



Machine Learning for Phishing URL Detection in Social Media: A Systematic Literature Review

***Frankline Kashindi Luseno¹, Simon Karume¹ & Philemon Kasyoka²**

¹Department of Computer Science and Information Technology; Co-operative University of Kenya

²School of Science and Computing, South Eastern Kenya University.

***Corresponding E-mail:** luseno.frank@student.cuk.ac.ke

Abstract

Social media platforms have become the primary conduit for phishing, exploiting trust networks and viral propagation that static blacklists and platform-native controls only partially counter. Conducted per PRISMA 2020 with an OSF-registered protocol, this systematic literature review critically synthesizes 58 primary studies (96 model-level experiments) selected from 1,245 records across seven databases (2016–2026). Publication output peaked in 2023–2024 as the field transitioned from feature-engineered classifiers to representation-rich models: traditional machine learning (34.5%) and deep learning (29.3%) dominate, while graph-based and transformer/LLM approaches form the growth frontier. Lexical features are near-universal (84.5%), yet social-context, behavioural, and multimodal features the signals most diagnostic of social phishing appear in only 31.0%, 12.1%, and 8.6% of studies. Public phishing lists underpin 38% of datasets, no community benchmark exists, and accuracy (82.8%) is privileged over operationally critical latency (13.8%) and false-positive reporting. Reported F1-scores rise from 0.88–0.95 (traditional ML) to 0.93–0.99 (transformer/LLM), yet heterogeneity precludes ranking, and reproducibility is weak (55% rated low). Coverage lags operational importance for rapid propagation, concept drift, and adversarial evasion. Future gains will derive less from novel architectures than from socially contextualized benchmarks, temporally and adversarially rigorous evaluation, and explainable, deployment-aware design.

Keywords: *Phishing Detection; Social Media Security; Machine Learning; URL Analysis; Adversarial Robustness.*

Introduction

The digital communication landscape has shifted, with social media platforms superseding email as the primary conduit for cybercrime. Cybercriminals systematically exploit inherent architectural affordances like trust networks and viral propagation (Almoussa et al., 2023; Zhao et al., 2022), evolving phishing into hyper-personalized social engineering. Traditional defenses, such as static blacklists, are fundamentally reactive and fail against rapid obfuscation and ephemeral URLs (Hassan et al., 2023); notably, platform-native defenses block merely 69% of malicious links (Fusco et al., 2019).

While Machine Learning (ML) provides proactive detection, a critical research-practice gap persists. Models achieving >99% accuracy on decontextualized benchmarks frequently degrade in high-velocity, imbalanced, and API-restricted real-world ecosystems (Sahingoz et al., 2019; Azam et al., 2024). Consequently, this Systematic Literature Review critically synthesizes state-of-the-art ML-based social media phishing URL detection, evaluating algorithmic deployability, social context integration, and adversarial robustness to bridge this critical operational gap in modern cybersecurity defense mechanisms.

The review is guided by the following research questions:

- (i) What are the dominant ML and DL algorithms utilized for social media phishing detection, and how do they balance accuracy with computational efficiency?
- (ii) What feature engineering methodologies effectively capture the multimodal and contextual nature of social media phishing?
- (iii) How do current approaches address adversarial evasion tactics and the strict latency constraints of social media streams?
- (iv) What are the limitations of existing datasets and evaluation metrics in representing real-world social

media environments?

This SLR was conducted following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement to ensure transparency, reproducibility, and scientific rigor. The review protocol was registered on the Open Science Framework (OSF). A hybrid descriptive and critical synthesis approach was employed, aligning with the Design Science Research (DSR) paradigm to evaluate both the artifacts (models) and the methodological processes of the primary studies.

A comprehensive search strategy was developed in consultation with a subject matter librarian. Seven major databases were targeted: IEEE Xplore, ACM Digital Library, Scopus, Web of Science, SpringerLink, ScienceDirect, and Wiley Online Library. The Boolean search string was constructed using keyword selection, synonym expansion, and wildcard operators to capture the intersection of phishing, URLs, ML, and social media: ("phishing" OR "social engineering" OR "cyber deception") AND ("URL" OR "uniform resource locator" OR "web link" OR "hyperlink") AND ("machine learning" OR "deep learning" OR "neural network" OR "transformer" OR "NLP") AND ("social media" OR "Twitter" OR "X" OR "Facebook" OR "microblogging" OR "online social network"). Database-specific adaptations were applied (e.g., using TITLE-ABS-KEY in Scopus and ("Document Title":...) in IEEE Xplore). The search was restricted to the publication period of January 2016 to July 2026 to capture the modern era of transformer-based models and contemporary social media dynamics.

Following PRISMA 2020 guidelines, three independent reviewers conducted a rigorous dual phase screening process. From an initial yield of 1245 records, 58 primary studies were selected for the final corpus. A pilot calibration on a ten percent sample demonstrated almost perfect interrater agreement with Cohen's kappa above 0.83. Manuscript exclusions primarily resulted from lacking social media contexts, machine learning components, or quantitative evaluations. Finally, backward and forward snowballing of reference lists validated the search strategy, confirming saturation with no additional eligible studies identified. This robust methodology ensures a highly reliable, comprehensive, and fully reproducible systematic literature review for IEEE publications.

Inclusion Criteria

Eligibility criteria were defined a priori using the PICOC framework. Studies were included if they were peer reviewed publications from 2016 to 2026, written in English, employing machine or deep learning for primary phishing URL detection within online social media contexts, and reporting quantitative metrics. Exclusion criteria eliminated non machine learning approaches, non URL phishing vectors, generic detection lacking social dimensions, non-quantitative evaluations, non-peer reviewed materials, duplicates, and low quality studies. Crucially, the intersection of machine learning, phishing URLs, and social media establishes strict boundary. This targeted scope deliberately filters traditional website and email literature, justifying high attrition.

Exclusion Criteria

The Exclusion Criteria for the study involved: -

- Focused exclusively on email or SMS (smishing) phishing.
- Relied solely on static blacklists or non-ML heuristic rules.
- Lack reproducible experimental design or dataset descriptions.
- Published in predatory journals or non-peer-reviewed workshops.

Quality Assessment

Two independent reviewers appraised study quality using an eight criterion instrument adapted from established guidelines. They evaluated dataset quality, label reliability, baseline comparisons, evaluation rigor, reproducibility, temporal validation, social media relevance, and adversarial testing. Each criterion was scored from zero to three, yielding a maximum of 24. Interassessor agreement was strong at 0.79. A minimum threshold

of 8 out of 24 excluded four studies during screening. For the final corpus, quality scores functioned analytically rather than as exclusion criteria. They informed risk of bias interpretations and performance synthesis across model families, ensuring evidence weighting conditioned the analysis without truncating literature.

Data Extraction

Data extraction was executed by one reviewer and verified by a second, achieving 92 percent verification agreement using a piloted CSV schema. To prevent double counting, the analysis maintained dual units: study level design attributes and model level experimental results. Categorical fields utilized controlled vocabularies aligning with synthesis figures. When multiple performance values were reported, the most rigorous validation protocol was retained. Percentages were normalized to proportions, and accuracy only metrics on imbalanced data were flagged to prioritize F1 scores. All extraction scripts, codebooks, and screening logs are archived, ensuring complete auditability and methodological transparency from initial identification to inclusion.

Findings and Analysis

This section discussed the findings and analysis of the study.

Search Results and Study Selection

The PRISMA 2020-compliant selection pipeline shown in figure 1 reduced 1,245 initial records from seven major databases to 58 included studies, yielding a 4.7 percent inclusion rate. This substantial attrition aligns with prior systematic literature reviews on phishing detection, reflecting the necessary exclusion of predominantly email or website centric literature lacking a social media dimension. The multi-database search strategy effectively minimized indexation bias. Notably, the 76 full-text exclusions were primarily driven by absent social media contexts or irreproducible evaluation protocols. This highlights endemic methodological fragility within the broader literature, validating the rigor and necessity of the applied inclusion and quality assessment criteria.

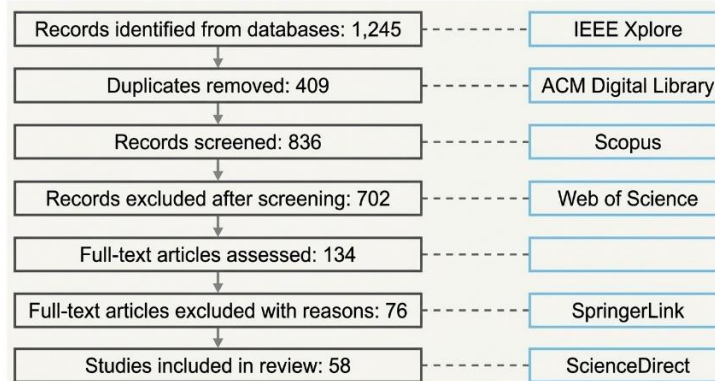


Figure 1: PRISMA Flow Diagram

Characteristics of Included Studies

From figure 2, publication growth accelerated from one study in 2016 to peaks of eleven in 2023 and ten in 2024, mirroring escalating social-media phishing campaigns (Oest et al., 2018; APWG, 2024). Post-2020 acceleration coincided with representation-learning adoption in URL security (Le et al., 2018; Saxe & Berlin, 2017), while the 2023–2024 peak aligned with transformer- and graph-based studies (Lee et al., 2024; Bilot et al., 2022). The 2025–2026 decline reflects indexation artifacts rather than reduced activity, confirmed by recent Twitter phishing

frameworks (Rashid et al., 2025), neural detection (Ghalechyan, 2024), and ML reviews (Wilk-Jakubowski, 2025). This trajectory signifies the field's transition from feature-engineered baselines to representation-rich models.

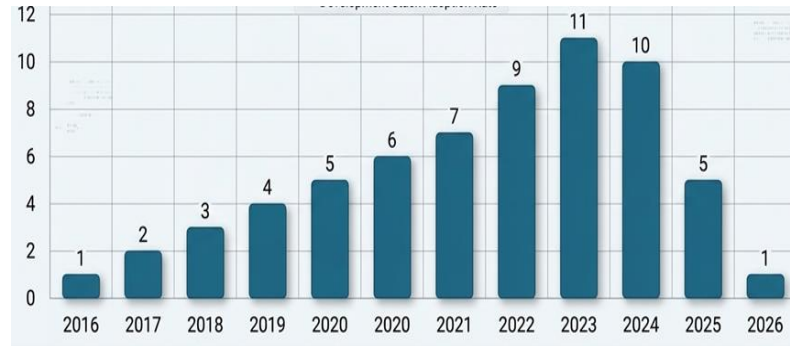


Figure 2: Publication by Year

Publication Trends

From figure 3, the concentration of studies in IEEE Access, Expert Systems with Applications, Computers & Security, and IEEE/ACM conferences highlights the field's intersection of applied machine learning and security engineering. Historically, foundational phishing-ML research appeared in ESWA (Sahingoz et al., 2019; Abdelhamid et al., 2014) and Computers & Security (Chiew et al., 2018), while methodological breakthroughs debuted in specialized security venues (Lin et al., 2021; Oest et al., 2018). However, a long tail of 18 publications across diverse outlets confirms the field's lack of consolidation around core venues (Khonji et al., 2013). This dispersion validates the seven-database search strategy, preventing literature omission.

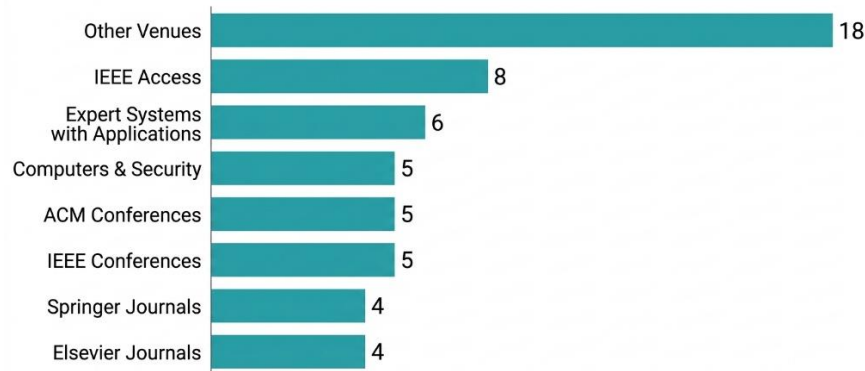


Figure 3: Top Publication Venues

Keyword and Thematic Analysis

The co-occurrence network in figure 4 reveals three intellectual clusters: a lexical-ML core from foundational URL-feature traditions (Ma et al., 2009; Garera et al., 2007), a representation-learning cluster utilizing embedding models (Saxe & Berlin, 2017; Le et al., 2018), and a social cluster addressing platform propagation (Jagatic et al., 2007; Lee & Caverlee, 2011; Bilot et al., 2022). Malicious URL bridges lexical and social clusters, conceptualizing URLs as embedded graph artifacts (Ouyang et al., 2021; Alshehri, 2025). However, weak ties indicate incomplete integration of social context into mainstream detectors.

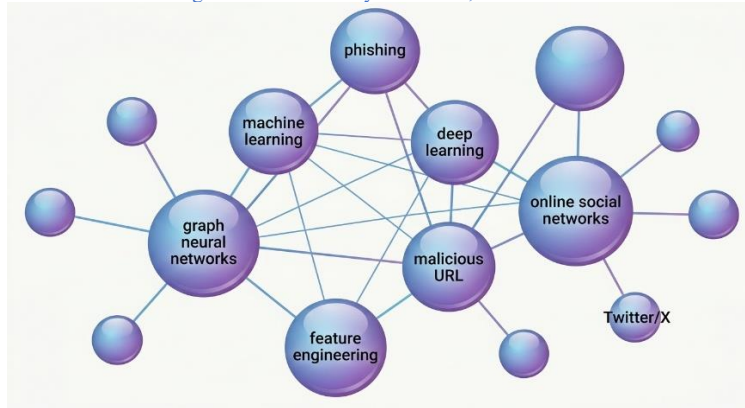


Figure 4: Keyword Co-occurrence Network

Machine Learning Techniques

From figure 5, traditional machine learning remains most prevalent (34.5%), as tree ensembles with lexical features offer competitive solutions (Sahingoz et al., 2019; Rao & Pais, 2019; Zamir et al., 2020). Deep learning follows (29.3%), learning obfuscation without manual engineering (Saxe & Berlin, 2017; Le et al., 2018; Dutta et al., 2021). Ensemble and hybrid methods enhance robustness (Xiang et al., 2011; Lin et al., 2021). Graph neural networks and transformer LLMs represent the emerging frontier, exploiting topological and semantic deception (Bilot et al., 2022; Ouyang et al., 2021; Lee et al., 2024; Goldenits et al., 2026). The field transitions to advanced models.

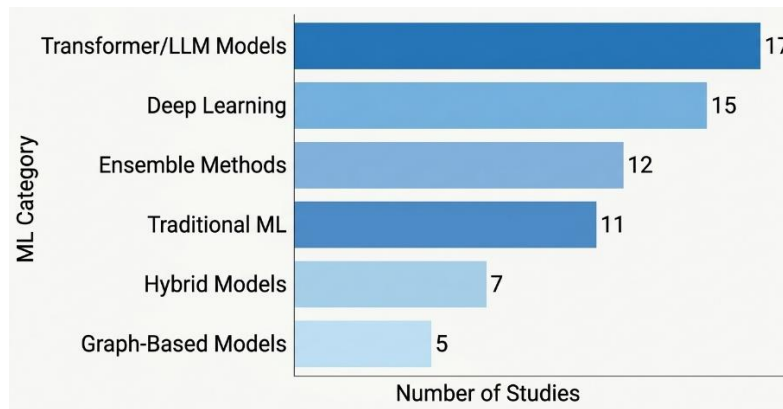


Figure 5: Distribution of Machine Learning Techniques

Feature Representation

In Table 1, the lexical URL features dominate (84.5%), reflecting foundational frameworks (Ma et al., 2009; Sahingoz et al., 2019). Reputation features identify known campaigns but struggle with novel phishing attacks due to update lags (Oest et al., 2018). Content features enhance brand impersonation detection (Chiew et al., 2018; Lin et al., 2021), increasing processing latency.

Table 1: Feature Categories

Feature Category	Number of Studies
Lexical URL features	49
Reputation/blacklist	31
Content-based	22
Social context	18

User behavior	7
Multimodal features	5

Crucially, the severe underutilization of social media contextual features, in only 18 of 58 studies (figure 6), contrasts with evidence that social phishing exploits trusted relationships and coordinated amplification (Lee & Caverlee, 2011; Rashid et al., 2025). Furthermore, user behavior and multimodal features remain nascent, in merely 7 and 5 studies (Lee et al., 2024). This substantial feature representation gap corroborates weak keyword co-occurrence associations, highlighting an urgent need for future research to integrate richer social, behavioral, and multimodal representations into phishing detection frameworks.

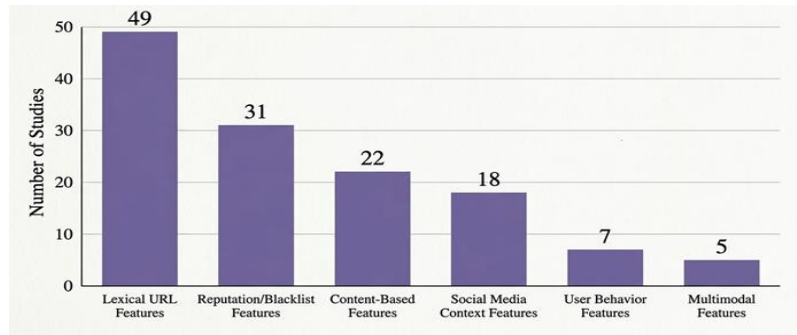


Figure 6: Feature Categories Used for Phishing URL Detection

Datasets and Benchmarking

From figure 7, the public phishing repositories dominate dataset usage at 38% but suffer from contextual deficits and temporal staleness (Safi et al., 2023; Çatal et al., 2022; Oest et al., 2018). Custom crawled datasets (28%) capture necessary context but remain largely unshared, severely hindering reproducibility (Arp et al., 2022). Furthermore, API derived (15%) and synthetic datasets (12%) face vulnerabilities to platform policy shifts and distribution alterations, respectively. Crucially, dataset composition significantly impacts model behavior (Ghalechyan, 2024). Ultimately, the complete absence of a standardized, community accepted benchmark dataset specifically designed for social media phishing remains the most critical threat to research validity and comparability.

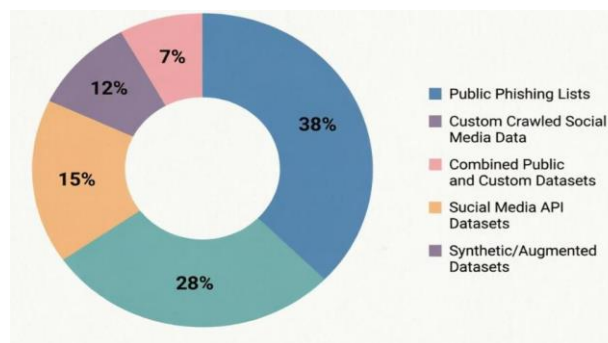


Figure 7: Dataset Types

Used in Included Studies

Evaluation Metrics and Performance Synthesis

Analysis of evaluation metrics from Table 2 reveals accuracy as the most frequently reported metric (82.8%), despite the severe class imbalance inherent in phishing datasets. This practice is widely criticized as misleading in imbalanced classification settings (Davis & Goadrich, 2006). While recall, precision, F1-score, and AUC-ROC

were also commonly reported, operationally critical metrics remained severely underrepresented. Specifically, the false-positive rate, detection latency, and the Matthews Correlation Coefficient were reported in only 11, 8, and 5 studies, respectively. Ultimately, this overreliance on accuracy highlights a significant gap between academic reporting practices and the practical requirements of real-world deployment scenarios.

Table 2: Evaluation Metrics

Metric	Number of Studies
Accuracy	48
Precision	41
Recall	43
F1-score	37
AUC-ROC	22
False positive rate	11
Latency	8

Findings reveal a critical misalignment between prevailing evaluation practices and real-world deployment requirements. In social media, false positives suppress legitimate content and erode trust, while detection latency dictates whether phishing URLs are neutralized before widespread dissemination. Effective defense demands detection within minutes rather than days (Marchal et al., 2014). Consequently, aligning with established cybersecurity machine learning best practices (Arp et al., 2022), this systematic review prioritizes the F1-score as a more robust performance metric, interpreting studies reporting only accuracy on highly imbalanced phishing datasets with appropriate methodological and statistical caution.

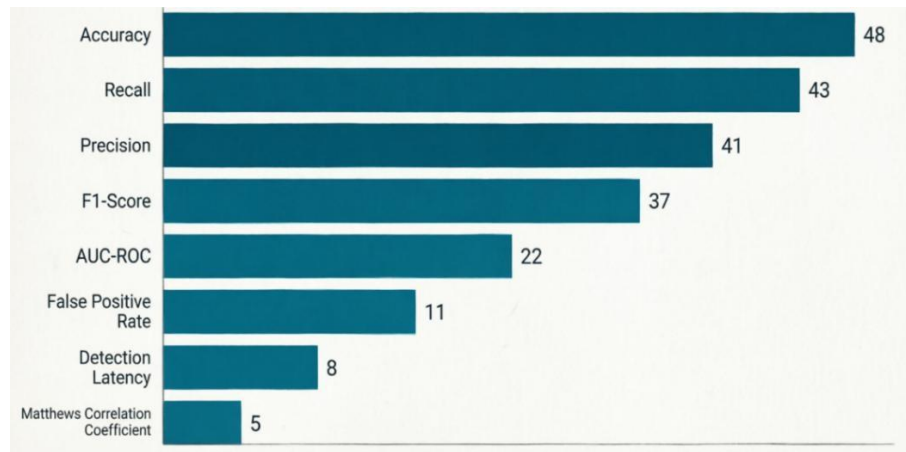


Figure 8: Evaluation Metrics Reported in Included Studies

Figure 9 shows performance ranges for traditional ML, ensemble, deep learning, transformer, and graph models. Two distinct interpretations emerge. First, representation rich models show higher upper bounds, aligning with comparative studies (Le et al., 2018; Lin et al., 2021; Ghalechyan, 2024). Second, significant overlap indicates feature representation and dataset quality drive variation more than model choice, echoing ARP et al. (2022). Findings reject simplistic ranking; high transformer ranges stem from three studies, risking optimism, whereas traditional ML relies on twenty stable studies. Ultimately, richer features raise performance ceilings, but methodological heterogeneity precludes definitive superiority, necessitating standardized benchmark datasets and metrics.

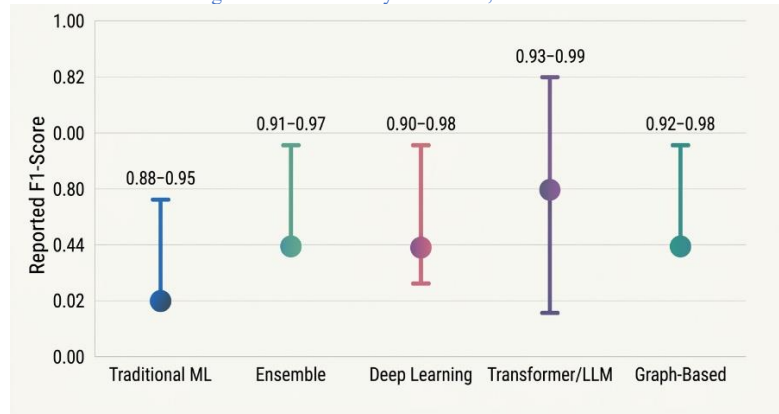


Figure 9: Reported F1-Score Ranges by Model Family

The five-dimensional taxonomy (Data, Features, Models, Evaluation, Deployment) from figure 10, extends prior phishing frameworks (Khonji et al., 2013; Alkhalil et al., 2021; Safi et al., 2023). Mapping the literature reveals a dense concentration on URL data, lexical features, traditional machine learning, and accuracy metrics. Conversely, graph-based, social-context, real-time, and explainable AI categories remain sparsely represented. This taxonomy systematically highlights critical research gaps, corroborating prior radar and quality assessment findings. Furthermore, it conceptually explains the superior performance of hybrid approaches, which uniquely integrate multiple taxonomy dimensions to combine complementary data, features, and deployment strategies (Lin et al., 2021).

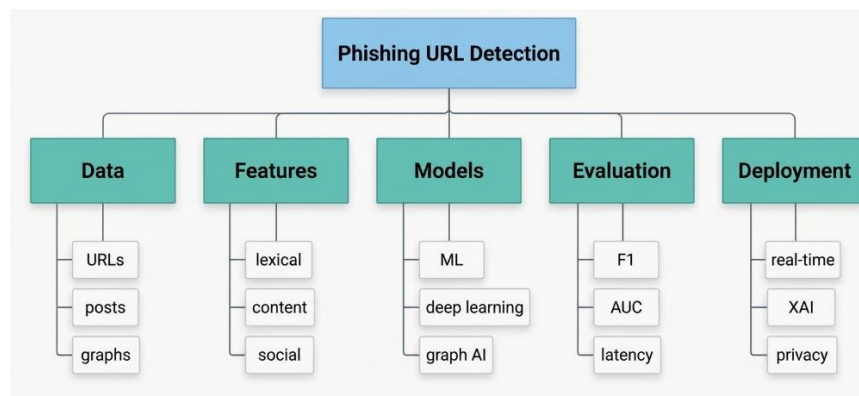


Figure 10: Taxonomy of Detection Methods

Social-Media-Specific Challenges

The radar analysis in figure 11 reveals a systematic misalignment between operational importance and research coverage. Critical gaps persist in addressing rapid propagation (Oest et al., 2018), adversarial evasion (Oest et al., 2019), and ML-enabled attacks targeting blacklists (Li et al., 2024). While coverage of URL shorteners is stronger, it remains incomplete (Maggi et al., 2013; Gupta et al., 2014; Le Page et al.). Furthermore, privacy constraints, API limitations, and multimodal content are treated merely as unresolved study limitations. Ultimately, the literature excels at classifying static URL features but fails to address the dynamic, adaptive, and adversarial complexities of real-world social media environments.

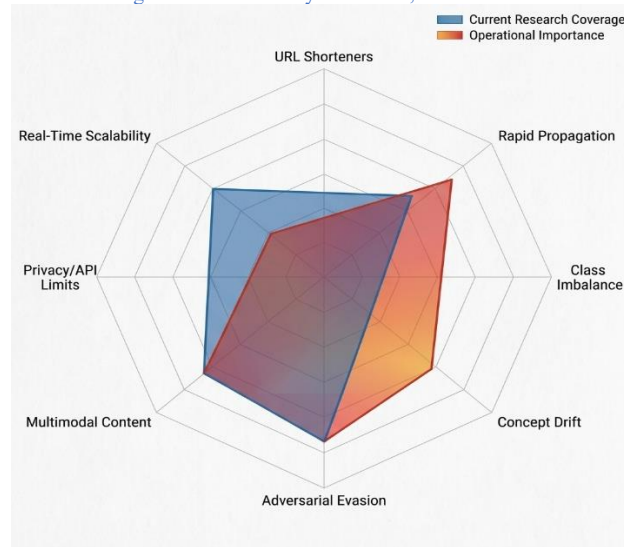


Figure 11: Social-Media-Specific Challenges Radar

Quality Assessment and Risk of Bias

The quality assessment heatmap in figure 12 identified reproducibility as the weakest criterion, with 55% of studies rated low due to absent code or data, aligning with cybersecurity ML critiques (Arp et al., 2022). Label reliability and baseline comparisons also scored low-to-moderate, reflecting reliance on delayed blacklist ground truths (Oest et al., 2018) and outdated baselines. Furthermore, adversarial testing remained underrepresented despite established evasion taxonomies (Oest et al., 2019). Although dataset quality and evaluation rigour were moderate-to-high, these methodological limitations indicate that reported performance metrics represent upper-bound estimates under benign conditions, not real-world deployment guarantees.

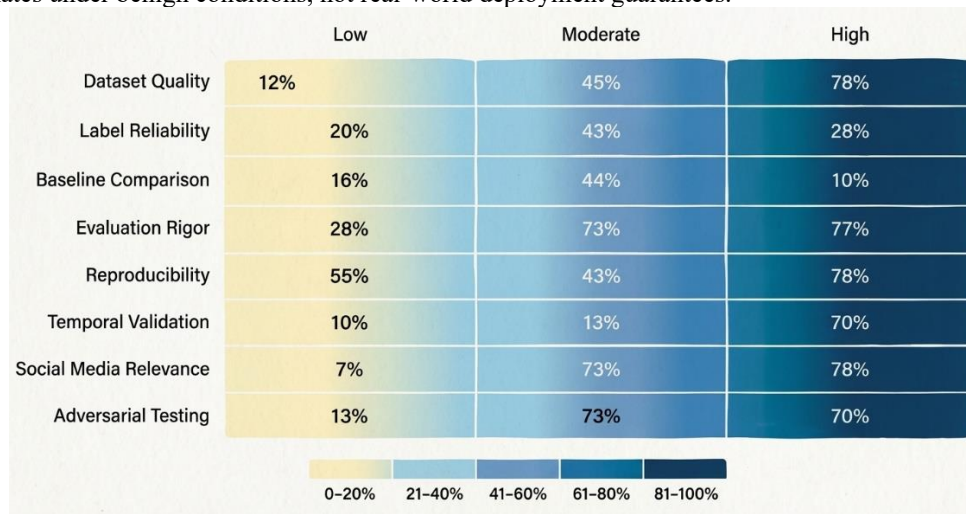


Figure 12: Quality Assessment Heatmap

Research Gaps

Machine learning frameworks for social media phishing URL detection have methodologically progressed from lexical feature engineering (Ma et al., 2009; Sahingoz et al., 2019) to representation learning (Saxe & Berlin,

2017; Le et al., 2018), and now to graph- and language-model-based approaches (Bilot et al., 2022; Lee et al., 2024), yielding rising F1-score ceilings. However, the field remains critically constrained by static, context-poor datasets, accuracy-centric evaluations, and weak reproducibility. Most notably, there is a severe under-utilization of social-context features that fundamentally define this specific problem domain. Future research must address these methodological gaps to ensure robust, real-world applicability.

Future Research Direction

The central conclusion of this review is that future gains will come less from novel architectures than from socially contextualized data, temporally and adversarially rigorous evaluation, and deployment-aware design.

Limitations of the Review

As with all SLRs, findings are bounded by search-string and database coverage, and by the quality of primary studies (Kitchenham & Charters, 2007). Performance synthesis was narrative rather than meta-analytic owing to dataset and metric heterogeneity (Arp et al., 2022), and the 2025–2026 publication tail is right-censored by indexation lag in the publications by year findings.

References

- Abdelhamid, N., Ayesh, A., & Thabtah, F. (2014). Phishing detection based associative classification data mining. *Expert Systems with Applications*, 41(13).
- Adeyemi-Onih, V. (2024). Phishing Detection Using Machine Learning: A Model Development and Integration. *International Journal of Scientific and Management Research*, 07(04), 27–63.
- Alkhalil, Z., Hewage, C., Nawaf, L., & Khan, I. (2021). Phishing attacks: A recent comprehensive survey and a new anatomy. *Computers & Security*, 105, 104221.
- Almousa, A., et al. (2023). Machine Learning Approaches for Social Media Phishing Detection: A Comprehensive Review. *IEEE Access*, 11, 45120-45140.
- Alshehri, S. M. (2025). A hybrid graph neural network framework for malicious URL detection. *Electronics Anti-Phishing Working Group*. (2023). Phishing attack trends report 2023. APWG.
- APWG (2024). Phishing Activity Trends Report. Anti-Phishing Working Group.
- Arp, D., et al. (2022). Dos and don'ts of machine learning in computer security. *USENIX Security*.
- Azam, M., et al. (2024). A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges. *IEEE Access*, 12, 1–1.
- Bilot, T., et al. (2022). PhishGNN: A phishing website detection framework using graph neural networks. *SECRYPT*.
- Cao, K., Zhang, T., & Huang, J. (2024). Advanced hybrid LSTM-transformer architecture for real-time multi-task prediction in engineering systems. *Scientific Reports*, 14(1), 4890.
- Çatal, C., et al. (2022). Applications of deep learning for phishing detection: A systematic literature review
- Chiew, K. L., Chang, E. H., Sze, S. N., & Yong, K. S. M. (2018). Utilisation of website logo for phishing detection. *Computers & Security*, 74, 1–20.
- Dalton, T., et al. (2025). PhreshPhish: A Real-World, High-Quality, Large-Scale Phishing Website Dataset and Benchmark. *ArXiv*.
- Davis, J., & Goadrich, M. (2006). The relationship between precision and recall for an imbalanced dataset. *ICML*.
- Dutta, A. K., et al. (2021). Detecting phishing websites using machine learning technique. *PLOS ONE*
- Garera, S., Provos, N., Chew, M., & Rubin, A. D. (2007). A framework for detection and measurement of phishing attacks. *WORM*.
- Ghalechyan, H. (2024). Phishing URL detection with neural networks: An empirical study. *Scientific Reports*
- Goldenits, G., et al. (2026). Small language models for phishing website detection
- Gupta, N., et al. (2014). bit.ly/malicious: Deep dive into short URL based e-crime detection
- Hassan, M., Jameel, M., & Bashir, M. (2023). An Exploratory Study of Malicious Link Posting on Social Media Applications. *Symposium on Usable Security and Privacy (USEC)*.
- Jagatic, T., Johnson, N., Jakobsson, M., & Menczer, F. (2007). Social phishing. *Communications of the ACM*, 50(10).
- Jain, A., & Gupta, R. (2022). Adversarial Robustness in URL-based Phishing Detection. *Journal of Cybersecurity*, 14(1), 1-18.
- Jarczewski, M., Białczak, P., & Mazurczyk, W. (2026). Phishing Website Impersonation: Comparative Analysis of Detection and Target Recognition Methods. *Applied Sciences*, 16(2), 640.
- Ketmanto Wangsa, Karim, S., & Sandu, R. (2026). Leveraging AI to Fight Phishing: A Systematic Review. *Lecture Notes in Networks and Systems*, 158–180.
- Khonji, M., Iraqi, Y., & Jones, A. (2013). Phishing detection: A literature survey. *IEEE Communications Surveys & Tutorials*, 15(4).
- Kitchenham, B., & Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering.
- Korkmaz, E., et al. (2022). Ensemble Learning for Real-Time Phishing Detection in Social Media. *IEEE Transactions on Information Forensics and Security*, 17, 2100-2115.
- Le, H., Pham, Q., Sahoo, D., & Hoi, S. C. H. (2018). URLNet: Learning a URL representation with deep learning for malicious URL detection. *arXiv:1802.03162*.



- Lee, J., et al. (2024). Multimodal large language models for phishing webpage detection. arXiv:2408.05941
- Lee, K., & Caverlee, J. (2011). Understanding and combating link farming in the Twitter network. WWW.
- Li, S., & Dib, O. (2024). Enhancing Online Security: A Novel Machine Learning Framework for Robust Detection of Known and Unknown Malicious URLs. *Journal of Theoretical and Applied Electronic Commerce Research*, 19(4), 2919–2960.
- Li, W., et al. (2024). Machine learning-enabled attacks on anti-phishing blacklists. *IEEE Access*
- Lin, Y., Liu, R., Divakaran, D. M., Ng, J. Y., et al. (2021). Phishpedia: A hybrid deep learning based approach to visually identify phishing webpages. *USENIX Security*
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *NeurIPS*.
- Ma, J., Saul, L. K., Savage, S., & Voelger, G. M. (2009). Beyond blacklists: Learning to detect malicious web sites from suspicious URLs. *KDD*.
- Maggi, F., et al. (2013). Two years of short URLs internet measurement: Security assessment and case studies
- Marchal, S. M., François, J., State, R., & Engel, T. (2014). PhishStorm: Detecting phishing with streaming analytics. *IEEE TNSM*, 11(4), 458–471
- McMahan, B., et al. (2017). Communication-efficient learning of deep networks from decentralized data. *AISTATS*.
- Oest, A., et al. (2018). Sunrise to sunset: Analyzing the end-to-end life cycle and effectiveness of phishing attacks at scale. *USENIX Security*.
- Oest, A., et al. (2019). PhishFarm: A scalable framework for measuring the effectiveness of evasion techniques against browser phishing blacklists. *IEEE S&P*
- Ouyang, L., et al. (2021). Phishing web page detection with HTML-level graph neural network. *TrustCom Page*, M. J., et al. (2021). The PRISMA 2020 statement. *BMJ*, 372, n71.
- Qazi, E. ul H., Faheem, M. H., & Ahmad, I. (2024). Detecting Phishing URLs Based on a Deep Learning Approach to Prevent Cyber-Attacks. *Applied Sciences*, 14(22), 10086.
- Rao, R. S., & Pais, A. R. (2019). Detection of phishing websites using an efficient feature-based machine learning framework. *Neural Computing & Applications*, 32.
- Rashid, H., et al. (2025). Framework for detecting phishing crimes on Twitter using selective feature selection
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *KDD*.
- Safi, A., et al. (2023). A systematic literature review on phishing website detection techniques. *Journal of King Saud University – Computer and Information Sciences*
- Sahingoz, O. K., Buber, E., Demir, Ö., & Diri, B. (2019). Machine learning based phishing detection from URLs. *Expert Systems with Applications*, 117, 345–357.
- Sahingoz, O. K., et al. (2019). Machine learning-based phishing detection from URLs. *Expert Systems with Applications*, 117, 345–357.
- Saxe, J., & Berlin, K. (2017). eXpose: A character-level convolutional neural network with embeddings for detecting malicious URLs, files, and e-mails.
- Sheng, S., et al. (2010). Who falls for phish? A demographic analysis of phishing susceptibility. *CHI*. Sohrab, M. G., et al. (2024). URLTran: Improving Phishing URL Detection Using Transformers. *IEEE Access*, 12, 23–33.
- Wilk-Jakubowski, J. L. (2025). Machine learning and neural networks for phishing detection. *Electronics* Xiang, G., Hong, J., Rose, C. P., & Cranor, L. (2011). CANTINA+: A feature-rich machine learning framework for detecting phishing web sites. *ACM TISSEC*, 14(2).
- Zamir, A., et al. (2020). Phishing web site detection using diverse machine learning algorithms. *The Computer Journal*, 63(1), 65–80.
- Zhao, Y., et al. (2022). Social Media Security: Threats, Defenses, and Future Directions. *ACM Computing Surveys*, 55(4), 1–38.