FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 10 No. 14, June, 2026, pp 367 – 372
DOI: https://doi.org/10.33003/fjs-2026-1014-5671

Review on Phishing Threat Advancement in Todays Digital Environments: Detection Techniques, Agents and Future Directions

***Buhari Dauda, Ismaila Idris, Noel Moses Dogonyaro and Suleiman Ahmad**

Department of Cyber Security Science, Federal University of Technology, Minna, Niger State, Nigeria.

*Corresponding authors' email: daudabuhari@futminna.edu.ng

Received: 05 Aug. 2026
Accepted: 17 Aug. 2026
Published: 27 Aug. 2026

ABSTRACT
Phishing threats serve as a continuous advanced digital security threat which remains active throughout all contemporary online platforms. The researchers conducted a systematic review to study the historical development of phishing link attacks by analyzing 26 peer-reviewed studies that researchers published between 2020 and 2025. The review employs the PRISMA 2020 framework to evaluate phishing agents and attack strategies and detection methods and upcoming research paths which it establishes through evidence from IEEE, ACM, ScienceDirect and Scopus databases. The research shows artificial intelligence combined with social engineering methods and multi-channel delivery systems which include email and SMS and voice calls and social media platforms resulting in more sophisticated phishing attacks. The research establishes machine learning and deep learning techniques which include CNNs and RNNs and transformers and GCNs and multimodal models as the most effective methods to identify phishing URLs and emails because they achieve more than 95% accuracy across different research studies. The deployment of machine learning solutions encounters difficulties because adversarial attacks and domain adaptation issues and the need for standardized benchmarking datasets remain unresolved. The review establishes that future defense systems need to incorporate explainable AI together with adversarial-resilient models and cross-dataset standardization and multimodal architectures and human-centric training strategies for effective defense against upcoming phishing threats.

Keywords: Artificial Intelligence, Cybersecurity, Deep Learning, Machine Learning, Phishing Detection, Phishing Links

INTRODUCTION

Phishing attacks operate as deceptive schemes through which cybercriminals develop fake emails and messages and websites that imitate authentic sources to deceive users into disclosing their private data. The current defensive systems against phishing attacks fail to protect organizations because attackers have developed increasingly advanced methods which result in annual worldwide financial losses that reach approximately $100 billion according to Taha et al. 2024.

The current phishing environment now uses multiple methods which include SMS (smishing), voice calls (vishing), social media networks, and collaboration platforms instead of relying solely on email as their primary attack method. Cybersecurity experts face an exceptional challenge because attackers use multiple channels together with artificial intelligence to create this particular problem. Phishing attacks now use machine learning algorithms to create tailored deceptive content which security systems and human experts find hard to identify (Hao et al., 2024).

The systematic review seeks to address the following research questions: Why do phishing attacks continue to pose a significant threat despite advances in cybersecurity technologies? Which areas have become the main targets of highly sophisticated phishing campaigns? Which detection techniques are most effective for identifying and classifying phishing attacks? What defense mechanisms can be adopted to prevent phishing attacks effectively in the future while ensuring practical solutions and protecting users' privacy?

MATERIALS AND METHODS

The researchers conducted their systematic review according to the PRISMA 2020 guidelines (Page et al., 2021), which enabled them to produce transparent results that could be verified through detailed and systematic research procedures. The review process included four stages, beginning with identification and screening, followed by eligibility assessment and inclusion. A comprehensive literature search was performed using four main academic databases: IEEE Xplore, ScienceDirect, ACM Digital Library, and Scopus. The search was limited to publications between January 2020 and December 2025 to capture recent advancements in phishing detection and attack methodologies. Boolean operators were used with key search terms including "phishing detection," "machine learning phishing," "AI-generated phishing," "URL detection," "spear phishing," and "phishing countermeasures."

Table 1: Inclusion and Exclusion Criteria

Inclusion criteria:	Exclusion criteria:
Peer-reviewed research articles, conference papers, or systematic reviews;	Non-English publications;
Focus on phishing attack methods, detection techniques, or defense mechanisms;	Publications prior to 2020;
Publication in English language;	Non-peer-reviewed articles or opinion pieces;
Empirical or theoretical contributions to phishing research;	Studies not directly relevant to phishing link threats;

Inclusion criteria:	**Exclusion criteria:**
Publication from 2020-2025.	Duplicate publications.

Study Selection Process

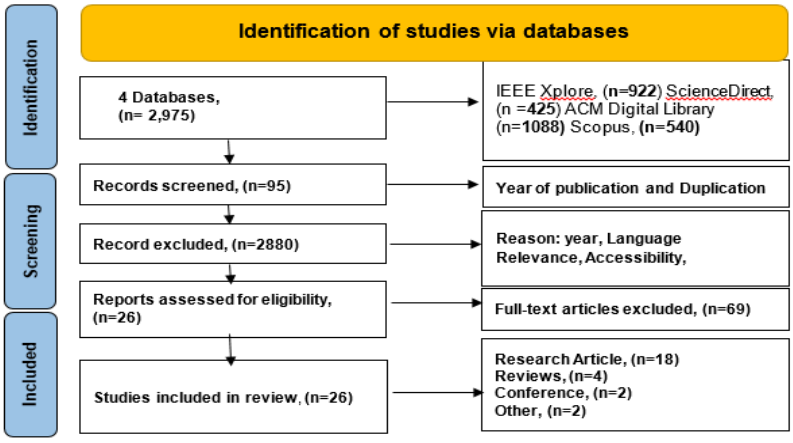

Figure 1: PRISMA 2020 Flow Diagram of Study Selection Process

The first search produced a total of 2975 records which were distributed across all databases. The research began with a total of 95 records which underwent full-text examination after duplicate records and title and abstract screening had been completed. After conducting eligibility evaluations 26 studies passed all inclusion requirements and were chosen for final assessment see Figure 1.

RESULTS AND DISCUSSION

Characteristics of Included Studies

A total of 26 studies were included in the systematic review after the identification, screening, and eligibility assessment of the retrieved literature. The initial search yielded 2,975 records, from which 95 records remained for further screening and assessment after the removal of duplicates and preliminary screening. Following the application of the predefined inclusion and exclusion criteria, 26 studies were considered sufficiently relevant to the objectives and scope of the review and were therefore included in the final analysis. Thus, the number of included studies was determined by the eligibility of the available literature rather than by an arbitrary sample size. The 26 included studies comprised various publication types: 18 research articles (69.2%), 4 review articles (15.4%), 2 conference papers (7.7%), and 2 other publications, including guidelines and preprints (7.7%). The publication data showed an increasing trend from 2020 to 2025, reaching its highest point in 2024, reflecting the growing research interest in phishing threats and their detection.

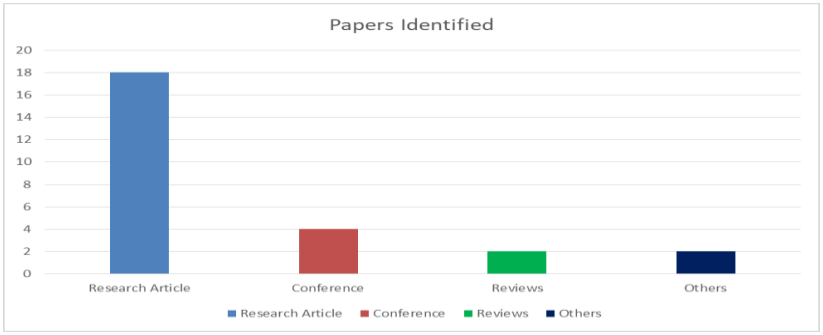

Figure 4(a): Article type distribution

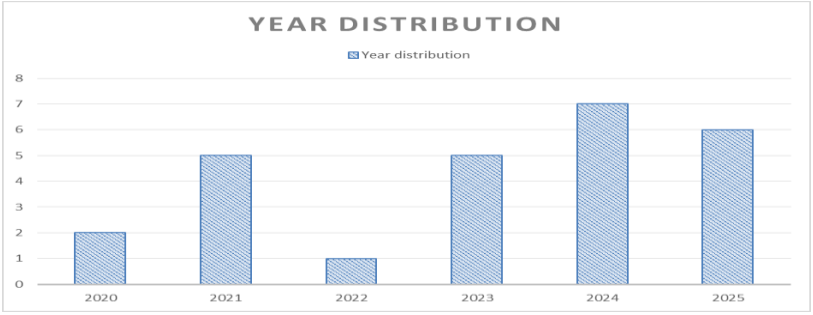

Figure 4(b): Year distribution of papers review

The publication timeline for the 26 studies that this systematic review assessed begins in 2020 and ends in 2025, which Figure 4(b) shows. The chart displays continuous research output growth throughout the years studied, and it shows a significant research output peak which occurs in 2024. The upward trend demonstrates that researchers and professionals are becoming more interested in phishing threats because these threats are developing into more complex and widespread attacks. The study includes 2025 research, which contains preprints and recent conference papers to demonstrate the most recent technological developments and emerging trends in phishing detection and defense.

Key Findings from Reviewed Studies

Table 2: Summary of Selected Studies and Their Key Contributions

Author(s)	Year	Focus Area	Key Contribution/Findings
Hannousse & Yahiouche	2020	Dataset creation and benchmarking	Produced balanced dataset of 11,430 URLs; Random Forest achieved 96.61% accuracy; emphasized need for standardized phishing datasets
Maneriker et al.	2020	Transformer-based URL detection	URLTran model achieved 86.80% true positive rate at 0.01% false positive rate; demonstrated transformer superiority over classical models
Gupta et al.	2021	Lightweight lexical URL detection	Achieved 99.57% accuracy using only 9 URL features; real-time performance with 22.5ms response time
Alhogail & Alsabih	2021	GCN + NLP for email detection	Graph Convolutional Network achieved 98.2% accuracy for phishing email detection
Zhou et al.	2023	Hybrid domain-feature detection	LightGBM model with hybrid features achieved 93.93% accuracy for website detection
Haq et al.	2024	1D CNN for URL detection	Deep learning approach demonstrated high accuracy for URL-based phishing detection
Tian et al.	2025	Multimodal BERT + HTML analysis	WebGuard++ achieved up to 7.9× higher detection rate at low false positive rates
Uddin & Sarker	2024	Explainable transformer for emails	DistilBERT model achieved 98.48% accuracy with LIME interpretability

Phishing Attack Evolution

The review highlights the continued evolution and increasing sophistication of phishing attacks. Phishing has increasingly expanded beyond conventional email and website-based attacks to social media and message-based platforms, where attackers exploit social trust, contextual cues, and malicious links to deceive users. Shuaibu *et al*. (2025) demonstrated this development through a message-based phishing detection study focused on social media platforms, using features such as message length, link presence, and urgency cues. Their study compared Random Forest, Support Vector Machine, and Deep Neural Network models and reported that the DNN achieved 99% across the evaluated performance measures. These findings demonstrate that the changing communication environment has created new phishing detection requirements, particularly for attacks delivered through rapidly changing social-media messages.

AI-generated phishing attacks have also enabled more effective and convincing phishing campaigns compared with conventional phishing methods. Attackers increasingly employ multiple communication channels, targeting organizations and individuals through email, text messaging, telephone calls, social media platforms, and collaboration software. In addition, sophisticated social engineering techniques, including deepfake technology, enable realistic voice phishing (vishing), while behavioural analysis facilitates more effective psychological manipulation (Wiemken et al., 2025). The review further identifies several evasion techniques that can undermine phishing detection systems, including adversarial attacks against machine learning models, domain adaptation, and the integration of zero-day exploit methods (Yamin et al., 2021).

Detection Technique Performance

The reviewed studies indicate that machine learning and deep learning approaches generally achieve better detection performance than traditional phishing detection methods. Traditional approaches, such as blacklists and heuristic methods, showed limited effectiveness against novel attacks, with an average detection rate of 65–75%. Machine learning techniques, including Random Forest, LightGBM, and ensemble methods, achieved accuracy levels of 93–99% in controlled environments (Gupta et al., 2021; Zhou et al., 2023). Deep learning approaches based on CNN, RNN, transformer, and GCN architectures achieved accuracy levels ranging from 95% to 99%, although their higher performance often required greater computational resources (Haq et al., 2024; Maneriker et al., 2020). Multimodal approaches that combine URL, HTML, and behavioural features demonstrated strong potential for improving detection and resistance to evasion attempts, as reported by Tian et al. (2025).

The Adaptability of Phishing Threats

Phishing threats are remarkably adaptable due to three interconnected factors: technological, human, and economic factors. The review found that attackers increasingly exploit new technologies, including artificial intelligence and automated systems, in ways that can challenge existing defence mechanisms. From a technological perspective, attackers have developed new methods for bypassing standard linguistic analysis techniques through AI-generated content. Hao et al. (2024) documented a 1,265% increase in AI-generated phishing emails between 2023 and 2024, with these campaigns demonstrating significantly higher engagement rates. From a human perspective, individuals can be influenced through psychological manipulation

techniques, with Wiemken et al. (2025) experimentally demonstrating that emotional states, particularly anger, and low physiological arousal increase phishing susceptibility by 34–48%. The study further suggested that security awareness training should incorporate education on emotional states. Economically, the relatively low operational cost of phishing attacks, combined with their potential to generate substantial profits, enables attackers to sustain their activities. The number of phishing kits available on dark web markets has also reportedly increased by 50% since 2022, lowering the technical barrier for individuals with limited expertise to conduct phishing attacks.

Detection Technique Limitations

The current detection methods show accurate results when tested in controlled experiments. However, these methods encounter major problems that prevent their implementation in real-world situations.

Table 3: Comparison of Detection Technique Limitations

Technique Category	Strengths	Limitations	Deployment Challenges
Traditional (Blacklists/Heuristics)	Low false positives, interpretable	Poor against novel attacks, maintenance intensive	Real-time updating, coverage gaps
Machine Learning	High accuracy, adaptable to patterns	Feature engineering required, adversarial vulnerability	Cross-dataset generalization, explainability
Deep Learning	Automated feature learning, state-of-art accuracy	Black-box nature, computational cost	Resource requirements, interpretability barriers
Multimodal Approaches	Robustness, comprehensive analysis	Integration complexity, data requirements	System integration, performance optimization

Generalization Challenges: Rashid et al. (2024) proved that machine learning models which trained on one phishing dataset experience performance decrease between 15 to 30 percent when they test on other datasets. The "domain shift" problem exists because feature distributions between different domains show variations which create obstacles for actual system implementation. Adversarial Vulnerabilities: Phishing attacks have evolved to use adversarial methods which help attackers bypass machine learning security systems. To test detection systems researchers found that even small changes to URLs and email messages could decrease detection rates by 40% (Yamin et al., 2021).

Emerging Attack Vectors

The review has identified several emerging attack vectors that demand immediate attention. AI-powered spear phishing uses personal information obtained from social media platforms and public databases to create highly targeted attacks, with artificial intelligence being used to generate specific attack materials for selected targets (Birthriya et al., 2025). Collaboration tools such as Microsoft Teams, Google Chat, and Slack are also increasingly exploited for phishing activities because users tend to place greater trust in these platforms. Mobile-first attacks involve fraudulent SMS messages and malicious links designed to deceive mobile users, particularly those using mobile banking services. Deepfake voice technology represents another emerging threat, enabling attackers to generate spoofed calls that imitate the voices of executives or family members. In addition, IoT platform infiltration involves the exploitation of smart device interfaces and home networking systems to facilitate phishing attacks.

Future Research Directions

The systematic analysis identified several important directions for future research in phishing detection. One major area is Explainable Artificial Intelligence (XAI), as the black-box nature of many deep learning models can limit transparency and trust in their decisions, particularly in critical applications. Future research should therefore focus on developing more interpretable architectures that support transparent decision-making. Uddin and Sarker (2024) demonstrated the potential of combining LIME with transformer interpretation techniques, although wider application and standardized practices are still required. Another important direction is the development of adversarial-resilient models capable of maintaining their performance against increasingly sophisticated attacks. Future studies should examine approaches such as adversarial training, defensive distillation, and ensemble methods to improve model robustness under attack conditions (Divakaran & Oest, 2022). Multimodal detection systems also represent an important research direction, particularly systems capable of integrating URL structure, HTML content, visual similarity, user behaviour patterns, and metadata. The WebGuard++ system developed by Tian et al. (2025) represents an important advancement in this area, although further research is needed to enable the real-time processing of multiple data streams.

Future research should also place greater emphasis on human-centric defence strategies and the standardization of datasets and evaluation procedures. Combining technical detection systems with behavioural and security measures can improve phishing protection, particularly through the development of emotion-aware learning systems that adapt to users' psychological states (Huic & Hitzler, 2025), just-in-time interventions that provide security warnings at appropriate times and in effective formats (Greco, 2025), and personalized nudges based on individual vulnerability patterns. In addition, researchers face difficulties when comparing findings because standardized datasets and evaluation metrics remain limited. Future efforts should therefore establish well-organized and balanced phishing datasets containing diverse attack vectors (Hannouche & Harouache, 2020), adopt standardized evaluation measures that consider accuracy alongside robustness, latency, and interpretability, and develop open benchmark challenges to encourage consistent evaluation and further innovation in phishing detection.

CONCLUSION

This systematic review examined 26 studies published between 2020 and 2025 to assess the evolution of phishing threats, current detection techniques, emerging attack vectors, and future directions for effective defence. The findings show that phishing has evolved from conventional email-based deception into a more sophisticated and adaptive threat involving artificial intelligence, social engineering, multiple communication channels, adversarial techniques, and

emerging technologies. Machine learning and deep learning approaches generally demonstrated higher detection performance than traditional methods, with several reviewed studies reporting accuracy levels between 93% and 99% under controlled conditions. However, the review also shows that high performance in controlled environments does not necessarily translate into reliable real-world protection because of adversarial attacks, domain-shift and cross-dataset generalization challenges, computational requirements, and limited interpretability. These limitations indicate that accuracy alone is insufficient for evaluating the effectiveness of phishing detection systems.

The review therefore establishes the need for more robust and comprehensive phishing defence mechanisms that combine advanced machine learning and deep learning techniques with explainability, adversarial resilience, multimodal analysis, human-centred security strategies, and standardized datasets and evaluation metrics. Future research should focus on developing detection systems that can adapt to emerging phishing techniques while maintaining reliable performance across different datasets and attack conditions. Greater attention should also be given to explainable and interpretable models, real-time multimodal detection, behavioural factors, and standardized benchmarking to enable meaningful comparison of detection systems. Although the review was limited to English-language publications retrieved from selected academic databases and therefore may not represent all available research, the findings provide a useful basis for understanding the current state and future direction of phishing detection. Given the rapid evolution of phishing techniques, continuous research and evaluation remain necessary to develop effective, reliable, and adaptable defence mechanisms against emerging phishing threats.

REFERENCES

Alhogail, A., & Alsabih, A. (2021). Applying machine learning and natural language processing to detect phishing email. *Computers & Security, 110*, 102414. https://doi.org/10.1016/j.cose.2021.102414

Alkhalli, Z., Hewage, C., Nawaf, L., & Khan, I. (2021). Phishing attacks: A recent comprehensive study and a new anatomy. *Frontiers in Computer Science*, 563060. https://doi.org/10.3389/fcomp.2021.563060

Aslam, S., Aslam, H., Manzoor, A., Hui, C., & Rasool, A. (2023). AntiPhishStack: LSTM-based stacked generalization model for optimized phishing URL detection. *Symmetry, 15*(10), 1986. https://doi.org/10.3390/sym15101986

Birthriya, S. K., Ahlawat, P., & Jain, A. K. (2025). Detection and prevention of spear phishing attacks: A comprehensive survey. *Computers & Security, 151*, 104317. https://doi.org/10.1016/j.cose.2025.104317

Divakaran, D. M., & Oest, A. (2022). Phishing detection leveraging machine learning and deep learning: A review. *IEEE Security & Privacy, 20*(5), 86–95. https://doi.org/10.1109/MSEC.2022.3175225

Greco, F., Desolda, G., Buono, P., & Piccinno, A. (2025). Enhancing phishing defenses: The impact of timing and explanations in warnings for email clients. *Computers in Human Behavior Reports, 93*, 103982. https://doi.org/10.1016/j.chbr.2025.103982

Guddanti, D. S., Chiramdasu, R., & Uppuluri, M. G. (2025). A multi-algorithm approach for phishing uniform resource locator's detection. *IAES International Journal of Artificial Intelligence, 14*(1), 358–367. https://doi.org/10.11591/ijai.v14.i1.pp358-367

Gupta, B. B., Yadav, K., Razzak, I., Psannis, K., Castiglione, A., & Chang, X. (2021). A novel approach for phishing URLs detection using lexical based machine learning in a real-time environment. *Computer Communications, 175*, 47–57. https://doi.org/10.1016/j.comcom.2021.04.023

Hannousse, A., & Yahiouche, S. (2020). Towards benchmark datasets for machine learning based website phishing detection: An experimental study. *Mendeley Data, V2*. https://doi.org/10.17632/c2gw7fy2j4.2

Hao, W., et al. (2024). *Study on AI-generated spam emails*. Columbia University.

Haq, Q. E. U., Faheem, M. H., & Ahmad, I. (2024). Detecting phishing URLs based on a deep learning approach to prevent cyber-attacks. *Applied Sciences, 14*(22), 10086. https://doi.org/10.3390/app142210086

Li, W., Manickam, S., Chong, Y.-W., Leng, W., & Nanda, P. (2024). A state-of-the-art review on phishing website detection techniques. *IEEE Access, 12*, 3514972. https://doi.org/10.1109/ACCESS.2024.3514972

Li, Y., et al. (2024). Quantifying the multidimensional impact of cyber-attacks in digital financial services: A systematic literature review. *IEEE Access*.

Maneriker, P., Stokes, J. W., Lazo, E. G., Carutasu, D., Tajaddodianfar, F., & Gururajan, A. (2020). URLTran: Improving phishing URL detection using transformers. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2927–2931). IEEE. https://doi.org/10.1109/ICASSP40776.2020.9052978

Mouncey, E., & Ciobotaru, S. (2025). Phishing scams on social media: An evaluation of cyber awareness education on impact and effectiveness. *Journal of Economic Criminology, 7*, 100125. https://doi.org/10.1016/j.jeconc.2025.100125

Nadeem, M., Zahra, S. W., Abbasi, M. N., Arshad, A., Riaz, S., & Ahmed, W. (2023). Phishing attack, its detections and prevention techniques. *International Journal of Wireless Security and Networks, 1*(2), 13–25.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

Ramesh, G., Lokitha, R. B., Monisha, R. R., & Neha, N. S. (2023). Phishing detection system using random forest algorithm. *International Journal for Research Trends and Innovation, 8*(4), 511–513.

Rashid, F., Doyle, B., Han, S. C., & Seneviratne, S. (2024). Phishing URL detection generalisation using unsupervised domain adaptation. *Computer Networks, 245*, 110398. https://doi.org/10.1016/j.comnet.2024.110398

Shuaibu, H. T., Adewumi, S. E., & Yemi-Peters, V. I. (2025). Phishing detection in social media platforms using machine learning. *FUDMA Journal of Sciences*, *9*(11), 435–440. https://doi.org/10.33003/fjs-2025-0911-3944

Taha, M. A., Jabar, H. D. A., & Mohammed, W. K. (2024). A machine learning algorithm for detecting phishing websites: A comparative study. *Iraqi Journal for Computer Science and Mathematics, 5*(3), 275–286. https://doi.org/10.52866/ijcsm.2024.05.03.015

Tian, Y., Zhang, Y., Jia, Y., Sun, J., & Wang, Y. (2025). WebGuard++: Interpretable malicious URL detection via bidirectional fusion of HTML subgraphs and multi-scale convolutional BERT. *arXiv*. https://arxiv.org/abs/2506.19356

Uddin, M. A., & Sarker, I. H. (2024). An explainable transformer-based model for phishing email detection: A large language model approach. *arXiv*. https://arxiv.org/abs/2402.13871

Wiemken, M., Jeworutzki, A., Hildebrandt, K., & Putzar, L. (2025). Emotional manipulation in phishing emails: Experimental study of affective responses and human classification errors in a simulated email environment. In *Proceedings of the 18th ACM International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '25)* (pp. 583–589). Association for Computing Machinery. https://doi.org/10.1145/3733155.3736796

Yamin, M. M., Ullah, M., Ullah, H., & Katt, B. (2021). Weaponized AI for cyber attacks. *Journal of Information Security and Applications, 57*, 102722. https://doi.org/10.1016/j.jisa.2020.102722

Zhou, J., Cui, H., Li, X., Yang, W., & Wu, X. (2023). A novel phishing website detection model based on LightGBM and domain name features. *Symmetry, 15*(1), 180. https://doi.org/10.3390/sym15010180