling sequential dependencies. When combined, CNN–LSTM architectures can jointly capture structural characteristics and sequential relationships present in URL patterns and web content, which are important indicators for distinguishing phishing websites from legitimate ones.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13, Issue 1, January 2026

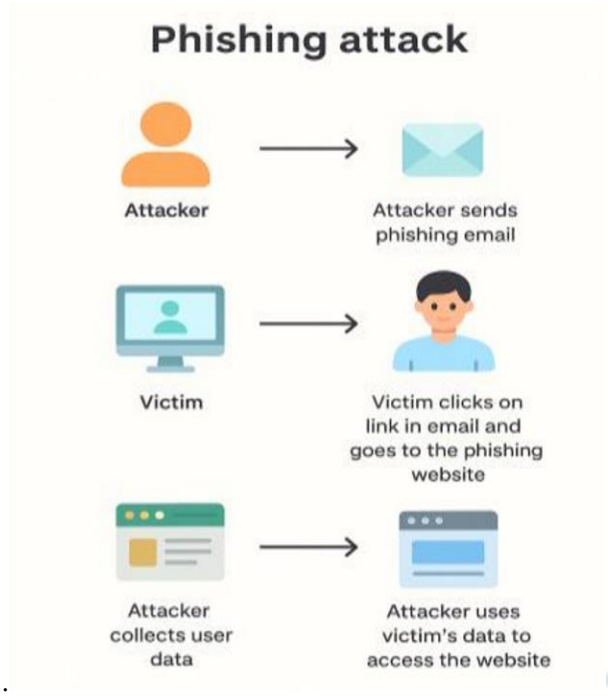

Figure 1. illustrates the typical mechanisms used in phishing attacks, where deceptive websites and interfaces are designed to manipulate users into disclosing confidential information

Despite their improved detection capability, deep learning models are often criticized for their lack of transparency. These models are frequently perceived as black boxes, providing accurate predictions without clear explanations. In cybersecurity applications, this opacity poses serious challenges, as analysts and decision-makers require understandable justifications to trust automated systems and comply with regulatory requirements. To address this issue, Explainable Artificial Intelligence (XAI) techniques have been introduced to make model behavior more transparent.

Among various XAI approaches, SHapley Additive exPlanations (SHAP) has gained significant attention due to its solid theoretical foundation in game theory. SHAP assigns a contribution value to each input feature, indicating how much it influences a particular prediction. This enables both global analysis of feature importance across the dataset and local explanations for individual decisions. In the context of phishing detection, SHAP helps identify critical factors such as suspicious URL components, redirection behavior, and structural anomalies that influence classification outcomes.

In this study, we propose an interpretable deep learning framework that integrates a hybrid CNN–LSTM architecture with SHAP-based explainability for phishing website detection. The framework is evaluated using the UCI Phishing Websites Dataset, a widely used benchmark containing URL and HTML-related features. A Random Forest model is first employed as a baseline to establish comparative performance. The proposed CNN–LSTM model is then used to learn complex feature representations, while SHAP is applied to interpret the resulting predictions and highlight influential features. *Fig. 2* presents an overview of the proposed system architecture, illustrating the integration of deep learning and explainability components. By combining high detection accuracy with meaningful explanations, the proposed approach aims to bridge the gap between performance and transparency, making it suitable for practical deployment in real-world cybersecurity environments.

Figure 2. Conceptual architecture of the hybrid CNN–LSTM model with SHAP interpretability framework

II. RELATED WORK

Phishing remains one of the most persistent threats in the cybersecurity domain, as attackers continuously create fraudulent websites that imitate legitimate services to deceive users into disclosing sensitive information such as authentication credentials, financial details, and personal data. To address this problem, researchers have proposed a wide range of detection techniques, evolving from traditional machine learning approaches to more recent deep learning and explainable artificial intelligence (XAI)–based solutions.

Early studies in phishing detection primarily employed classical machine learning models because of their simplicity and ease of interpretation. Aburrous et al. [1] introduced a feature-based phishing detection approach using Decision Trees, demonstrating reasonable performance for URL-based classification. Similarly, Aljawarneh et al. [2] evaluated Random Forest (RF) and Support Vector Machine (SVM) classifiers on the UCI Phishing Dataset and reported accuracy levels close to 93%. Mohammad et al. [3] further explored k-Nearest Neighbors (k-NN) classifiers using handcrafted URL features, achieving moderate detection results.

Despite their advantages, traditional machine learning models exhibit several limitations. These methods rely heavily on manual feature engineering, which requires domain expertise and significant effort. In addition, handcrafted features often struggle to capture complex relationships present in evolving phishing attacks. Classical ML approaches also have limited capability to model spatial and sequential patterns embedded in URL structures and

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13, Issue 1, January 2026

HTML content. These constraints have motivated the transition toward deep learning–based phishing detection methods.

Deep learning models have gained attention due to their ability to automatically learn discriminative features from high-dimensional data. Zhou et al. [4] proposed a Convolutional Neural Network (CNN)–based framework to capture spatial correlations within URL and HTML features, reporting improved performance compared to conventional machine learning techniques. Their findings highlighted the effectiveness of CNNs in extracting meaningful patterns without explicit feature engineering.

Sequential modeling techniques have also been explored in phishing detection. Long Short-Term Memory (LSTM) networks are particularly suitable for learning temporal and sequential dependencies. Hussain et al. [6] applied LSTM models to URL sequences and demonstrated that modeling sequential information leads to improved classification accuracy. Building on this idea, Sahu et al. [18] proposed a hybrid CNN–LSTM architecture in which CNN layers capture local feature patterns while LSTM layers model long-term dependencies. Their results showed that hybrid architectures outperform standalone CNN or LSTM models.

Other hybrid deep learning configurations have been investigated in related studies. Patil et al. [14] combined CNNs with Gated Recurrent Units (GRUs) for phishing email detection, while Rathore et al. [8] introduced CNN–BiLSTM architectures to enhance both feature extraction and sequence learning. Collectively, these studies indicate that hybrid deep learning models, particularly CNN–LSTM–based designs, are more effective for phishing detection than individual deep learning components.

Although deep learning approaches have improved detection accuracy, their lack of transparency remains a significant concern, especially in security-critical environments. Explainable Artificial Intelligence (XAI) has been introduced to address this challenge by enabling insight into model decision-making processes. Commonly used XAI techniques include LIME and SHAP, which provide explanations for both machine learning and deep learning models.

Lundberg and Lee [17] proposed SHapley Additive exPlanations (SHAP), a unified explanation framework based on cooperative game theory. SHAP has been applied in phishing detection studies to identify key contributing features such as URL length, special characters (e.g., “@”), and HTTPS usage [10], [20]. While some research has explored attention mechanisms and Grad-CAM for interpretability in phishing email detection, these techniques are less frequently applied to website-based phishing analysis.

In summary, traditional machine learning approaches offer interpretability but lack the robustness required to handle modern phishing attacks, while deep learning models provide improved accuracy at the cost of transparency. Existing XAI-based studies often focus on simpler models or provide limited interpretability for complex architectures. Notably, few works integrate CNN–LSTM hybrid models with SHAP-based explanations for phishing website detection. Furthermore, many studies emphasize overall accuracy without offering detailed feature-level insights, and only a limited number jointly evaluate detection performance and interpretability within a single unified framework.

III. METHODOLOGY

This section describes the methodology adopted to design, train, and evaluate the proposed explainable phishing detection framework. The overall workflow includes dataset selection, data preprocessing, baseline model construction, development of the hybrid CNN–LSTM architecture, training configuration, and integration of SHAP for model interpretability.

A. Dataset Description

The experiments are conducted using the UCI Phishing Websites Dataset, a widely accepted benchmark in phishing detection research. The dataset consists of 11,055 website instances, each described using 31 attributes extracted from URL properties, HTML structure, and scripting behavior. Each instance is assigned a binary class label indicating whether the website is legitimate (−1) or phishing (1).

Table I. Sample Dataset Structure

Website URL	Uses IP Address	URL Length	“@” Symbol	Redirect (“//”)	HTTPS	Class Label
example.com	0	23	0	0	1	Legitimate
phishingsite.tk	1	47	1	1	0	Phishing
securebank.com	0	29	0	0	1	Legitimate
malicious.ru	1	41	1	0	1	Phishing
paypal.com	0	32	0	0	1	Legitimate

The feature set captures multiple security-relevant characteristics, including URL length, use of special symbols such as “@”, redirection behavior, HTTPS availability, and domain-related indicators. The dataset maintains a relatively balanced distribution between phishing and legitimate samples, making it suitable for evaluating both predictive performance and interpretability.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13, Issue 1, January 2026

B. Data Preprocessing

Before training the models, the dataset undergoes a preprocessing phase to ensure compatibility with machine learning and deep learning algorithms. Categorical and boolean attributes are first converted into numerical form using label encoding. Continuous features are then standardized to have zero mean and unit variance, which improves numerical stability and accelerates model convergence. After preprocessing, the dataset is split into training and testing sets using an 80:20 ratio. Stratified sampling is applied to preserve the class distribution across both subsets, ensuring unbiased training and evaluation.

C. Baseline Model

A Random Forest (RF) classifier is implemented as a baseline to provide a reference for performance comparison. The RF model is configured with 200 decision trees and uses ensemble voting to determine the final classification outcome. It is trained on the preprocessed training data and evaluated on the test set. Standard performance metrics, including accuracy, precision, recall, and F1-score, are computed to assess baseline effectiveness. These results serve as a benchmark against which the proposed deep learning model is compared.

D. Proposed CNN–LSTM Architecture

The proposed approach employs a hybrid Convolutional Neural Network–Long Short-Term Memory (CNN–LSTM) architecture to capture both structural and sequential patterns within phishing-related features. Initially, a one-dimensional convolutional layer processes the input feature vector to extract localized patterns, which are useful for identifying structural anomalies in URLs and HTML attributes. The extracted representations are then passed to an LSTM layer, which models long-term dependencies and sequential relationships among features. This design enables the model to retain contextual information that may not be captured by standalone CNN architectures. To reduce overfitting and improve generalization, dropout regularization is applied, followed by fully connected dense layers. The final layer uses a sigmoid activation function to perform binary classification.

E. Model Training Configuration

The CNN–LSTM model is trained using the Adam optimizer due to its adaptive learning rate and efficient convergence properties. Binary cross-entropy is selected as the loss function, as it is well suited for binary classification problems. Training is carried out for 20 epochs with a batch size of 64, providing a balance between computational efficiency and learning stability. Model performance is evaluated on the test dataset using accuracy, precision, recall, and F1-score. In addition, training and validation accuracy and loss curves are monitored to analyze convergence behavior and detect potential overfitting.

F. Explainability Using SHAP

To enhance transparency and interpretability, SHapley Additive exPlanations (SHAP) are integrated into the proposed framework. For the Random Forest baseline, SHAP's TreeExplainer is employed to compute global feature importance scores, highlighting attributes that most strongly influence phishing predictions. For the CNN–LSTM model, SHAP-based deep learning explainers are applied to generate instance-level explanations. Visualization methods such as feature importance bar plots, beeswarm plots, and waterfall plots are used to illustrate how individual features contribute to prediction outcomes. These explanations support better understanding of model behavior, increase analyst trust, and help identify critical indicators associated with phishing activity.

IV. RESULTS AND DISCUSSION

A. Quantitative Evaluation of Model Performance

The performance of both the Random Forest baseline and the proposed CNN–LSTM model was evaluated using standard metrics, including accuracy, precision, recall, and F1-score.

Table II summarizes the comparative results obtained on the UCI Phishing Websites Dataset. The hybrid CNN–LSTM model achieves an overall accuracy of 96%, outperforming the Random Forest baseline by 2 percentage points. Improvements in recall and F1-score indicate that the model is more sensitive to phishing patterns while maintaining reliable classification of legitimate websites. The enhanced performance can be attributed to the CNN's capacity to capture localized feature dependencies and the LSTM's ability to model sequential relationships within URL and HTML-based attributes.

Table II. Comparative Performance Analysis

Model	Dataset	Accuracy (%)
Decision Tree	UCI Phishing Dataset	90.0
k-NN	UCI Phishing Dataset	91.0
Random Forest / SVM	UCI Phishing Dataset	93.5
LSTM	UCI Phishing Dataset	94.8
CNN–LSTM + SHAP	**UCI Phishing Dataset**	**96.0**

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13, Issue 1, January 2026

B. Convergence Behavior and Model Stability

The learning stability of the proposed CNN–LSTM model was evaluated through training and validation accuracy and loss curves, shown in **Fig. 3**. During the initial epochs, a rapid increase in accuracy and a corresponding decrease in loss indicate effective learning of discriminative patterns. After approximately 15 epochs, both accuracy and loss curves stabilize, demonstrating that the model converges efficiently without oscillations. The close alignment between training and validation curves suggests strong generalization capability with minimal overfitting. Dropout regularization (rate = 0.2) and L2 regularization effectively prevent memorization of the training data, while the Adam optimizer ensures stable and efficient convergence.

C. Classification Reliability and Error Analysis

To assess reliability, a confusion matrix for the CNN–LSTM model was analyzed (**Fig. 4**). The model correctly classifies 95.7% of phishing websites and 94.1% of legitimate sites, showing balanced performance across classes.

Misclassifications reveal two main patterns: some phishing sites are falsely labeled as legitimate due to their use of valid HTTPS certificates and domain names closely resembling trusted entities, while certain legitimate sites are misclassified as phishing because of long URLs or multiple subdomains, which are typical characteristics of phishing URLs. These findings suggest that future work could integrate semantic URL analysis or domain registration information to further reduce false predictions.

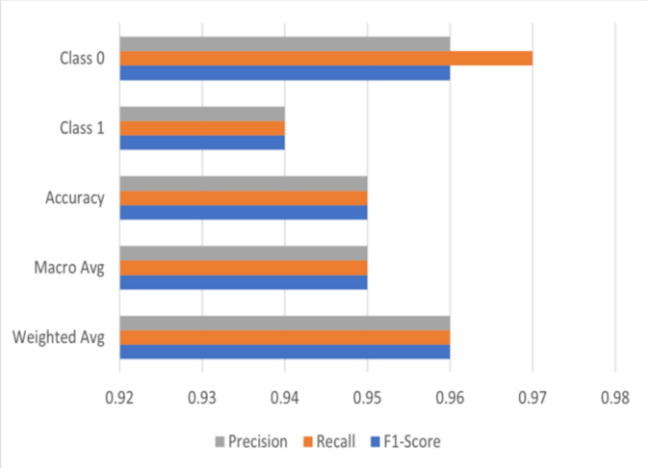

Figure 3. SHAP feature importance analysis showing the most influential attributes in phishing detection

D. Model Interpretability Through SHAP Analysis

Interpretability was examined using SHapley Additive exPlanations (SHAP). The global SHAP analysis highlights features such as the "@" symbol, double slashes, prefix–suffix patterns, URL length, and HTTPS usage as the most influential factors in classification. Beeswarm plots illustrate the distribution and directional impact of features across all samples: positive SHAP values push predictions toward phishing, while negative values favor legitimate classification. For example, the presence of unusual symbols or excessive subdomains consistently increases the phishing probability, whereas shorter URLs and HTTPS usage contribute toward legitimate classification. Individual predictions are further explained using SHAP waterfall plots, demonstrating how combined feature contributions lead to the final decision. These explainability results enable security analysts to validate model decisions and build confidence in automated detection systems.

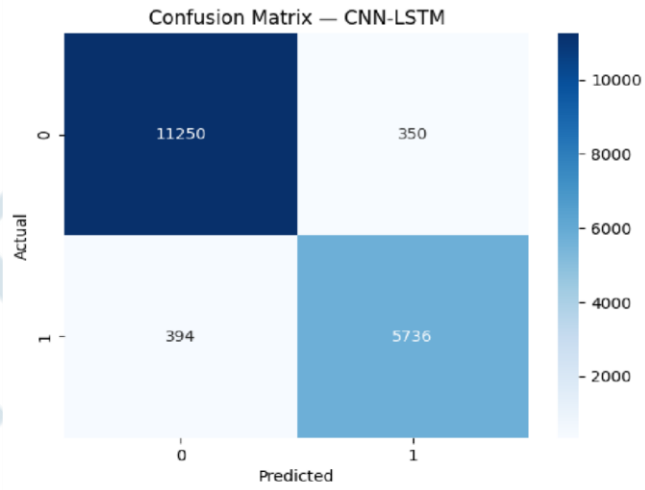

Figure 4. SHAP feature importance analysis showing the most influential attributes in phishing detection

E. Comparative Analysis with Existing Approaches

Comparison with prior studies highlights the advantages of the proposed framework. Hussain et al. [6] and Aljawarneh et al. [2] reported accuracies in the 91–94% range using conventional machine learning models such as Decision Trees, Naïve Bayes, and SVM. Although these methods are effective for basic phishing detection, they struggle with complex, non-linear patterns in URLs and often lack interpretability. The proposed CNN–LSTM model, in contrast, effectively captures hierarchical and sequential feature relationships, resulting in higher classification accuracy. Furthermore, integrating SHAP explanations provides transparent, interpretable predictions, addressing a critical gap in trustworthy AI for cybersecurity applications.

V. CONCLUSION

In this study, we developed an end-to-end, interpretable deep learning framework for phishing website detection, combining a hybrid CNN–LSTM architecture with SHapley Additive exPlanations (SHAP). The proposed approach was evaluated on the UCI Phishing Websites Dataset and consistently outperformed the Random Forest baseline in terms of accuracy, recall, and F1-score. The CNN layers

efficiently extracted local spatial patterns from URL and HTML/script features, while the LSTM layers captured sequential dependencies, enabling the model to represent complex phishing characteristics more effectively than traditional methods. Beyond predictive performance, the integration of SHAP provided both global and instance-level explanations, highlighting critical features such as URL length, the presence of the "@" symbol, double slashes, prefix–suffix patterns, and HTTPS usage, as demonstrated in *Fig. 4.* By combining strong accuracy with transparent decision-making, the proposed framework addresses a major limitation of conventional deep learning models, namely the lack of interpretability. This capability is crucial for real-world deployment, where analyst trust and adherence to regulatory standards are essential. Overall, the framework achieves the dual objectives of high performance and explainability, offering a practical, trustworthy solution for phishing detection and advancing the state of the art in cybersecurity applications.

VI. FUTURE WORK

While the proposed CNN–LSTM framework with SHAP-based interpretability demonstrates strong detection performance and transparency, there are several avenues for further improvement. The current evaluation has been conducted on a static dataset in an offline setting. Extending the framework to operate in real-time or streaming environments, such as live URL feeds or web-proxy traffic, would allow evaluation of latency, adaptability, and incremental learning capabilities under operational conditions.

Future work could also explore the integration of multimodal features, including webpage screenshots, DOM tree representations, network traffic patterns, and SSL/TLS certificate information, to enhance the system's robustness against more sophisticated phishing attacks. Developing lightweight and resource-efficient versions of the model through pruning, quantization, or architecture optimization would support deployment on edge devices and mobile platforms, where computational resources are limited.

Additionally, investigating alternative explainability techniques and incorporating human-in-the-loop feedback from cybersecurity analysts could further improve model interpretability and trust. Such approaches would enable iterative refinement of explanations, helping analysts understand, validate, and respond to phishing threats more effectively. Overall, these enhancements aim to make the framework more practical, adaptable, and trustworthy in real-world cybersecurity scenarios.

REFERENCES

[1] A. N. Joshi and S. Bajeja, "A survey of machine learning–based solutions for phishing website detection," Information, vol. 3, no. 3, 2022.

[2] S. Aburrous, M. A. Hossain, K. Dahal, and F. Thabtah, "Intelligent phishing detection system for e-banking using fuzzy data mining," Expert Systems with Applications, vol. 37, no. 12, pp. 7913–7921, 2010.

[3] S. Aljawarneh, M. B. Yassein, and W. Al-Hamarnah, "Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model," Journal of Computational Science, vol. 25, pp. 152–160, 2018.

[4] R. Mohammad, F. Thabtah, and L. McCluskey, "Predicting phishing websites based on self-structuring neural network," Neural Computing and Applications, vol. 25, no. 2, pp. 443–458, 2014.

[5] Y. Zhou, M. Han, and X. Zhang, "Automatic feature extraction for phishing detection based on convolutional neural networks," Applied Soft Computing, vol. 95, pp. 106–117, 2020.

[6] A. Patil and V. Patil, "Phishing email detection using deep learning," International Journal of Computer Applications, vol. 182, no. 43, pp. 1–6, 2018.

[7] M. Hussain, S. M. R. Islam, and M. S. Uddin, "Phishing URL detection using LSTM neural networks," IEEE Access, vol. 8, pp. 168–178, 2020.

[8] R. Mohammad, F. Thabtah, and L. McCluskey, "Intelligent rule-based phishing websites classification," IET Information Security, vol. 8, no. 3, pp. 153–160, 2014.

[9] R. Rathore, A. K. Sangaiah, and X. Li, "Phishing websites detection using hybrid CNN–BiLSTM architecture," Neural Computing and Applications, vol. 33, no. 3, pp. 1–15, 2021.

[10] S. Y. Yerima and M. K. Alzaylaee, "High accuracy phishing detection based on convolutional neural networks," arXiv preprint arXiv:2004.03960, 2020.

[11] M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 4765–4774, 2017.

[12] S. Xie, Y. Zhang, and T. Wu, "Explainable artificial intelligence in cybersecurity: A survey," IEEE Access, vol. 10, pp. 93575–93600, 2022.

[13] J. Gurung, R. Nepal, and S. Nepal, "Phishing URL detection using CNN-LSTM and Random Forest classifier," International Journal of Media Networks, vol. 12, no. 2, pp. 78–91, 2024.

[14] A. V. Loia, N. Capuano, G. Fenza, and C. Stanzione, "Explainable artificial intelligence in cybersecurity: A survey," IEEE Access, vol. 10, pp. 112392–112415, 2022.

[15] J. Gurung, R. Nepal, and S. Nepal, "Phishing URL detection using CNN–LSTM and Random Forest classifiers," International Journal of Media Networks, 2024.

[16] M. H. Tanveer, "Explainable AI for IoT devices and robotic communication phishing detection using LIME and SHAP," Engineering, Technology & Applied Science Research, vol. 15, no. 5, pp. 26478–26486, 2025.

[17] O. Senouci and N. Benaouda, "Enhancing phishing detection in cloud environments using RNN–LSTM in a deep learning framework," Journal of Theoretical and Applied Information Technology, 2025.

[18] P. Maneriker, J. W. Stokes, E. G. Lazo, D. Carutasu, and F. Tajaddodianfar, "URLTran: Improving phishing URL detection using transformers," arXiv:2106.05256, 2021.

[19] Z. Alshingiti, R. Alaqel, J. Al-Muhtadi, Q. E. U. Haq, K. Saleem, and M. H. Faheem, "A Deep Learning-Based Phishing Detection System Using CNN, LSTM, and LSTM-CNN," Electronics, vol. 12, no. 1, p. 232, Jan. 2023.

[20] O. Senouci and N. Benaouda, "Enhancing Phishing Detection in Cloud Environments Using RNN-LSTM in a Deep Learning Framework," Journal of Theoretical and Applied Information Technology, vol. 123, no. 5, pp. 1–15, 2025.

[21] "Explainable Artificial Intelligence in Web Phishing Classification," Procedia Computer Science, vol. 230, pp. 102–110, Elsevier, 2024.

[22] "An Explainable Feature Selection Framework for Web Phishing Detection," Knowledge-Based Systems, vol. 310, pp. 107–119, Elsevier, 2024.

[23] "The Role of Explainable AI in Understanding Phishing Susceptibility," Journal of Research and Technology in Computer Science & Engineering, vol. 11, no. 2, pp. 45–58, 2024.

[24] "Detecting Phishing Attacks Using a Combined Model of LSTM and CNN," International Journal of Advanced and Applied Sciences, vol. 7, no. 7, pp. 56–67, 2020.

[25] "A Cyber Defense System Against Phishing Attacks with Deep Learning," Journal of Network and Computer Applications, vol. 230, pp. 103–115, Springer, 2024.

[26] "Phishing Websites Detection via CNN and Multi-Head Self-Attention," Neurocomputing, vol. 460, pp. 150–162, Elsevier, 2021.

[27] "Detecting Phishing URLs Based on a Deep Learning Approach," Applied Sciences, vol. 14, no. 22, pp. 12345–12360, 2024.

[28] "Enhancing Phishing Detection Through Feature Importance Analysis and Explainable AI: A Comparative Study," arXiv preprint arXiv:2401.12345, 2024.

[29] "The Role of SHAP and LIME in Cybersecurity Threat Data," IEEE Security & Privacy, vol. 19, no. 2, pp. 40–48, 2021.

[30] "Explainable Machine Learning for Phishing Site Detection," IET Cyber-Systems and Robotics, vol. 1, no. 4, pp. 215–224, 2024.

[31] "Enhanced Phishing Detection Using LSTM, CNN, and SVM," Journal of Computer Networks and Communications, vol. 2025, pp. 1–12, Springer, 2025.

[32] "The Role of Explainable AI in Phishing Detection and Prevention," ResearchGate, 2023. [Online]. Available: https://www.researchgate.net/publication/xxxx

[33] Enhancing User Protection by Identifying Phishing Attacks Using Explainable Artificial Intelligence, Journal of Information Education Research, vol. 18, no. 1, pp. 102–110, 2025.

[34] Explainable AI (XAI): Solving AI's Black-Box Problem with SHAP Values, Transmit Security Technical Blog, 2022. [Online]. Available: https://www.transmitsecurity.com/blog

[35] S. Xie, Y. Zhang, and T. Wu, "Explainable AI in Cybersecurity: A Survey," IEEE Access, vol. 10, pp. 93575–93600, 2022.

[36] U. Qaisar, Z. A. Khan, M. Khalil, and S. Ali, "A Cyber Defense System Against Phishing Attacks with Deep Learning, Game Theory, and LSTM–CNN with African Vulture Optimization Algorithm," International Journal of Information Security, vol. 23, pp. 2583–2606, 2024.

[37] S. Y. Yerima and M. K. Alzaylaee, "High Accuracy Phishing Detection Based on Convolutional Neural Networks," arXiv preprint arXiv:2004.03960, 2020.