



International Journal of Research Publication and Reviews

Journal Homepage: www.ijrpr.com ISSN 2582-7421

Analyzing Applicability of Phishing Analysis and Prevention Techniques

Khushi Kumari¹, Mr. Nilesh Khare², Kritika Yadav³

¹⁻³ School of Computer Applications, Babu Banarasi Das University, Lucknow, Uttar Pradesh, India

DOI : <https://doi.org/10.5281/zenodo.23556037>

ABSTRACT

Phishing protection must operate across heterogeneous attack channels, changing adversarial tactics, and stringent deployment constraints. Although recent studies report substantial gains from machine learning, deep learning, transformer-based language models, and ensemble classifiers, predictive accuracy alone does not establish whether a detector is appropriate for a particular cybersecurity setting. This article critically synthesizes fifteen peer-reviewed studies published during 2022–2026, together with authoritative authentication and e-mail security standards, to examine the applicability of conventional and artificial-intelligence-based phishing analysis and prevention techniques. A structured taxonomy distinguishes reputation and rule systems, lexical and structural URL analysis, e-mail semantics, sequence models, transformer architectures, ensembles, continual learning, and explainability mechanisms. This article introduces a context-aware applicability framework that separates minimum operational requirements from multi-criteria comparison of security efficacy, cross-source generalization, false-alarm burden, latency, resource cost, interpretability, dependency, maintainability, and preventive actionability. A reproducible evaluation protocol is specified for public URL and e-mail corpora, with safeguards against duplicate and domain leakage, temporal drift, misleading class prevalence, and incomplete end-to-end latency accounting. The evidence supports conditional rather than universal selection: inexpensive lexical models are attractive at browser and mobile endpoints; contextual text analysis adds value at e-mail gateways; and ensemble, adaptive, and explainable components are most useful when their operational overhead can be justified. The proposed defense-in-depth architecture couples detection to warnings, quarantine, phishing-resistant authentication, reporting, and incident response.

Keywords— Phishing; cyber threat intelligence; machine learning; deep learning; concept drift; security evaluation.

1. INTRODUCTION

Phishing is a form of social engineering in which an attacker impersonates a legitimate institution, person, service, or web interface to induce disclosure of sensitive information or execution of a harmful action. Modern delivery mechanisms extend beyond e-mail to websites, shortened links, messaging applications, mobile notifications, social platforms, and QR-code mediated redirection. Technical features such as deceptive subdomains, visually confusable identifiers, page cloning, and legitimate hosting infrastructure are combined with urgency, authority claims, and personalization. Consequently, a phishing detector must interpret both the digital artifact and the context in which a user encounters it.

The traditional defensive perimeter relies on blocklists, domain reputation, signature matching, URL and header heuristics, and user reports. These controls can be inexpensive, interpretable, and fast, but recognition depends on observable indicators already known to the system. Learning-based models seek to discover patterns that transcend individual signatures. In the URL domain, feature-based classifiers and character-sequence networks can operate before a web page is fully rendered; in e-mail, contextual natural-language models examine intent, wording, and relationships that keyword rules frequently miss [1]–[5].

The principal challenge is no longer whether a technique can discriminate two classes in a laboratory benchmark. Operational performance also depends on the composition and recency of training data, the prevalence of attacks, processing delay, device resources, external-service availability, explanation quality, and the reliability of an intervention. Continual-learning results demonstrate deterioration of static detectors under changing attack patterns [6]. Multilingual evaluations reveal that aggregate performance can obscure harmful false alarms for underrepresented languages [7]. Browser and mobile studies show that an apparently strong classifier must also meet the latency and integration requirements of the point of decision [2], [8].

This article therefore addresses a specific research question: under what technical and organizational conditions are different phishing analysis and prevention techniques practically applicable? The contribution is a critical synthesis of the evidence and a Context-Aware Phishing Applicability (CAPA) framework that translates model assessment into deployment choices. Its originality is methodological and integrative, rather than the claim that a new classifier has already been trained or experimentally validated.

2. RESEARCH DESIGN AND SCOPE

2.1. Review strategy and research questions

The paper uses a structured, critical narrative review rather than claiming an exhaustive systematic review or meta-analysis. Studies were organized by input modality, detector family, deployment environment, and operational measures. Because the source papers differ in sampling, labels, feature visibility, train-test partitioning, and hardware, their headline metrics are not statistically pooled or treated as a common leaderboard.

Four guiding questions structure the analysis: (RQ1) Which evidence types and algorithms best support URL-, webpage-, and e-mail-oriented detection? (RQ2) Which evaluation criteria explain the gap between benchmark performance and real deployment? (RQ3) How should resource constraints, drift, and false-alarm costs shape technique selection? (RQ4) How can a detector's output be linked to preventive controls without assuming that classification alone prevents compromise?

2.2. Threat and operational boundaries

The considered attacker controls deceptive message text, a phishing URL, and potentially a landing page or lookalike domain; may rapidly move infrastructure or alter wording; and may seek credentials, authentication approval, financial transfer, or malicious code execution. The defender may observe only the raw URL, or may additionally have headers, page artifacts, hosting intelligence, and organizational context. The review distinguishes a maliciousness signal from proof of compromise: a suspicious URL score does not establish that credentials were exposed, and authenticated sender infrastructure does not guarantee benign message content.

A high-confidence online blocking workflow and a lower-confidence analyst-review workflow also have different failure costs. An erroneous browser block impedes a legitimate user, whereas a missed targeted phishing message can create a disproportionately serious incident. This asymmetry motivates explicit reporting of false-positive and false-negative rates and context-specific intervention thresholds rather than maximization of accuracy alone.

3. TAXONOMY OF PHISHING ANALYSIS METHODS

Detection families are differentiated by what they observe, how they generalize, and which external resources they require. The following taxonomy is a functional classification, not a ranking of inherent quality. Hybrid systems may combine multiple families; prevention controls are addressed separately because they act after or independently of a classification decision

3.1. Reputation and heuristic analysis

Known-indicator services provide a valuable first-stage filter because they can block a previously reported malicious destination before webpage retrieval. Transparent heuristics further detect IP-address hosts, excessive tokenization, unusual encodings, deceptive hostnames, and credential-oriented terms. Yet no single token is decisive: long URLs are sometimes legitimate; encrypted HTTPS pages can still be malicious; and attackers can place phishing content on compromised legitimate domains. A rule-based output is accordingly most defensible as one signal in a composite workflow, with update frequency and query availability recorded as operational dependencies.

3.2. Feature-based and sequence learning

Logistic regression and linear support vector machines provide comparatively reproducible baseline models for lexical URL or TF-IDF text vectors. Decision trees and random forests can learn nonlinear feature interactions and offer a useful foundation for feature attribution. Neural approaches instead learn character or token patterns using convolutional and recurrent layers. Zhou et al. combine convolutional attention, BiLSTM, and multiscale feature handling for malicious URLs [4]; Ghalechyan et al. demonstrate repeated validation on incoming URL observations, reinforcing the importance of evaluating beyond a static split [1]. These advantages must be assessed against preprocessing cost and the actual inputs available at prediction time.

3.3. Language models and contextual signals

E-mail phishing frequently requires semantic interpretation because sender claims, payment requests, threats, and credential prompts cannot always be identified from a URL alone. Work on language-pack-tuned transformers emphasizes nonstandard spelling and e-mail language patterns [3], while hybrid attention architectures combine BERT representations with convolutional and recurrent processing [5]. A language model can be appropriate for a gateway with server-side compute, but it may be unnecessarily expensive for local URL checks, and its semantic decisions remain vulnerable to distribution shift, adversarial wording, privacy constraints, and incomplete context.

3.4. Ensembles, adaptation, and interpretation

The mobile-oriented Phish-Jam system shows how handcrafted features, transformer representations, and ensemble learning can be integrated for a constrained application [2]. Gandotra et al. show that classifier integration can be more effective when base learners are selected by error behavior rather than combined indiscriminately [13]. Continued retraining is necessary when attack distributions change, but naive online adaptation risks catastrophic forgetting, feedback loops, and learning from adversarial labels [6]. Few-shot approaches may reduce the delay before novel attacks are recognized [14]; their practical benefit must be demonstrated on genuinely unseen campaigns. SHAP-based explanations can help an analyst inspect the contribution of URL-derived features [15], although explanations are not, by themselves, guarantees of correctness or causal validity.

4. CRITICAL SYNTHESIS OF PUBLISHED EVIDENCE

The examined literature indicates growing attention to deployment, temporal generalization, and explainability. Piracema provides a substantial snapshot-oriented research resource with historical technical metadata, highlighting the importance of provenance and time-aware corpus construction [16]. Ejaz et al. directly examine model deterioration across time [6]. Ghalechyan et al. use recurrent validation with newer URL records rather than relying exclusively on a single historical benchmark [1]. Komosny shows why multilingual or regionally localized benign webpages matter when calculating false alarms [7].

Several recent studies report strong classification scores, but the values in Table 2 are individual authors' reported results under heterogeneous conditions. They must not be interpreted as measurements from the same corpus or computational environment. For example, a highly balanced benchmark can produce an attractive F1-score without establishing a tolerable false-alert rate at low attack prevalence. Similarly, a latency figure may omit feature acquisition, tokenization, model loading, network requests, and user-interface rendering.

Table 2. Selected results reported in source publications (not directly comparable)

Study	Technique / context	Reported finding	Cross-study caution
Ghalechyan et al. (2024) [1]	URL neural classifiers	~97% mean daily validation accuracy	Mixed-source labels and production setting
Rao et al. (2025) [2]	Mobile hybrid ensemble	98.93% accuracy (reported)	Mobile implementation; different corpus
Abiramasundari and Ramaswamy (2025)	E-mail transformer + recurrent models	95.47% accuracy (reported)	Specialized ransomware-phishing task
Hosseinzadeh et al. (2025) [5]	BERT-CNN-GRU e-mail model	96.8% accuracy; 96.3% F1 (reported)	Different corpus, tuning, and class balance
Komosny (2025) [7]	Language-aware webpage URL analysis	~99% accuracy on 12 countries (reported)	Local-language case and prevalence
Dandotiya et al. (2026) [8]	ML browser-extension detection	98.7% accuracy; MCC 0.96 (reported)	Browser implementation-specific assessment
Gandotra et al. (2026) [13]	Error-ranked classifier integration	Weighted F-measure 0.984 (reported)	Benchmark-specific ensemble comparison

The evidence also exposes interpretive limitations. Rawat et al. examine feature reduction and runtime constraints in a hybrid phishing framework [17], while Gobinath and Manikandan assess optimized URL intelligence with attention to false positives and delay [18]. These are promising directions, but reductions in feature count or parameter count do not automatically imply lower total latency, energy consumption, or maintenance effort. Quantum-enhanced methods in particular require a distinction between simulation-based claims and deployability on ordinary organizational infrastructure [19]. Hence, method selection should examine acquisition time, hardware, lifecycle cost, and failure modes, not merely report a model-size or headline-accuracy benefit.

Three recurring evidence gaps are especially important. First, many experiments use random partitions that can expose almost identical hostnames, domains, message templates, or campaign structures to both training and test sets. Second, the effects of current phishing prevalence and intervention cost are seldom integrated with classifier evaluation. Third, practical controls—safe browsing warnings, quarantine, identity assurance, and incident response—are often absent from detector-focused papers. These gaps motivate an applicability model centered on deployment context.

5. PROPOSED CONTEXT-AWARE PHISHING APPLICABILITY (CAPA) FRAMEWORK

5.1. Applicability as a constrained decision

Let T denote a candidate technique and C a deployment context consisting of an input modality, threat tolerance, hardware budget, decision deadline, privacy requirement, and response capability. The proposed CAPA model first rejects techniques that violate non-negotiable conditions. Examples include dependence on webpage source code when a browser extension must decide before navigation; reliance on an external reputation service when the endpoint must work offline; or p95 latency above the application's permitted delay. The admissible set is $F(C) = \{T : g_k(T, C) \leq 0 \text{ for every mandatory constraint } k\}$. This feasibility gate prevents a high benchmark score from compensating for an impossible deployment requirement.

For feasible techniques, construct a normalized criterion vector containing phishing recall, precision or false-alert burden at target prevalence, cross-source performance, temporal stability, calibrated risk quality, end-to-end latency, resource consumption, explanation fidelity, dependency resilience, update burden, and intervention effectiveness. A transparent context-dependent utility may be defined as $U(T|C) = \sum w_i(C) s_i(T, C)$, where $w_i \geq 0$ and $\sum w_i = 1$, and where larger s_i means better utility. Each normalization rule, threshold, and stakeholder weight must be stated in advance. Because weights reflect operational priorities, U is a decision aid—not an objective universal ranking. Pareto-front inspection and sensitivity analysis should accompany the weighted score.

5.2. Metrics and deployment trade-offs

For positive class phishing, TP, TN, FP, and FN retain their conventional meanings. Recall = $TP/(TP+FN)$; FPR = $FP/(FP+TN)$; FNR = $FN/(TP+FN)$; precision = $TP/(TP+FP)$; and $F1 = 2 \times \text{precision} \times \text{recall} / (\text{precision} + \text{recall})$. ROC-AUC summarizes discrimination over false-positive rates, whereas PR-AUC is often more informative when phishing is relatively rare. Matthew's correlation coefficient (MCC) and calibration diagnostics supplement accuracy under imbalance. All zero-denominator cases must be reported explicitly rather than silently assigned perfect values.

Base-rate sensitivity is pivotal. If attack prevalence is π , phishing recall is r , and legitimate-message false-positive rate is f , then positive predictive value is $PPV = (r\pi)/(r\pi + f(1-\pi))$. This identity implies that even a detector with high recall may generate many unjustified alerts when phishing forms a small fraction of daily traffic. Reporting precision from an artificially balanced benchmark without specifying the intended operational prevalence can therefore mislead deployment decisions. Decision thresholds should be chosen from validation data using explicit costs for missed phishing and wrongful blocking, and then assessed on an untouched test set.

Efficiency is measured at the complete system boundary: data ingestion, normalization, feature extraction, third-party lookup, model inference, explanation generation, and issuance of an actionable warning. Per-sample p50/p95/p99 latency, peak memory, model artifact size, CPU/GPU utilization, and service failure behavior should be evaluated separately for cold and warm starts. These measures must be recorded on named hardware, with and without connectivity where relevant. Analyst-facing explanations may be cached or displayed only for uncertain decisions rather than imposed on every browser navigation.

5.3. Risk-sensitive multi-criteria reporting

Rather than reducing all measurements to a single opaque index, CAPA proposes an applicability profile: (i) security discrimination; (ii) robustness and drift; (iii) user-facing error impact; (iv) response time; (v) resource use; (vi) interpretability; (vii) integration and third-party dependence; (viii) model maintenance; and (ix) prevention coverage. Each dimension retains its measured unit alongside any categorical interpretation. A scorecard can then identify dominated alternatives, hard-constraint violations, and trade-offs that require institutional judgment. Models that disagree strongly, encounter missing evidence, or exhibit low calibration confidence can abstain and route to a safer secondary control or analyst review.

6. REPRODUCIBLE EVALUATION PROTOCOL

6.1. Data acquisition and safe preparation

The experimental protocol is designed for licensed public phishing and legitimate URL and e-mail sources, including repositories derived from PhishTank or OpenPhish and appropriately permitted benign-web and mail corpora [16], [1]. Corpus documentation must state source, collection interval, label policy, licensing, observation time, and whether records represent individual URLs, registrable domains, messages, or full campaigns. Malicious URLs are handled as inert strings; any webpage artifacts must be accessed from previously captured research material or an isolated environment. Reports should display links in defanged form to limit accidental interaction.

URL preprocessing should preserve security-relevant distinctions such as percent encoding, case-sensitive path segments, punycode, full host boundaries, suspicious delimiters, and redirection indicators. A reproducible parser can derive length, digit fraction, token count, subdomain depth, domain/host indicators, entropy, and character n-grams without making a live request to the destination. E-mail preparation should separate headers, subject, plain-text body, rendered-body artifacts, and linked URLs; remove user-identifying information where necessary; and build TF-IDF or contextual text features only from training data.

6.2. Leakage-controlled evaluation design

A baseline stratified split is useful for comparability but insufficient for claims about emerging attacks. The proposed protocol additionally requires group-aware partitions that prevent a registrable domain, campaign, or near-duplicate template from appearing across train and test folds. A separate cross-source evaluation assesses portability to a repository not used in model selection. When timestamps are reliable, chronological evaluation trains on earlier material and evaluates later arrivals without using future labels or future external-intelligence records. All feature scaling, vocabulary estimation, resampling, selection, and threshold fitting must occur within the training or validation portion only.

A feasible comparison set consists of transparent rules, logistic regression, linear SVM, random forest, one compact URL sequence architecture (1D-CNN or BiLSTM), one e-mail contextual transformer (for example DistilBERT), and a compact ensemble of complementary baselines. URL and text models should first be compared within their own modality; multimodal fusion may then be examined only on records containing aligned evidence. Hyperparameter optimization uses the same evaluation budget where possible, with multiple random seeds for stochastic models. Quantized or compressed models should be identified as separate deployment variants, not casually equated to full-size counterparts.

Table 3. Minimum reporting requirements for a fair comparison

Evaluation component	Reason
Source and labels	Trace provenance and temporal applicability
Partitioning	Expose contamination and distribution shift
Security metrics	Account for asymmetric classification errors
Prevalence / calibration	Estimate likely operational alert burden
Compute and timing	Test practical decision feasibility
Robustness	Identify localized failure modes
Explainability	Support analyst review without overclaiming causality
Reproducibility	Allow independent verification

6.3. Statistical and operational validation

Report 95% confidence intervals for primary class-sensitive outcomes, using suitable binomial or bootstrap procedures, with the resampling unit matched to campaign or domain dependence. For paired comparison on shared held-out samples, a test such as McNemar's test may assess a difference in binary errors; uncertainty in PR-AUC or cost metrics may be estimated using grouped bootstrap replicates. Multiple-comparison control is required when many candidates are evaluated. Any claimed improvement must be distinguished from a small, unstable difference that lacks practical meaning at the target error cost.

A shadow-mode deployment may subsequently collect time-stamped scores, warning rates, user overrides, and analyst-confirmed outcomes before a blocking threshold is enabled. Online monitoring should track changes in input distributions, label lag, calibration, escalation volume, and model rollback conditions. Confirmed false positives and new phishing campaigns should feed a controlled retraining queue, with a retained historical validation set to detect forgetting [6]. The planned lightweight Flask, Streamlit, or browser-facing interface can expose a classification, concise reason, confidence qualifier, and report option, but an interface demonstration cannot substitute for live efficacy testing.

7. CONTEXT-SPECIFIC APPLICABILITY ANALYSIS

Applying CAPA qualitatively to the technical characteristics described in the literature yields the deployment hypotheses in Table 4. These are reasoned mappings, not measured outcomes from a common experiment. Their purpose is to determine which families should enter the shortlist for subsequent institution-specific evaluation, rather than to assert one globally superior architecture.

Table 4. Conditional technique recommendations by deployment scenario

Operational context	Non-negotiable evaluation focus	Escalation / prevention
Browser pre-navigation	Decision before credential entry; low p95 latency; offline fallback	Interstitial/block warning; report and recheck
Mobile endpoint	CPU, memory, energy, network use; accessibility of warning	Minimal warning; safe exit and reporting
Enterprise e-mail gateway	Throughput, multilingual FPR, privacy,	Quarantine, inspection, user warning

	attachments and links	
SOC/analyst triage	Attribution stability, traceability, case throughput	Case enrichment, human review, IOC propagation
Emerging-campaign watch	Temporal recall, poisoning, rollback	Shadow evaluation, supervised update, alert escalation

A browser-side detector should prioritize features available before potentially dangerous navigation and should not silently fetch a suspicious webpage simply to obtain features. A mobile model faces stricter compute and battery constraints. The mobile ensemble literature is promising [2], but pre-deployment profiling is necessary to determine whether a simpler local model is preferable. At an e-mail gateway, authenticated-domain evidence, linked destinations, and contextual text can be combined because server-side resources and quarantine controls are available. However, e-mail authentication primarily addresses domain misuse, not every case of account compromise, lookalike domains, or malicious content originating from an authenticated account [11].

Security-operations analysts benefit from model explanations and richer external intelligence, but those benefits are conditional on auditability and manageable false-alert volume. An explanation that states only “high risk” provides little operational value; a better analyst view records a suspicious host/domain mismatch, risky lexical features, inconsistent sender context, and which evidence is missing. For emerging attacks, the literature supports continual or few-shot adaptation [6], [14], while recognizing the need for controlled update cadence, poisoning protection, human adjudication, and evaluation against retained earlier periods.

8. INTEGRATION WITH PHISHING PREVENTION CONTROLS

Prevention requires at least three linked capabilities: identify potentially harmful content, interrupt or contain risky interaction, and limit damage if detection fails. The architecture proposed in Fig. 1 connects URL and text evidence to a staged decision engine. A low-cost first layer checks sender authentication, known indicators, and transparent heuristics. Uncertain records enter a modality-appropriate classifier. Contextual scoring then considers policy, confidence, recipient exposure, and evidence availability before choosing allow, warn, quarantine, block, or analyst review. Events and verified incidents feed governance and model maintenance.

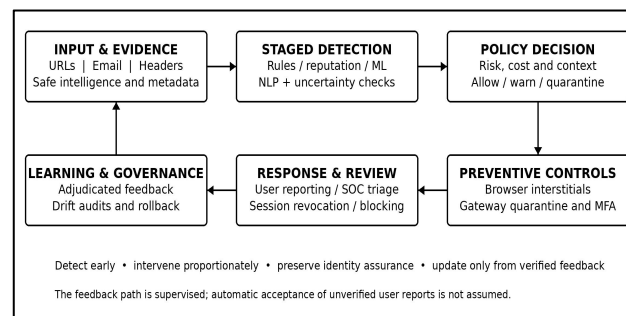


Figure 1. Conceptual layered pipeline linking detection, prevention, incident response, and validated feedback.

At the e-mail layer, SPF, DKIM, and DMARC help authenticate and align sender-domain evidence; the current DMARC specification is RFC 9989, which explicitly limits the mechanism’s direct protection to particular forms of domain spoofing [11]. At the identity layer, verifier-bound cryptographic authentication provides phishing resistance that passwords and manually entered one-time codes do not supply [9]. These measures reduce attack success even when an adversary’s wording evades a content detector. Browser warnings and message quarantine limit immediate exposure; concise reporting paths and staff training help escalate suspicious cases; and incident response must support credential reset, session termination, domain blocking, evidence preservation, and downstream notification as appropriate [12].

Prevention must be evaluated separately from classification. Relevant operational endpoints include the proportion of malicious interactions stopped before credential submission, wrongful-block incidents, warning override behavior, quarantine release time, time to credential/session revocation, and recurrence of already-known indicators. A detector might improve recall but perform poorly if its warning is delayed or routinely ignored. Conversely, phishing-resistant authentication can prevent some credential-replay attacks without changing a detector’s apparent benchmark performance.

9. DISCUSSION

The central result of this critical synthesis is a decision principle: applicable phishing protection is a system-level property, not a classifier label. The most advanced architecture is not inherently the best choice. When a local device lacks network access and must decide quickly, inexpensive features and calibrated warnings can have greater preventive value than a stronger but unavailable remote model. When a gateway has rich textual evidence and sufficient resources, a transformer may offer a justified second-stage benefit. When an analyst must explain a false positive to a business owner, traceability and stable feature attribution may carry more value than a small increase in headline accuracy.

Several design tensions remain unresolved. Contextual models may improve linguistic detection while introducing privacy and computation costs. Continual learning can reduce drift but increases vulnerability to mislabeled feedback and forgetting. High-capacity fusion may improve discriminatory evidence while creating brittle dependencies on third-party lookups. Explainability can help analysts diagnose error patterns but does not remove adversarial adaptation or guarantee faithful causal interpretations. A strong next-stage experimental study should therefore measure changes in total preventive loss under realistic workflows rather than compare model accuracy in isolation.

10. CONCLUSION

Phishing analysis and prevention require a common framework connecting predictive performance to security operations. The reviewed work establishes valuable capabilities in URL lexical modeling, sequence learning, contextual language analysis, ensembles, continual adaptation, and explainability, while also demonstrating the hazards of temporal drift, heterogeneous data, cross-language errors, and unmeasured operational costs. The CAPA framework responds by combining hard deployment constraints with transparent, context-specific multi-criteria evaluation and a staged prevention architecture. Its principal recommendation is to adopt a layered and measurable defense: use fast interpretable checks where early decisions are essential, add context-rich analysis when evidence and resources permit, and retain phishing-resistant authentication and incident response as safeguards against detection failure. The next research step is to validate these conditional recommendations with leakage-controlled, time-aware, resource-measured experiments and real prevention outcomes rather than relying on cross-paper accuracy rankings.

REFERENCES

- [1] H. Ghalechyan, E. Israyelyan, A. Arakelyan, G. Hovhannisyanyan, and A. Davtyan, "Phishing URL detection with neural networks: an empirical study," *Scientific Reports*, vol. 14, art. no. 25134, 2024. doi: 10.1038/s41598-024-74725-6.
- [2] R. S. Rao, C. Kondaiah, A. R. Pais, and B. Lee, "A hybrid super learner ensemble for phishing detection on mobile devices," *Scientific Reports*, vol. 15, art. no. 16839, 2025. doi: 10.1038/s41598-025-02009-8.
- [3] S. Abiramasundari and V. Ramaswamy, "Cacography based Ransomware Email Phishing Attack Prevention using Language Pack Tuned Transformer Language Model," *Scientific Reports*, vol. 15, art. no. 21526, 2025. doi: 10.1038/s41598-025-06530-8.
- [4] J. Zhou, K. Zhang, A. Bilal, et al., "An integrated CSPPC and BiLSTM framework for malicious URL detection," *Scientific Reports*, vol. 15, art. no. 6659, 2025. doi: 10.1038/s41598-025-91148-z.
- [5] M. Hosseinzadeh, U. Ali, S. Ali, et al., "Improving phishing email detection performance through deep learning with adaptive optimization," *Scientific Reports*, vol. 15, art. no. 36724, 2025. doi: 10.1038/s41598-025-20668-5.
- [6] A. Ejaz, A. N. Mian, and S. Manzoor, "Life-long phishing attack detection using continual learning," *Scientific Reports*, vol. 13, art. no. 11488, 2023. doi: 10.1038/s41598-023-37552-9.
- [7] D. Komosny, "Phishing detection on webpages in European non-English languages based on machine learning," *Scientific Reports*, vol. 15, art. no. 37472, 2025. doi: 10.1038/s41598-025-21384-w.
- [8] M. Dandotiya, N. K. Goyal, A. Khunteta, and B. Tiwari, "Real time identification of phishing attacks through machine learning enhanced browser extensions," *Scientific Reports*, vol. 16, art. no. 6612, 2026. doi: 10.1038/s41598-026-35655-7.
- [9] National Institute of Standards and Technology, "Digital Identity Guidelines: Authentication and Authenticator Management," NIST SP 800-63B-4, July 2025. doi: 10.6028/NIST.SP.800-63b-4.
- [10] E. Tabassi, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," NIST AI 100-1, Jan. 2023. doi: 10.6028/NIST.AI.100-1.
- [11] T. Herr and J. Levine, Eds., "Domain-Based Message Authentication, Reporting, and Conformance (DMARC)," RFC 9989, May 2026. doi: 10.17487/RFC9989.
- [12] Cybersecurity and Infrastructure Security Agency, "Implementing Phishing-Resistant MFA," Fact Sheet, 2022. Available: <https://www.cisa.gov/sites/default/files/2023-01/fact-sheet-implementing-phishing-resistant-mfa-508c.pdf>.
- [13] E. Gandotra, D. Gupta, M. Dhalaria, N. Gupta, and S. Redkar, "PhishDetect: A ranking-based classifier integration approach for improving phishing website detection," *Scientific Reports*, vol. 16, art. no. 29386, 2026. doi: 10.1038/s41598-026-57807-5.
- [14] M. Tawfik, A. A. Abu-Ein, A. H. Abdelhaliem, Y. M. Al-Sharo, and I. S. Fathi, "Explainable few-shot learning with modern BERT for detecting emerging phishing attacks using XF PhishBERT," *Scientific Reports*, vol. 15, art. no. 42821, 2025. doi: 10.1038/s41598-025-27500-0.
- [15] T. Kehkashan, M. Abdelhaq, A. S. Al-Shamayleh, et al., "Explainable phishing website detection for secure and sustainable cyber infrastructure," *Scientific Reports*, vol. 15, art. no. 41751, 2025. doi: 10.1038/s41598-025-27984-w.
- [16] J. C. Gomes de Barros et al., "Piracema: a Phishing snapshot database for building dataset features," *Scientific Reports*, vol. 12, art. no. 15149, 2022. doi: 10.1038/s41598-022-19442-8.

-
- [17] R. Rawat, H. Rawat, A. Rawat, and A. Rajavat, "An entropy-guided hybrid framework for real-time phishing detection in digital communication systems," *Scientific Reports*, vol. 16, art. no. 17222, 2026. doi: 10.1038/s41598-026-46430-z.
- [18] R. Gobinath and S. Manikandan, "Deep learning-based phishing classification framework for accurate detection using optimized URL intelligence," *Scientific Reports*, vol. 16, art. no. 15794, 2026. doi: 10.1038/s41598-026-46481-2.
- [19] B. B. Gupta, S.-H. Pan, A. Gaurav, et al., "Phishing threat mitigation in E-commerce using a quantum-enhanced hybrid AI framework," *Scientific Reports*, vol. 15, art. no. 42282, 2025. doi: 10.1038/s41598-025-27515-7.