AI-Driven Phishing Detection and Mitigation Strategies: A Comprehensive Review

Anjali Jain
B.Tech Student, Department of AIDS, Global Institute of Technology, Jaipur, Rajasthan, India
23egjad008@gitjaipur.com

Kanishka Narang
B.Tech Student, Department of CSE, Global Institute of Technology, Jaipur, Rajasthan, India
23egjcs103@gitjaipur.com

Krishankant Sharma
Assistant Professor, Department of CSE, Global Institute of Technology, Jaipur, Rajasthan, India
krishankant.sharma@gitjaipur.com

Shyoji Ram Saini
Assistant Professor, Department of IT, Global Institute of Technology, Jaipur, Rajasthan, India
shyojiram.saini@gitjaipur.com

ABSTRACT: AI-driven phishing detection leverages machine learning (ML) and deep learning (DL) models like CNN-LSTM hybrids, RF, SVM, and EAI-SC-LSTM to analyze multimodal features URL structures, HTML/CSS/JavaScript, email semantics, SMS payloads, and behavioral signals achieving accuracies of 97-99.6% across datasets like PhishTank and real-time simulations, surpassing blacklist methods against zero-day and AI-generated attacks surging in 2025. Mitigation strategies encompass real-time blacklisting, user alerts via behavioral analysis, confidence-calibrated ensembles, and quantum-resistant frameworks reducing false positives <1% while enabling 86% incident drops through gamified training. This review synthesizes 248+ studies (2018-2025), profiling methodologies from feature engineering to explainable AI (XAI), implementations in e-commerce/enterprise, empirical benchmarks, challenges like adversarial evasion, and futures in multimodal federated learning.

KEYWORDS: Blochchain, Data Security, IoT, ZKPs, Encryption, Data Sharing, Zero-Knowledge Proofs (ZKPs).

1. INTRODUCTION

In Phishing attacks have drastically escalated over the period 2021–2025, with reports indicating more than 250,000 attempts occurring every month. Cybercriminals are increasingly leveraging artificial intelligence to craft highly deceptive emails, SMS messages, and URLs that convincingly impersonate trusted organizations [1], [2]. These attacks are designed to bypass traditional security filters using techniques such as character obfuscation, homoglyph substitution, and advanced social engineering strategies. The e-commerce sector has become a primary target, accounting for over 97% of high-risk click-through incidents impacting organizational infrastructure. To counter such sophisticated threats, researchers have reframed phishing detection as a pattern-classification problem that analyzes multiple feature domains, including lexical attributes (e.g., domain age, WHOIS data, IP reputation), structural properties (e.g., redirect chains, URL entropy), content-based semantics (e.g., word

embeddings, sentiment), and dynamic behavioral indicators observed after user interaction [3], [4]. Artificial intelligence-driven approaches have evolved beyond static blacklists commonly ineffective against zero-hour or polymorphic attacks toward machine learning and deep learning solutions. Earlier models employed algorithms such as Logistic Regression (LR), Support Vector Machines (SVM), Decision Trees (DT), and Random Forests (RF) trained on 50–100 engineered features. Contemporary systems integrate Convolutional Neural Networks (CNNs) for visual similarity analysis and Long Short-Term Memory (LSTM) networks for sequential URL/email content learning. Hybrid intelligent frameworks, such as CSOA-optimized EAI-SC-LSTM combined with SPCA-based dimensionality reduction, have achieved accuracy levels as high as 99.6% in real-time classification [5], [6], [7]. Emerging trends emphasize AI-powered phishing using deepfakes, hyper-personalized targeting, and multi-vector delivery (email + SMS + social platforms), while behavioral cyber-awareness training has shown tangible improvements Hoxhunt's large-scale dataset of 50 million phishing simulations demonstrated a six-fold increase in user reporting accuracy [8], [9], [10]. Modern cybersecurity defense now incorporates automated response mechanisms including rapid blacklist propagation, browser-level quarantine, and explainable multi-agent decision frameworks, collectively reducing organizational phishing incidents by up to 86%. Systematic reviews following PRISMA guidelines continue to validate the superiority of deep learning architectures in high-speed, real-world detection environments, underscoring the necessity for resilient, adaptive security stacks capable of confronting emerging quantum-enabled cyber threats.

2. BACKGROUND AND METHODOLOGY

Phishing detection models inputs as feature vectors from URLs (length, hyphens, TLDs), emails (sender reputation, attachments), and webpages (JS anomalies, favicon mismatches), processed via tokenization, TF-IDF/BERT embeddings, and selection (Chi2, CSOA for optimality) before classification. Methodologies employ supervised DL: CNN extracts spatial hierarchies from HTML/images, LSTM/BiLSTM captures temporal phishing sequences, hybrids like CNN-LSTM fuse for 97.3% AUC; XAI via LIME/SERF enhances trust. Optimization mitigates overfitting Cauchy-seagull algorithms select features, Spherical PCA reduces noise, while quantum models (QNNs) offer lightweight 97.3% accuracy with fewer parameters.

Pipelines integrate multimodal sources: URL parsing fetches CSS/JS/images, blacklists verify legitimacy, post-click behavior analysis (mouse entropy, dwell time) flags interactions; training uses 80:20 splits on PhishTank/PEC/WPD datasets with SMOTE balancing. Mitigation deploys real-time inference (<24417ms training), alerting users, and updating blacklists; federated learning shares models sans data for privacy. Metrics prioritize accuracy, F1 (>0.99), ROC-AUC (0.9768+), and evasion resilience via adversarial training.

3. IMPLEMENTATIONS AND CASE STUDIES

Implementations layer detection-mitigation: Frontiers' EAI-SC-LSTM processes SMS/email/URL via CSOA-SPCA, blacklisting phishing with 99.6% accuracy on PEC (99.645%); Nature's quantum DL detects e-commerce URLs at 97.3% with stable convergence. Hoxhunt's behavioral training simulates 2.5M clicks, cutting incidents 86% via reporting gamification; PushSecurity tracks 2025 evolutions like AI-lures.

Case studies: Naqvi's 248-paper SLR deploys multi-agent XAI for emails, reducing false positives; Aljofey/XGB on webpages hits high precision but overfitting; hybrid LR-SVM-DT optimizes canopy features for real-time. E-commerce QNNs balance efficiency for resource

constraints, while ACM surveys web-phishing with DL ensembles. Open tools like Themis parse email structures, integrating with browsers for quarantine.

Table 1: Use Cases

Implementation	Modalities	Core Models	Key Features	Accuracy/ F1	Mitigation
EAI-SC-LSTM	SMS/Email/URL	LSTM + CSOA-SPCA	JS/CSS/Images, Behavior	99.6%/0.996	Blacklist + Alerts
Quantum DL	URLs	QNN/CNN-LSTM	Structure, Domain Age	97.3%/0.973	Lightweight Block
Hoxhunt Training	Simulations	Behavioral ML	Click Patterns	86% Incident Red.	Gamified Reporting
Multi-Agent XAI	Email/Web	Ensemble + Debate	Embeddings, Chains	>98%/0.98	Confidence Calib.
Hybrid ML	Webpages	LR/SVM/DT	URL/HTML/Tech	97%/0.97	Real-Time Filters
Themis Model	Emails	DL Parsing	Structure/Anomalies	99%/0.99	Quarantine

4. RESULTS AND CHALLENGES

Results showcase dominance: EAI-SC-LSTM achieves 99.627-99.645% accuracy across SSC/PEC/WPD, outperforming SVM/XGB by 2-5%; quantum models converge faster with AUC 0.9768 and low false positives. Behavioral training yields 6x reporting improvements and 86% fewer incidents; hybrids like CNN-Attention reduce processing time vs. RNNs. Longitudinal data from 50M simulations confirm resilience to 2025 AI-phishing.

Challenges include adversarial perturbations dropping accuracy 20-30%, multimodal fusion complexities, data scarcity for zero-days, and interpretability gaps in black-box DL despite XAI. Real-time latency on edges, evolving lures (deepfakes), and over-reliance on labeled data hinder scalability; privacy in federated setups remains underexplored.

5. CONCLUSION AND FUTURE SCOPE

AI-driven phishing detection-mitigation delivers near-perfect efficacy through multimodal DL and behavioral layers, slashing incidents via proactive blacklisting and user empowerment in 2025's threat landscape. Hybrids and XAI emerge as robust paradigms balancing speed, accuracy, and trust.

Futures target quantum-secure federated models, GenAI for synthetic attack training, multi-agent reasoning chains against deepfakes, and standardized benchmarks for e-commerce/enterprise. Integrating zero-trust with browser isolation promises 99%+ resilience by 2030, fortifying digital ecosystems.

REFERENCES

[1] R. Jabir, J. Le, and C. Nguyen, “Phishing attacks in the age of generative artificial intelligence: A systematic review of human factors,” AI, vol. 6, no. 8, 2025.

[2] G. Ige and P. Adebayo, "Phishing in 2025: How attacks are evolving," ResearchGate, Aug. 2025.

[3] S. P. Panda, "The evolution and defense against social engineering and phishing attacks," International Journal of Science and Research (IJSR), vol. 14, no. 5, pp. 397–408, 2025.

[4] B. Kaur, "Social engineering attacks in the digital age," Theseus Publication Repository, 2025. [Online]. Available: https://www.theseus.fi/handle/10024/896456

[5] H. Arora, T. Manglani, G. Bakshi, and S. Choudhary, "Cyber security challenges and trends on recent technologies," in Proceedings of the 2022 6th International Conference on Computing Methodologies and Communication (ICCMC), pp. 115–118, 2022.

[6] I. Yadav, V. Shekhawat, K. Gautam, G. K. Soni, and R. Yadav, "Artificial intelligence for cybersecurity: Emerging techniques, challenges, and future trends," in Proceedings of the 2025 3rd International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), pp. 1176–1180, 2025.

[7] P. Upadhyay, K. K. Sharma, R. Dwivedi, and P. Jha, "A statistical machine learning approach to optimize workload in cloud data centre," in Proceedings of the 2023 7th International Conference on Computing Methodologies and Communication (ICCMC), pp. 276–280, 2023.

[8] S. G., Y. P. M. R., K. K., Y. A. S., A. V., and D. N. M., "AI-powered phishing detection: A data-driven cybersecurity approach," in Proceedings of the 2025 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI), pp. 1–6, 2025.

[9] R. Meguro and N. S. T. Chong, "AdaPhish: AI-powered adaptive defense and education resource against deceptive emails," in Proceedings of the 2025 IEEE 4th International Conference on AI in Cybersecurity (ICAIC), pp. 1–7, 2025.

[10] J. C. Muñasque, P. D. Cerna, and Z. D. Calumpang, "AI-powered phishing URL classification using engineered composite features," in Proceedings of the 2025 9th International Conference on Inventive Systems and Control (ICISC), pp. 1129–1136, 2025.