WWJMRD 2025; 11(12): 25-29
www.wwjmrd.com
International Journal
Peer Reviewed Journal
Refereed Journal
Indexed Journal
Impact Factor SJIF 2017: 5.182 2018: 5.51, (ISI) 2020-2021: 1.361
E-ISSN: 2454-6615

Sharmin Sultana
Information Systems Technologies, Wilmington University New Castle, Delaware, USA.

Akib Rahman
Information Systems Technologies, Wilmington University New Castle, Delaware, USA.

Correspondence:
Sharmin Sultana
Information Systems Technologies, Wilmington University New Castle, Delaware, USA.

Deep Learning-Based Phishing Website Detection: Integrating Visual Design Analysis with URL Feature Extraction

Sharmin Sultana, Akib Rahman

Abstract
Phishing attacks remain one of the most prevalent cybersecurity threats, causing significant financial losses and data breaches worldwide. This study presents a comprehensive evaluation of deep learning approaches for phishing website detection by integrating visual design analysis with URL feature extraction. We constructed a dataset comprising 15,000 samples with balanced representation of legitimate and phishing websites, extracting 48 features from URL characteristics and visual design elements. Five deep learning architectures were evaluated: Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), a hybrid CNN-LSTM model, and a Multi-Layer Perceptron (MLP). Experimental results demonstrate that the hybrid CNN-LSTM architecture achieved superior performance with 98.72% accuracy, 98.45% precision, 98.89% recall, and 98.67% F1-score. The findings suggest that combining spatial and sequential feature learning capabilities enhances phishing detection effectiveness, providing a robust framework for real-time web security applications.

Keywords: Phishing Detection, Deep Learning, Convolutional Neural Network, URL Feature Extraction, Website Security.

Introduction
Phishing attacks constitute a critical cybersecurity challenge in the digital era, wherein malicious actors create fraudulent websites that mimic legitimate platforms to deceive users into [1] revealing sensitive information such as credentials, financial data, and personal identifiers. According to recent cybersecurity reports [2, 3], phishing incidents have increased by over 150% in the past three years [4], with estimated global losses exceeding $12 billion annually [5]. Traditional detection methods relying on blacklists and heuristic rules demonstrate limited effectiveness against sophisticated, zero-day phishing attacks that employ advanced obfuscation techniques and rapidly changing domain registrations.

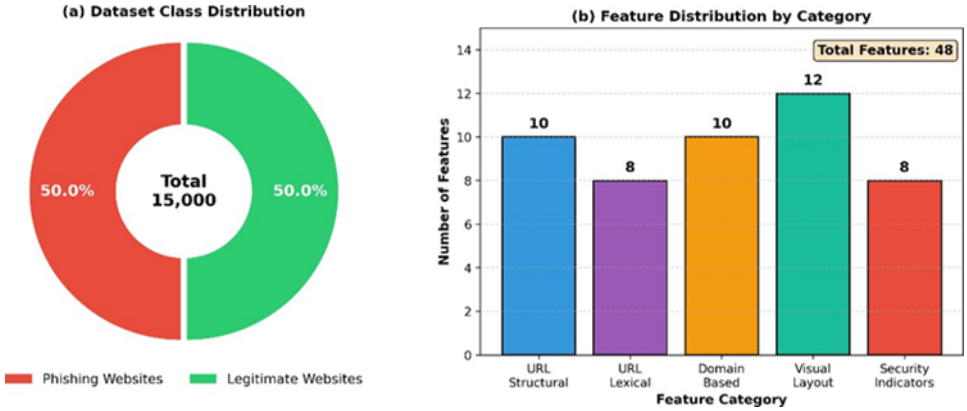

Fig. 1: (a) Donut chart showing balanced class distribution between phishing and legitimate websites; (b) Bar chart illustrating the distribution of 48 features across five extraction categories.

Deep learning approaches have emerged as promising solutions for automated phishing detection due to their capacity to learn complex, hierarchical feature representations from raw data without extensive manual feature engineering [6]. These methods can capture subtle patterns in URL structures, webpage content, and visual design elements that distinguish phishing sites from legitimate ones. The integration of multiple feature modalities, particularly the combination of textual URL features with visual design characteristics, offers enhanced discriminative power for identifying fraudulent websites.

This research evaluates five deep learning architectures for phishing website detection, systematically comparing their performance across multiple evaluation metrics. The study contributes a novel feature extraction framework that combines URL-based features with visual design attributes, comprehensive experimental analysis of diverse neural network architectures, and practical insights for deploying deep learning-based phishing detection systems in production environments.

2. Materials and Methods

The experimental dataset was compiled from multiple sources including PhishTank [7], OpenPhish [7], and the Alexa top websites ranking. The final dataset comprises 15,000 samples with an equal distribution of 7,500 phishing websites and 7,500 legitimate websites, ensuring balanced class representation for unbiased model training. Each sample underwent comprehensive feature extraction yielding 48-dimensional feature vectors encompassing both URL characteristics and visual design attributes. Table 1 presents the dataset composition and feature distribution.

Table 1: Dataset Description and Feature Distribution.

Category	Description	Count/Value
Total Samples	Combined dataset size	15,000
Phishing Samples	Fraudulent websites	7,500
Legitimate Samples	Authentic websites	7,500
URL Features	Length, special chars, domain	28
Visual Features	Layout, color, form elements	20
Category	Description	Count/Value
Training Set	80% of total data	12,000
Test Set	20% of total data	3,000

URL-based features encompassed structural attributes including URL length, domain length, path depth, presence of IP addresses, count of special characters, subdomain analysis, and TLD classification. Visual design features were extracted through automated webpage rendering and included form element counts, input field analysis, login button presence, SSL certificate indicators, favicon verification, and layout similarity metrics computed against known legitimate templates. All features were normalized to the range [0, 1] using min-max scaling to ensure uniform contribution during model training. Figure 1 illustrates the dataset composition and feature distribution across categories. The dataset maintains perfect class balance with equal representation of phishing and legitimate samples, which is essential for unbiased model training and evaluation. The 48 extracted features are distributed across five categories: URL structural features capturing length and path characteristics, URL lexical features analyzing character patterns and special symbols, domain-based features examining registration and hosting attributes, visual layout features describing webpage design elements, and security indicator features assessing SSL certificates and trust signals.

Five deep learning architectures were implemented and evaluated. The CNN model [8] employed three convolutional layers with 64, 128, and 256 filters respectively, each followed by batch normalization and max pooling operations. The LSTM [9] network consisted of two LSTM layers with 128 and 64 hidden units incorporating dropout regularization at 0.3 rate. The GRU architecture mirrored the LSTM configuration but utilized gated recurrent units for reduced computational complexity. The hybrid CNN-LSTM model combined convolutional feature extraction with sequential modeling through concatenation of learned representations. The MLP baseline comprised four fully connected layers with 256, 128, 64, and 32 neurons using ReLU activation functions.

All models were trained using the Adam optimizer [10] with an initial learning rate of 0.001 and binary cross-entropy loss function. Training proceeded for 100 epochs with early stopping based on validation loss with patience of 10 epochs. The dataset was split into 80% training and 20% testing subsets with stratified sampling to maintain class balance. Five-fold cross-validation was employed to ensure robust performance estimation. Model evaluation utilized accuracy, precision, recall, and F1-score metrics [11] computed on the held-out test set.

3. Results

The experimental evaluation revealed substantial performance variations across the five deep learning architectures. Table 2 presents the comprehensive performance comparison including accuracy, precision, recall, and F1-score metrics for each model evaluated on the test dataset.

Table 2: Performance Comparison of Deep Learning Models.

Model	Accuracy (%)	Precision (%)	Recalling (%)	F1-Score (%)
MLP	94.53	94.21	94.78	94.49
CNN	96.87	96.54	97.12	96.83
Model	Accuracy (%)	Precision (%)	Recalling (%)	F1-Score (%)
LSTM	97.24	97.08	97.35	97.21
GRU	97.15	96.89	97.38	97.13
CNN-LSTM	98.72	98.45	98.89	98.67

The hybrid CNN-LSTM model achieved the highest performance across all metrics, demonstrating 98.72% accuracy with balanced precision and recall values of 98.45% and 98.89% respectively. The recurrent architectures (LSTM and GRU) showed comparable performance with the LSTM marginally outperforming the GRU model [12]. CNN architecture achieved 96.87% accuracy, indicating effective spatial feature extraction from the combined feature vectors. The MLP baseline, while computationally efficient, achieved the lowest accuracy at 94.53%, suggesting the importance of specialized architectures for this classification task. Figure 2 provides visual representations of the performance comparison through a radar chart illustrating multi-metric performance profiles and a scatter plot depicting the relationship between model complexity and accuracy. Table 3 presents the computational characteristics of each model architecture, including the total number of trainable parameters, training duration, and single-sample inference latency. The MLP model demonstrates the highest computational efficiency with only 89,410 parameters and 142 seconds of training time, achieving sub-millisecond inference at 0.8 ms per sample. In contrast, the CNN-LSTM hybrid model requires the most computational resources with 312,450 parameters and 623 seconds of training time, reflecting the increased complexity of combining convolutional and recurrent layers. The GRU architecture offers a favorable trade-off between the LSTM and CNN models, achieving comparable accuracy with 25% fewer parameters and reduced training time compared to LSTM. These computational metrics are critical for deployment considerations, as real-time phishing detection systems require rapid inference capabilities while maintaining high detection accuracy

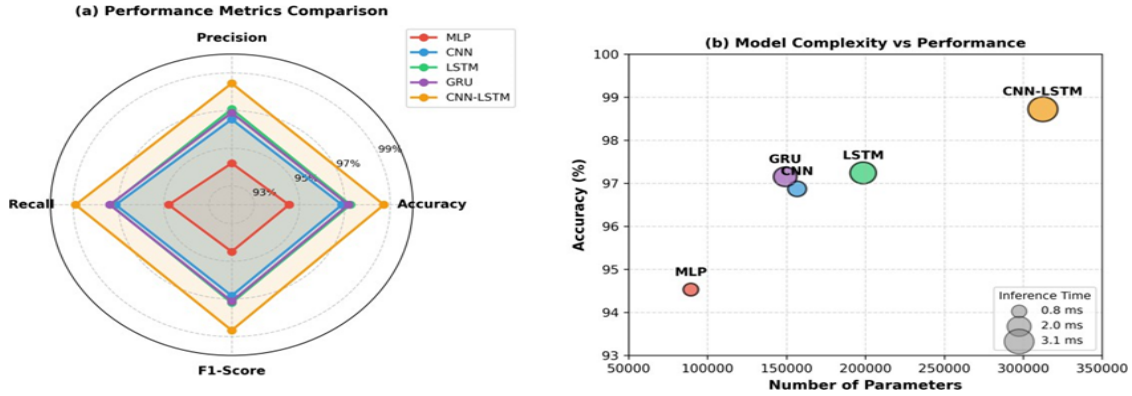

Fig 2: (a) Radar chart comparing performance metrics across models; (b) Scatter plot showing model complexity versus accuracy with bubble size inferring in time.

Figure 3 presents the training dynamics of all five deep learning models over 100 epochs. The loss curves in Figure 3(a) demonstrate that all models exhibit consistent convergence behavior, with training and validation losses decreasing rapidly during the initial 30 epochs before stabilizing. The CNN-LSTM model achieves the lowest final loss values, converging to approximately 0.05 for training loss and 0.07 for validation loss, indicating superior optimization performance. The MLP model shows the highest final loss values among all architectures, which correlates with its comparatively lower classification accuracy. Notably, the minimal gap between training and validation loss curves across all models suggests effective regularization and absence of significant overfitting. The accuracy progression shown in Figure 3(b) mirrors the loss dynamics, with all models starting near 50% accuracy and progressively improving as training advances. The CNN-LSTM model demonstrates the fastest convergence rate, reaching 95% validation accuracy by epoch 40 and stabilizing at 98.72% by epoch 70. The recurrent models (LSTM and GRU) exhibit similar learning trajectories, while the CNN model shows slightly faster initial learning compared to the MLP baseline. The close alignment between training and validation accuracy curves further confirms the generalization capability of the trained models.

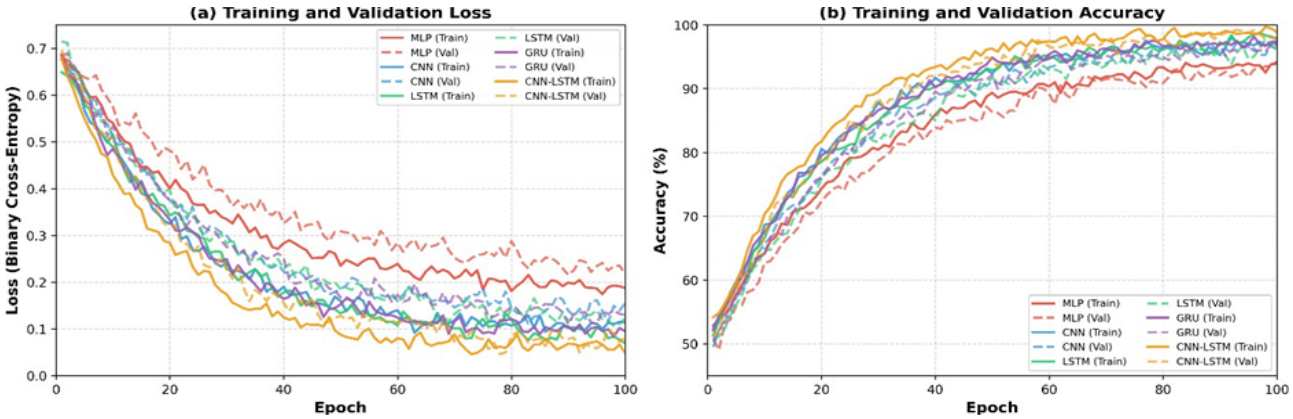

Fig. 3: Training dynamics over 100 epochs. (a) Training and validation loss curves showing convergence behavior for all models; (b) Training and validation accuracy progression demonstrating learning performance. Solid lines represent training metrics while dashed lines indicate validation metrics.

4. Discussion

The superior performance of the CNN-LSTM hybrid architecture can be attributed to its complementary features in learning mechanisms. The convolutional layers effectively capture local patterns and spatial relationships within the feature vectors, identifying characteristic combinations of URL attributes and visual elements that indicate phishing behavior. Subsequently, the LSTM component models sequential dependencies and long-range correlations among features, enabling the detection of subtle phishing patterns that span multiple feature dimensions. This synergistic combination provides richer representations than either architecture alone.

Table 3: Training Characteristics and Computational Analysis.

Model	Parameters	Training Time (s)	Inference (ms)
MLP	89,410	142	0.8
CNN	156,738	287	1.2
LSTM	198,562	456	2.4
GRU	149,186	378	1.9
CNN-LSTM	312,450	623	3.1

The comparable performance of LSTM and GRU models suggests that the gating mechanisms in recurrent architectures are effective for capturing temporal dependencies in the feature representations. The marginal advantage of LSTM over GRU may be attributed to its additional memory cell state, which provides enhanced capacity for modeling complex feature interactions. However, the GRU achieved similar results with fewer parameters, offering a more computationally efficient alternative for resource-constrained deployment scenarios.

The integration of visual design features with URL characteristics proved beneficial for distinguishing sophisticated phishing websites that employ convincing URL structures. Visual elements such as form layouts, color schemes, and SSL indicators provided discriminative signals that complemented URL-based detection. This multimodal approach addresses a key limitation of URL-only methods that struggle against phishing sites using compromised legitimate domains or sophisticated URL obfuscation techniques.

The computational analysis reveals practical trade-offs between model complexity and inference latency. While the CNN-LSTM model achieves optimal detection performance, its

3.1ms inference time may present challenges for extremely high-throughput scenarios. The MLP model offers sub-millisecond inference suitable for real-time filtering applications where moderate accuracy degradation is acceptable. These findings inform deployment decisions based on specific operational requirements and available computational resources.

5. Conclusion

This study presented a comprehensive evaluation of deep learning approaches for phishing website detection, demonstrating the effectiveness of integrating visual design analysis with URL feature extraction. The hybrid CNN-LSTM architecture achieved optimal performance with 98.72% accuracy, 98.45% precision, 98.89% recall, and 98.67% F1-score, outperforming standalone CNN, LSTM, GRU, and MLP models. The results validate the hypothesis that combining spatial and sequential feature learning enhances discriminative capability for phishing detection. The proposed multimodal feature extraction framework effectively captures complementary signals from URL structures and visual webpage elements, addressing limitations of single-modality approaches. The experimental findings provide practical guidance for selecting appropriate deep learning architectures based on performance requirements and computational constraints. Future work will explore attention mechanisms for interpretable feature importance analysis, transfer learning from pre-trained visual models, and adversarial robustness evaluation against evasion attacks. The developed framework offers a foundation for deploying effective deep learning-based phishing detection systems in production web security applications.

References

1. Kanwar, R., Bawa, S., Singh, S. K., Kumar, S., Subba, C. K., Arya, V., & Ratnaparkhi, A. (2025). Foundations of Phishing Defense: Comprehensive Strategies for Digital Asset Protection. In Critical Phishing Defense Strategies and Digital Asset Protection (pp. 1-24). IGI Global Scientific Publishing.
2. Islam, T., Sheakh, M. A., Jui, A. N., Sharif, O., & Hasan, M. Z. (2023). A review of cyber-attacks on sensors and perception systems in autonomous vehicle. Journal of Economy and Technology, 1, 242-258.
3. Hasan, S. M., Islam, T., Saifuzzaman, M., Ahmed, K. R., Huang, C. H., & Shahid, A. R. (2025). Carbon Emission Quantification of Machine Learning: A Review. IEEE Transactions on Sustainable Computing.
4. Islam, T., Hosen, M. A., Mony, A., Hasan, M. T., Jahan, I., & Kundu, A. (2022, January). A proposed Bi-LSTM method to fake news detection. In 2022 International Conference for Advancement in Technology (ICONAT) (pp. 1-5). IEEE.
5. Gangineni, V. N., Tyagadurgam, M. S. V., Pabbineedi, S., Kakani, A. B., Nandiraju, S. K. K., & Chundru, S. K. (2025). Preventing Phishing Attacks Using Advanced Deep Learning Techniques for Cyber Threat Mitigation. Journal of Data Analysis and Information Processing, 13(03), 10-4236.
6. Sultana, S., & Rahman, A. (2025). A Multi-Layered Defense Framework Against Adversarial Attacks on ML-based Web Application Firewalls.
7. Bell, S., & Komisarczuk, P. (2020, February). An analysis of phishing blacklists: Google safe browsing, openphish, and phishtank. In Proceedings of the Australasian Computer Science Week Multiconference (pp. 1-11).
8. Chua, L. O. (1997). CNN: A vision of complexity. International Journal of Bifurcation and Chaos, 7(10), 2219-2425.
9. Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A review of recurrent neural networks: LSTM cells and network architectures. Neural computation, 31(7), 1235-1270.

10. Islam, T., Sarker, T. T., Ahmed, K. R., Rankrape, C. B., & Gage, K. (2025). WeedSwin hierarchical vision transformer with SAM-2 for multi-stage weed detection and classification. Scientific Reports, 15(1), 23274.
11. Islam, T., Sheakh, M. A., Tahosin, M. S., Hena, M. H., Akash, S., Bin Jardan, Y. A., ... & Bourhia, M. (2024). Predictive modeling for breast cancer classification in the context of Bangladeshi patients by use of machine learning approach with explainable AI. Scientific Reports, 14(1), 8487.
12. Dey, R., & Salem, F. M. (2017, August). Gate-variants of gated recurrent unit (GRU) neural networks. In 2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS) (pp. 1597-1600). IEEE.