



International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

The Rise of Sophisticated Phishing Attacks: Detection, Prevention, and Future Directions

Amit Sharma¹ and Khushboo Bansal²

¹Research Scholar, Desh Bhagat University, Mandi Gobindgarh, Punjab, India, Email: as901842@gmail.com, ORCID: 0009-0008-0672-4559

²Associate Professor, Desh Bhagat University, Mandi Gobindgarh, Punjab, India, Email: kbansal@deshbhagatuniversity.in, ORCID: 0000-0001-7492-4626

Abstract

Phishing has evolved from crude, error-prone mass email campaigns into a highly sophisticated class of multi-channel social-engineering attacks that exploit artificial intelligence, real-time session hijacking, and highly convincing brand impersonation. Traditional defence mechanisms, including blacklists, simple URL heuristics, and signature-based filters, are no longer sufficient to counter this rapidly expanding threat landscape. This review synthesizes recent literature on sophisticated phishing attacks and makes four primary contributions. First, it presents a systematized taxonomy of modern phishing attack vectors, encompassing email, SMS, voice, QR code, business email compromise (BEC), and adversary-in-the-middle (AiTM) techniques. Second, it provides a critical and comparative review of detection approaches spanning rule-based methods, classical machine-learning algorithms, deep-learning models, transformer-based architectures, visual analysis techniques, and behavioural detection methods, highlighting their reported strengths, limitations, and levels of operational maturity. Third, it proposes a conceptual four-layer hybrid detection framework that integrates lexical URL analysis, HTML and visual content inspection, transformer-based semantic understanding, and behavioural session analytics through a stacked meta-learner to improve detection accuracy and robustness. Finally, the review discusses current prevention strategies, including FIDO2 phishing-resistant authentication, DMARC enforcement, and human-centric security awareness training, while identifying key research directions such as LLM-driven adversarial phishing, deepfake voice attacks, privacy-preserving detection, explainable artificial intelligence, and resilient security workflows for next-generation phishing defence.

Keywords: Phishing detection, social engineering, cybersecurity, machine learning, deep learning, transformer models, adversary-in-the-middle, FIDO2, DMARC.

1. INTRODUCTION

Phishing has shifted, over the past five years, from a high-volume but low-skill nuisance into one of the principal initial-access vectors for serious cybercrime. The 2023 Verizon Data Breach Investigations Report attributes a substantial fraction of confirmed breaches to social-engineering precursors, with phishing remaining the dominant technique [2]. The Anti-Phishing Working Group recorded record-high phishing volumes in 2023, with an order-of-magnitude increase observed over the five-year period beginning in 2018 [1]. What has changed is not merely the scale but the sophistication: attackers now routinely deploy large language models (LLMs) to craft grammatically perfect, contextually relevant lures, run adversary-in-the-middle (AiTM) reverse proxies that defeat traditional multi-factor authentication, and embed malicious QR codes in printed documents to evade URL-based filters entirely [15], [16].

Classical defences were not designed for this threat surface. Blacklist-based filters, which served well when phishing infrastructure persisted for days or weeks, are largely circumvented by short-lived domains that live only hours [22]. Heuristic URL rules, designed by hand and effective against grammatically poor or hastily registered domains, struggle with attacker-controlled assets that look identical to legitimate ones [21]. Even mature machine-

learning classifiers built on engineered URL and HTML features begin to show degradation as adversaries gain access to generative tools capable of producing arbitrary, semantically rich content at low cost [15]. The result is a widening gap between the sophistication of offensive tooling and the assumptions embedded in widely deployed defensive technology.

The consequences of this gap are not abstract. Phishing is repeatedly identified as the most common entry point for financially motivated intrusions, ransomware deployment, and large-scale data breaches, and the losses attributed to BEC alone run to billions of dollars annually across reporting jurisdictions [2], [17]. Beyond direct financial loss, successful phishing imposes substantial secondary costs—incident response, regulatory reporting, credential resets, and erosion of customer trust—that fall disproportionately on organisations without mature detection and containment capabilities. The economic asymmetry is therefore stark: an attacker can launch a plausible, AI-assisted campaign for a negligible marginal cost, while the defender must sustain a layered, continuously updated programme to contain even a modest fraction of attempts. This asymmetry, rather than any single technical deficiency, is the central strategic challenge the remainder of this paper addresses.

This paper presents a structured review of the current state of sophisticated phishing and its detection, with four goals. First, we systematize the modern phishing landscape, mapping out attack vectors that have emerged or matured since 2020 and identifying the underlying capability shifts (AI generation, AiTM, multi-channel pivoting) that distinguish them from earlier generations. Second, we critically review detection research across rule-based, classical-ML, deep-learning, transformer-based, visual, and behavioural families, surfacing both their reported strengths and the operational limitations that often go unreported. Third, we synthesize these findings into a proposed conceptual four-layer hybrid detection framework that combines lexical, content-based, semantic, and behavioural signals in a stacked-ensemble architecture. We position the framework as a design template that the community can instantiate, evaluate, and extend; empirical validation is identified as the principal next step. Fourth, we discuss prevention strategies and outline open research directions.

Scope and method. This review is narrative and synthetic rather than an exhaustive systematic review. We surveyed peer-reviewed literature, standards documents, and industry threat-intelligence reports indexed up to early 2024, prioritizing works that (i) introduced a widely cited detection technique or dataset, (ii) reported benchmarking or survey evidence across multiple techniques, or (iii) documented a capability shift in attacker tradecraft. Sources span venues in computer security, machine learning, and network management, together with normative references from NIST, the IETF, the FIDO Alliance, and ENISA. Where the literature reports quantitative accuracy figures, we treat them cautiously because datasets and evaluation protocols differ widely across studies; consequently we characterize the maturity of each technique family qualitatively rather than ranking families by headline accuracy.

The remainder of the paper is organized as follows. Section 2 surveys the taxonomy of sophisticated phishing attacks and their lifecycle. Section 3 reviews detection techniques and the methodological practices used to evaluate them. Section 4 introduces the proposed conceptual framework and maps it to the threat model. Section 5 discusses prevention strategies and human factors. Section 6 outlines open research directions and the limitations of this review. Section 7 concludes. Two appendices provide a glossary of key terms and a summary of representative public datasets.

2. TAXONOMY OF SOPHISTICATED PHISHING ATTACKS

A. Vector-based classification

Modern phishing campaigns are no longer characterized solely by the delivery channel; they are increasingly hybrid, beginning in one channel and pivoting to another. Table I summarizes the principal vectors observed across recent threat reports [1], [17], [30]. Email remains the dominant initial channel, but SMS-based phishing (smishing) and voice phishing (vishing) have grown rapidly in financial-services targeting [17]. QR-code phishing—often called

—quishing—circumvents URL-scanning gateways because the malicious URL is encoded inside an image and is only resolved after the user scans it on a personal device that is typically outside the corporate security perimeter [17]. Business Email Compromise (BEC) attacks, in which an attacker impersonates a senior executive or trusted vendor to authorize a fraudulent wire transfer, are responsible for a disproportionate share of financial losses despite representing a small fraction of total phishing volume [2].

Alongside these headline vectors, several older techniques persist in modernized form. Clone phishing replaces a legitimate attachment or link in a previously delivered message with a malicious equivalent, exploiting the recipient's prior trust in the thread. Pharming redirects victims from correctly typed domains to attacker-controlled servers through DNS or host-file poisoning, bypassing the user's scrutiny of the address bar entirely. Pretexting establishes a fabricated but plausible narrative—an overdue invoice, a security alert, a payroll

change—that lowers the victim’s guard before any malicious payload is presented [3], [4]. These techniques are relevant precisely because they are frequently combined with the newer capabilities described below.

B. Capability-based shifts

Three capability shifts mark the current generation of phishing as qualitatively distinct from its predecessors. First, the operational availability of large language models has eliminated grammatical and stylistic giveaways that previously served as reliable detection signals; AI-generated lures are linguistically indistinguishable from legitimate corporate correspondence [15], [25]. Second, AiTM frameworks based on reverse-proxy phishing kits intercept user credentials and session cookies in real time, defeating SMS-based and push-notification MFA by relaying the authentication transaction to the legitimate service while harvesting the resulting session token [16]. Third, multi-channel pivoting—where an attacker initiates contact via email but moves the victim to SMS, instant messaging, or a phone call to complete the deception—breaks the assumption that any single channel can be defended in isolation [17].

C. Economic and ecosystem drivers

The sophistication shift is supported by an underground ecosystem in which phishing-as-a-service kits, AI prompt libraries, and stolen-credential marketplaces lower the technical barrier for new entrants [30]. Phishing kits with built-in AiTM functionality and template galleries for major brands are now offered on a subscription basis, allowing operators with limited technical skill to deploy campaigns that would previously have required custom development. This commoditization parallels the broader trend toward as-a-service offerings in cybercrime and has direct implications for defenders: the volume of low-cost, plausible phishing campaigns is increasing faster than manual analyst capacity can scale [4], [17].

Targeting is also increasingly selective. Rather than distributing a single template indiscriminately, sophisticated operators tailor campaigns to sectors and roles—finance and payroll staff for BEC and wire fraud, IT help-desk personnel for MFA-fatigue and reset abuse, and executives for high-value authorisation fraud. Seasonal and event-driven lures (tax deadlines, benefits-enrolment windows, high-profile outages, and major software releases) exploit predictable moments of elevated user expectation, when an unsolicited message is more likely to be accepted at face value. This selectivity compounds the detection problem, because narrowly targeted, context-aware messages generate fewer samples for defenders to learn from and are less likely to match signatures derived from mass campaigns [4], [17], [30].

D. The phishing lifecycle and kill chain

It is useful to view sophisticated phishing not as a single event but as a lifecycle that mirrors the intrusion kill chain [4], [21]. In the reconnaissance stage, attackers harvest names, roles, reporting lines, and ongoing projects from social media, corporate websites, and breached data to increase contextual relevance. In the weaponization stage, they register look-alike domains, provision TLS certificates, and stand up phishing kits—now often with AiTM proxying and brand-accurate templates. Delivery exploits the vectors of Table I, frequently pivoting across channels. Exploitation occurs when the victim interacts—entering credentials, approving an MFA prompt, or scanning a QR code. Credential and session capture follows, with AiTM kits stealing live session cookies rather than static passwords. Finally, monetization ranges from immediate wire fraud in BEC cases to lateral movement, data exfiltration, and ransomware staging. Framing detection and prevention against this lifecycle clarifies why single-point defences are insufficient: each stage presents distinct, and largely non-overlapping, opportunities for interruption, and a control that is strong at one stage may be irrelevant at another.

Table I. Taxonomy of sophisticated phishing attack vectors and representative threat indicators.

Attack Vector	Key Characteristics	Representative Threat Indicators
Email phishing	Mass-distribution emails impersonating trusted brands; rely on display-name spoofing and look-alike domains.	DMARC failures, SPF/DKIM mismatch, homoglyph domains, urgent calls-to-action.
Spear phishing	Targeted, OSINT-driven messages aimed at specific individuals; high contextual relevance.	Personalized salutation, reference to internal projects, attachment-based payloads.
Smishing / Vishing	SMS- or voice-based social engineering; exploits trust in telephony channels.	Short URLs, caller-ID spoofing, AI-generated voice clones.
Business Email Compromise (BEC)	Impersonation of executives or vendors to authorize fraudulent transactions.	Look-alike sender domains, wire-transfer urgency, change of payment details.
AI-generated / LLM	Use of large language models to produce	Absence of grammatical

phishing	grammatically correct, context-aware lures at scale.	anomalies, semantic coherence, polymorphic content.
Adversary-in-the-Middle (AiTM)	Reverse proxy intercepts authentication, harvesting session cookies and bypassing MFA.	Unrecognized session tokens, anomalous IP/device fingerprints, replayed cookies.
QR-code phishing (Quishing)	Malicious QR codes embedded in emails or printed media to bypass URL-based filters.	QR images in unsolicited mail, redirect to credential-harvesting pages.

3. DETECTION TECHNIQUES: A CRITICAL REVIEW

A. Rule-based and blacklist approaches

Early phishing detection systems relied on cooperatively maintained blacklists (e.g., Google Safe Browsing, PhishTank) and on hand-crafted heuristic rules over URL structure [22], [29]. PhishNet [22] extended pure blacklisting with five heuristics that generated variants of blacklisted URLs to anticipate near-duplicates. While computationally cheap and interpretable, these methods are now largely ineffective against zero-hour campaigns whose infrastructure has a median lifetime of hours rather than days [1], [22]. The principal value of blacklists in modern systems is as a fast, deterministic exclusion layer that can be combined with more sophisticated techniques rather than as a primary defence.

B. Classical machine learning

Subsequent work moved to supervised classification over engineered features. Sahingoz et al. [9] benchmarked Random Forest, k-NN, AdaBoost, and SVM over lexical URL features and reported strong accuracy on PhishTank-derived datasets. PhishStorm [7] used streaming analytics to flag phishing URLs in real time. Jain and Gupta [8], [23] proposed hyperlink-based feature engineering, exploiting the structural anomalies of phishing pages that often reference resources hosted on unrelated legitimate domains. Abu-Nimeh et al. [27] and Fette et al. [28] established many of the foundational features and benchmarking conventions still used today. More recently, there has been a growing interest in cybersecurity research for multi-layer machine-learning and deep-learning intrusion detection, which has been extended to other cyber threats and has been shown to be applicable to a wider range of cyber threats [32-34].

The principal weakness of this family is feature drift: adversaries who become aware of the heuristics in use can trivially adjust URL length, subdomain count, and certificate properties to avoid them. Aassal et al. [10] provided an in-depth benchmarking exercise that found substantial accuracy degradation when classical detectors trained on one dataset were applied to temporally disjoint test sets, underscoring the importance of temporal validation in phishing-detection research. Feature-engineered models nonetheless remain attractive in resource-constrained settings because of their low inference cost, interpretability, and modest data requirements, and they continue to serve effectively as a first-pass triage stage.

C. Deep learning and representation learning

URLNet [11] pioneered the application of character-level CNNs to URLs, removing the need for hand-crafted features and learning obfuscation-robust representations directly from raw strings. Huang et al. [12] combined CNN feature extraction with an attention-based hierarchical RNN, allowing the model to weight the most discriminative segments of a URL. These models are robust to obfuscation strategies (URL shortening, character substitution, homoglyph insertion) that defeat rule-based systems, but they are data-hungry and offer limited interpretability. The trade-off between feature engineering and deep representation learning remains active: deep models generalize better within a domain but require labelled corpora of a size that not all defenders have access to, and they can degrade sharply when the deployment distribution diverges from the training distribution.

D. Transformer and LLM-based detection

With the maturation of transformer architectures [26], detection has expanded from URL strings to the semantic content of emails and pages. Tang and Mabry [15] fine-tuned BERT and RoBERTa on phishing email corpora and reported strong performance against AI-generated lures that defeated classical filters. Verma et al. [5] showed early evidence that NLP-based intent classification—labels such as credential request, urgency, and financial action—is informative independent of URL features. Brown et al. [25] established the general-purpose, few-shot capabilities of large language models, which are exploited by attackers and defenders alike: the same underlying technology that produces sophisticated phishing also enables the most capable semantic detectors.

Transformer models, however, demand substantial compute, are themselves susceptible to adversarial perturbations of the input text, and require curated labelled data that remain expensive to obtain. There is also a more fundamental tension: as both sides adopt transformer-based tools, the asymmetry shifts toward whichever

party has access to more training data and inference compute, which may not always favour the defender [10], [15]. Because semantic detectors operate on message content rather than on delivery infrastructure, they are also the family most directly exposed to privacy and data-governance constraints, a point revisited in Section 6.

E. Visual and behavioural detection

Visual-similarity approaches compare a screenshot of a candidate page to a gallery of legitimate brand pages using image embeddings; they are effective against credential-cloning sites that copy a known visual layout [6]. Behavioural detection complements visual and content-based methods by analysing post-click signals: anomalous mouse and keyboard dynamics, impossible-travel logins, and replayed session cookies are signatures of AiTM compromise even when the underlying phishing page evades static detection [16]. The behavioural family is particularly important because static defences alone cannot, in principle, detect AiTM attacks in which the phishing page is an exact, transparent proxy of a legitimate one; the anomaly manifests only in the resulting session, not in the page itself.

F. Synthesis and gap analysis

Table II summarizes the comparative strengths and limitations of these technique families. Two persistent gaps emerge from the literature. First, no single technique dominates across all operational metrics; URL-only detectors cannot see the AI-generated body of a BEC email, content-only detectors cannot detect AiTM session hijacks, and behavioural detectors cannot prevent the initial credential harvest. Second, real-time deployment requires careful orchestration: fast, cheap signals must be evaluated first so that expensive deep-learning inference is reserved for ambiguous cases, but the literature largely reports results in isolation without considering this orchestration layer [10], [21]. More recent AI-driven security research also focuses on adaptive defence architectures, which integrate machine-learning based threat detection with automated response capabilities, underscoring the need for continuous adaptation of security systems, rather than solely on static detection rules [31].

G. Methodological considerations across the literature

A separate, methodological gap deserves explicit attention. Many published phishing-detection results are reported on random train-test splits of static, single-source corpora, which tend to overstate operational performance because adjacent phishing URLs in such corpora often share infrastructure, brand templates, and even attacker tooling. Aassal et al. [10] demonstrated that temporally disjoint evaluation—training on older data and testing on strictly later data—can reduce reported accuracy by several percentage points, particularly for feature-engineered classical models. A second methodological concern is class imbalance handling: real-world phishing-to-benign ratios are highly skewed, and detectors evaluated on artificially balanced datasets may not generalize. Reviewers and future authors should therefore look for, and report, three methodological features: temporal split, realistic class balance, and evaluation on samples that postdate any external blacklist incorporated as a feature [10], [13], [21].

Datasets, benchmarks, and evaluation metrics

Progress in the field is tightly coupled to the datasets on which methods are trained and compared. Commonly used resources include PhishTank and OpenPhish for live phishing URLs, the UCI phishing-websites feature set [14], and curated collections such as those released by Vrbančič et al. [13], while benign samples are typically drawn from ranked domain lists. Each carries limitations: feature-only datasets fix a particular feature engineering and cannot be re-derived for new representations; URL-only datasets omit page content and session behaviour; and label provenance is often tied to blacklists, which biases evaluation if the same blacklist is later used as a feature. Reported metrics also vary—accuracy, precision, recall, F1, and area under the ROC curve are all common—yet accuracy alone is misleading under the severe class imbalance of real traffic, where a trivial always-benign classifier can appear highly accurate. For operational relevance, the false-positive rate at a fixed, low value is frequently more informative than headline accuracy, because analyst alert fatigue and user friction are driven by false positives. The absence of a shared, temporally versioned, multi-modal benchmark that includes AI-generated and AiTM samples is, in our assessment, one of the most consequential gaps in the field, and it directly motivates the evaluation agenda in Section 6.

Table II. Comparative analysis of phishing-detection techniques.

Technique	Detection Basis	Strengths	Limitations	Maturity
Blacklist / URL reputation	Static lists of known-malicious URLs.	Low overhead, deterministic.	Useless against zero-hour attacks; high maintenance.	Established
Heuristic rules	Hand-crafted URL features (length, IP-in-URL, age).	Interpretable; deployable on endpoints.	Brittle; high false positives.	Established

Classical ML (RF, SVM, XGBoost)	Engineered URL, content and host features.	Good accuracy on benchmarks; modest data needs.	Feature drift; limited semantics.	Mature
Deep learning (CNN, LSTM)	Learned representations of URLs and HTML.	Robust to obfuscation; less manual feature work.	Data-hungry; opaque.	Mature
Transformer / LLM-based	Contextual semantics of email and page content.	Effective against AI-generated lures; multilingual.	High compute; adversarial perturbation risk.	Emerging
Visual similarity	Screenshot embedding vs legitimate brand pages.	Effective against credential-cloning sites.	Slow; fails on novel layouts.	Mature
Behaviour / network anomaly	User and traffic behaviour analytics.	Catches post-click compromise (AiTM).	Requires baselines; alert fatigue.	Emerging

4. PROPOSED CONCEPTUAL MULTI-LAYER FRAMEWORK

Drawing on the gap analysis of Section 3, we propose a conceptual four-layer hybrid framework for phishing detection. The framework is presented as a design template intended to guide future implementation and evaluation rather than as a system that has been independently validated in this paper; empirical evaluation is identified explicitly as a follow-up research item (see Section 6).

A. Design rationale

The review in Section 3 reveals that single-modality detectors each fail against at least one class of sophisticated attack. The proposed framework therefore arranges signals in four cooperating layers, with cheap signals evaluated first and more expensive signals invoked conditionally on the ambiguity of upstream layers. This cascaded organisation is not merely an accuracy choice but an economic and latency choice: sub-millisecond lexical checks can clear the large majority of obviously benign traffic, reserving costly page rendering, transformer inference, and behavioural correlation for the minority of ambiguous candidates. Per-layer outputs are combined by a stacked meta-learner so that no single layer's failure modes dominate the final decision. Figure 1 illustrates the architecture.

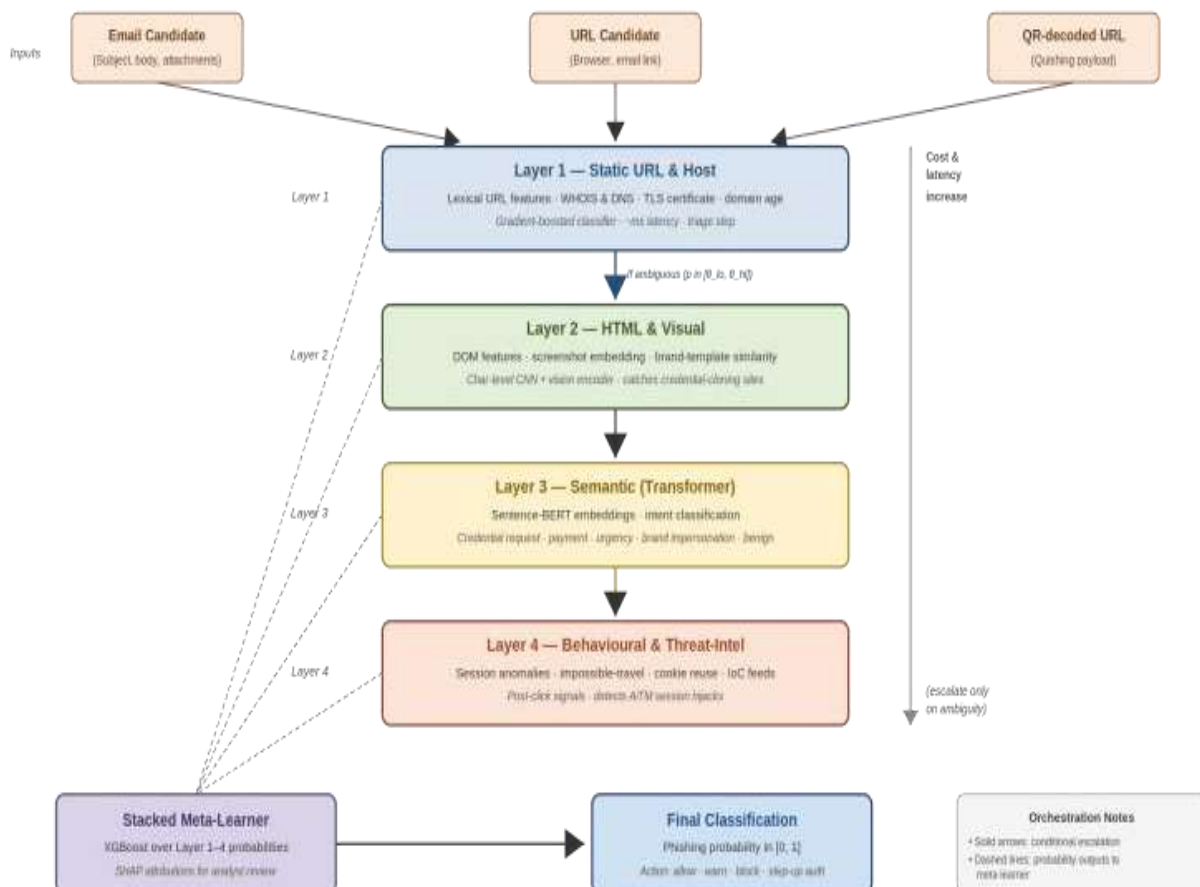


Figure 1. Architecture of the proposed four-layer conceptual phishing detection framework. Incoming candidates (URL, email, or QR-decoded payload) are first passed through Layer 1 (lexical and host features) for a fast triage. Ambiguous cases are escalated to Layer 2 (HTML and visual analysis) and Layer 3 (transformer-based semantic understanding). Layer 4 (behavioural and threat-intel) provides post-click signals. A stacked meta-learner combines layer probabilities and produces the final classification.

B. Layer 1 — Static URL and host

Layer 1 extracts lexical and WHOIS/DNS features (Table III) at sub-millisecond latency. A gradient-boosted classifier trained on these features alone would serve as both a triage step and a fallback when downstream layers are unavailable. Confidence scores in a defined ambiguity band escalate the candidate to Layer 2, while high-confidence benign and malicious verdicts short-circuit further processing to conserve resources.

C. Layer 2 — HTML and visual

Layer 2 fetches the candidate page in a sandboxed environment and extracts DOM-level features and a screenshot. A character-level CNN over the URL is combined with a vision encoder over the screenshot; the screenshot embedding is compared to a precomputed bank of legitimate brand templates via cosine similarity. This layer is principally responsible for detecting credential-cloning sites that pass URL triage by registering on plausible-looking domains [6].

D. Layer 3 — Semantic (transformer)

Layer 3 applies a fine-tuned transformer (for example Sentence-BERT) to the body of email candidates or the visible text of web candidates. Intent labels include credential request, payment redirection, urgency, brand impersonation, and benign. This layer is the principal proposed defence against AI-generated lures whose lexical and structural URL features may appear unremarkable [15], [25]. Because it operates on message content, its deployment must be governed by the privacy considerations discussed in Section 6.

E. Layer 4 — Behavioural and threat-intel

Layer 4 ingests post-click signals from the identity provider and the network: time-to-click distribution, session-cookie reuse, impossible-travel detection, IP and ASN reputation, and matches against curated indicator-of-compromise (IoC) feeds. The output is a continuous risk score for the active session, allowing the orchestration layer to revoke tokens or enforce step-up authentication when behavioural anomaly co-occurs with a borderline static score [16]. This layer is the only one in the framework capable, in principle, of detecting AiTM compromise

after the initial page has been served.

F. Stacked meta-learner

Per-layer probabilities are fed to a meta-learner (such as XGBoost or logistic regression) trained with cross-validation. The meta-learner produces a final phishing probability together with interpretable feature attribution via SHAP, supporting analyst review. This design intentionally separates per-layer model design from the cross-layer combination policy, allowing each layer to be replaced or upgraded independently as attacker tradecraft evolves, without retraining the entire system.

G. Feature inventory

Table III. Feature groups across the four layers of the proposed conceptual framework.

Feature Group	Representative Features	Layer
Lexical URL	URL length; entropy of path; digit ratio; punycode; suspicious TLD; @-symbol; subdomain count.	Layer 1 — Static URL
Host / WHOIS	Domain age; registrar reputation; geographic mismatch; DNS TTL; TLS	Layer 1 — Static URL

Feature Group	Representative Features	Layer
	certificate issuer and validity.	
HTML / DOM	External-resource ratio; form-action mismatch; obfuscated JavaScript; iframe count; favicon mismatch.	Layer 2 — Content
Visual similarity	CNN embedding distance to top-N legitimate brand templates; logo detection.	Layer 2 — Content
NLP semantics (transformer)	Sentence-BERT embeddings of body text; intent classification (credential request, payment, urgency).	Layer 3 — Semantic
Behavioural / session	Typing cadence; mouse dynamics; IP geolocation drift; impossible-travel; cookie reuse.	Layer 4 — Behavioural
Threat-intel correlation	Matches against IoC feeds; passive DNS; certificate-transparency anomalies.	Layer 4 — Behavioural

H. Threat-model coverage

A useful test of any layered design is whether its layers jointly cover the attack classes of Section 2, and where residual exposure remains. Table IV maps the framework's layers to representative attack classes. Classical mass and spear phishing are addressed primarily at Layers 1 and 2, where infrastructure and cloned-page signals are strongest. AI-generated lures, which are engineered precisely to defeat lexical and grammatical cues, are addressed principally at Layer 3 through intent-level semantic analysis. QR-code phishing requires decoding the embedded payload before Layer 1 can act, and therefore depends on an upstream image-processing step feeding the URL pipeline. AiTM attacks are the hardest case: because the phishing page is a faithful proxy of the legitimate site, Layers 1–3 may all pass, and detection falls to Layer 4 behavioural analytics and, ultimately, to phishing-resistant authentication (Section 5). This mapping makes the framework's residual risk explicit: no arrangement of detection layers fully neutralises AiTM, which is why detection must be paired with cryptographic, origin-bound authentication rather than treated as a standalone solution.

Table IV. Illustrative mapping of framework layers to representative attack classes.

Attack class	Primary detection locus	Residual risk
Mass / spear phishing	Layer 1 (host/URL) + Layer 2 (visual clone).	Low–Medium
BEC (no link/attachment)	Layer 3 (intent semantics: payment, urgency).	Medium
AI-generated lures	Layer 3 (semantic) + Layer 2 (page content).	Medium
QR-code phishing	Image decode → Layer 1/2 on resolved URL.	Medium
AiTM session hijack	Layer 4 (behavioural, cookie reuse, impossible travel).	High (needs FIDO2)

5. PREVENTION STRATEGIES AND HUMAN FACTORS

A. Phishing-resistant authentication

Detection alone is insufficient. The single most impactful preventive measure available today is the deployment of phishing-resistant authentication based on the FIDO2 standard [20]. Unlike SMS or push-based MFA, FIDO2 binds the authentication ceremony cryptographically to the origin domain, defeating AiTM reverse-proxy attacks at the protocol level. Organizations that have completed FIDO2 rollouts have reported substantial reductions in credential-phishing-driven account takeover [16], [20]. As Table IV makes clear, FIDO2 is the control that closes the residual gap left by even a well-designed detection stack.

B. Email-authentication enforcement

Properly configured DMARC enforcement (`—p=reject`) together with SPF and DKIM prevents external attackers from spoofing the organization's own domain in email envelopes [19]. Adoption remains uneven; a substantial fraction of high-value domains still publish DMARC in monitor-only mode, providing visibility without protection [19], [30]. Industry tracking suggests that the gap between published-and-monitored and published-and-enforced is one of the most consequential, and most fixable, weaknesses of the current email-security landscape.

C. Human-centric training

Anti-phishing training programmes have been shown to reduce click rates but plateau at a non-zero residual risk; sophisticated attackers, particularly those using AI-generated content, can still defeat trained users [17]. The operational implication is that training reduces but does not eliminate the need for technical controls. Training and technical controls should be deployed together rather than as alternatives, and training should explicitly cover the newest vectors (QR phishing, AI-generated lures, voice clones) rather than rely on older indicators that no longer apply [17], [30].

D. NIST guidelines and standards alignment

The NIST SP 800-63B guidelines on digital identity [18] and the IETF DMARC specification [19] together provide a normative baseline for organizational anti-phishing posture. NIST explicitly classifies SMS-based one-time passwords as a `—restricted` authenticator, recognizing their vulnerability to SIM-swap and AiTM attacks. ENISA's Threat Landscape reports [30] similarly emphasize the importance of moving away from knowledge-based and SMS-based factors. These standards documents represent a useful, defensible reference point for organizations under audit or regulatory scrutiny.

E. Layered defence and economic trade-offs

Effective anti-phishing posture is layered. Email-authentication enforcement reduces the volume of spoofed inbound mail; client-side detectors and gateways filter what gets through; phishing-resistant authentication contains the damage when a user is nevertheless deceived; behavioural and threat-intel analytics detect post-click compromise; and trained users contribute a residual but non-trivial reduction in click rate. Each layer has marginal cost and marginal benefit, and organizations with constrained security budgets benefit from explicit prioritization. The literature consistently identifies FIDO2 deployment and DMARC enforcement as the two highest-impact, lowest-relative-cost interventions for most enterprises [18], [19], [20], whereas large-scale custom-detection R&D is typically justifiable only at organizations of substantial size or in sectors of high attacker interest [17], [30].

A common pitfall is the substitution fallacy: deploying a single advanced technical control and treating it as a replacement for layered defence. The history of phishing detection suggests that no single layer—however sophisticated—has yet survived more than a few years before attacker adaptation degraded its effectiveness. Defensive architectures that assume continued attacker adaptation, and that combine multiple weak signals in defensible ways, are more durable than those that depend on a single technique remaining effective indefinitely [4], [21].

F. Zero-trust, isolation, and incident response

Three organisational controls complement the technical measures above. First, a zero-trust access posture—continuous verification of identity, device health, and context on every request rather than one-time perimeter authentication—limits the blast radius of a stolen credential or session, since a captured token no longer grants broad, standing access. Second, remote browser isolation and sandboxed link-rewriting execute untrusted content away from the endpoint, so that even a successful click does not directly expose the user's device or credentials. Third, a mature reporting and incident-response capability shortens the window between compromise and containment: one-click user reporting, automated triage, rapid session-token revocation, and coordinated takedown of malicious infrastructure convert individual detections into organisation-wide protection. Crucially, a healthy reporting culture reframes the user from a point of failure into an active sensor; each reported message enriches threat intelligence and can be fed back into the detection layers of Section 4. Together these controls acknowledge that some fraction of sophisticated phishing will always reach users, and they focus on limiting and reversing its consequences rather than on prevention alone [4], [16], [21].

6. OPEN RESEARCH DIRECTIONS AND LIMITATIONS

A. Open research directions

Several research questions are particularly pressing. (i) LLM-driven adversarial phishing: as attackers fine-tune their own LLMs, detection systems will need to defend against prompt-engineered evasion attacks that target the transformer detectors directly, motivating research on adversarial training and robustness certification for security NLP. (ii) Deepfake voice and video phishing: vishing campaigns using AI-cloned voices of executives have moved from proof-of-concept to active exploitation; detection requires audio and video forensic capabilities

outside the scope of current text-and-URL pipelines. (iii) Privacy-preserving detection: high-quality detection often relies on inspecting email content and user behavioural signals, raising regulatory and privacy concerns; federated learning and on-device inference are promising but require further research on robustness and communication efficiency. (iv) Explainability: production deployment requires that analyst-facing alerts come with human-readable explanations, motivating work on SHAP-based, counterfactual, and attention-based explanation methods specific to security workflows.

Three further directions extend beyond the immediate detection problem. (v) Supply-chain and vendor-impersonation phishing: as direct domain spoofing is closed off by DMARC enforcement, attackers increasingly compromise or impersonate trusted third parties, so detection must incorporate cross-organisational trust and vendor-relationship signals. (vi) Standardised, temporally versioned benchmarks: the field lacks a shared, multi-modal, time-stamped dataset that includes AI-generated and AiTM samples, and building one—together with agreed reporting of false-positive rates at fixed thresholds—would materially improve the comparability of future work [10], [13]. (vii) Forward-looking authentication resilience: as authentication ecosystems evolve, sustained attention to the migration path toward phishing-resistant and, ultimately, cryptographically agile factors will be needed to ensure that today's defences remain valid against tomorrow's capabilities.

B. Limitations of this review

Three limitations of this review should be acknowledged. First, sophisticated phishing is a rapidly evolving area; the literature surveyed here reflects work indexed up to early 2024, and substantial relevant work may have appeared since. Second, indicative comparisons across detection-technique families (Table II) draw on benchmark studies that use heterogeneous datasets and evaluation protocols, and direct numerical comparison across families is therefore not reliable; we have accordingly characterized maturity qualitatively rather than reporting cross-study accuracy figures. Third, the proposed conceptual framework (Section 4) is presented as a design template; its empirical evaluation—on temporally split datasets that include AI-generated and AiTM samples—is the principal follow-up research activity that this review identifies.

7. CONCLUSION

Sophisticated phishing now exploits AI-generated content, real-time session interception, and multi-channel social engineering in combinations that defeat any single defensive technique. This review has surveyed the modern attack taxonomy and lifecycle, critically reviewed the detection literature across rule-based, classical-ML, deep-learning, transformer-based, visual, and behavioural families, examined the methodological practices used to evaluate that literature, and proposed a conceptual four-layer hybrid framework as a design template for future implementation and evaluation. We mapped the framework to the threat model to make its residual exposure to AiTM explicit, and argued that detection must be paired with phishing-resistant FIDO2 authentication, DMARC enforcement, human-centric training, and zero-trust containment. We outlined open research directions including adversarial robustness against LLM-driven attacks, deepfake voice phishing, privacy-preserving detection, explainability, standardised benchmarks, and authentication resilience. The trajectory of the threat is clear: phishing will continue to absorb every new generative-AI capability that emerges, and defensive research will need to do the same.

APPENDIX A. GLOSSARY OF KEY TERMS

- **Adversary-in-the-Middle (AiTM):** A real-time reverse-proxy attack that relays a victim's authentication to a legitimate service while capturing credentials and session cookies, potentially bypassing conventional multi-factor authentication.
- **Business Email Compromise (BEC):** A type of fraud in which an attacker impersonates an executive, employee, or trusted vendor to induce fraudulent payments, disclose sensitive information, or perform other unauthorized actions.
- **DMARC:** Domain-based Message Authentication, Reporting, and Conformance; an email-authentication framework based on SPF and DKIM that enables domain owners to specify how receiving servers should handle authentication failures.
- **FIDO2:** An authentication standard comprising WebAuthn and CTAP that provides phishing-resistant authentication by cryptographically binding authentication to the legitimate service's origin.
- **Quishing:** A form of QR-code phishing in which a malicious URL is embedded within a QR code and accessed through a device, potentially bypassing conventional URL-scanning mechanisms.
- **Smishing/Vishing:** Phishing attacks conducted through SMS messages (smishing) or voice calls (vishing), increasingly incorporating AI-generated or manipulated voice content.

- **Stacked Meta-Learner:** A second-stage machine-learning classifier that combines the probability outputs generated by multiple detection layers to produce a final classification or decision.
- **Zero Trust:** A security model based on continuous verification of identity, device posture, and contextual factors for each access request rather than assuming trust based on previous authentication or network location.

APPENDIX B. REPRESENTATIVE PUBLIC DATASETS

- **PhishTank/OpenPhish:** Publicly accessible collections of live or recently identified phishing URLs commonly used for URL-based phishing detection and classification. Their principal limitation is that entries may be short-lived and can introduce blacklist-related sampling bias.
- **UCI Phishing Websites Dataset [14]:** A widely used dataset containing engineered website and URL-related features for classical machine-learning benchmarking. Its main limitation is the use of a fixed feature set that may not fully represent current phishing techniques.
- **Vrbančič et al. Dataset [13]:** Provides URL-related data and engineered feature sets used in reproducible machine-learning studies. Because the data represent a particular collection period, performance may be affected by temporal drift.
- **Ranked Domain Lists:** Collections of benign URLs or domains commonly used for constructing the negative class in phishing-detection experiments. These datasets may exhibit popularity bias and may contain relatively few uncommon or edge-case websites.
- **Enterprise Mail Corpora:** Collections of email messages used for semantic, natural-language-processing, and phishing-detection research. Their primary limitation is limited public availability because of privacy, confidentiality, and organizational security requirements.

REFERENCES

- [1] Anti-Phishing Working Group (APWG), —Phishing Activity Trends Report, Q4 2023,|| APWG, Cambridge, MA, USA, Tech. Rep., 2024.
- [2] Verizon, —2023 Data Breach Investigations Report (DBIR),|| Verizon Business, New York, NY, USA, 2023.
- [3] K. L. Chiew, K. S. C. Yong, and C. L. Tan, —A survey of phishing attacks: Their types, vectors and technical approaches,|| *Expert Systems with Applications*, vol. 106, pp. 1–20, 2018.
- [4] A. Aleroud and L. Zhou, —Phishing environments, techniques, and countermeasures: A survey,|| *Computers & Security*, vol. 68, pp. 160–196, 2017.
- [5] R. Verma, N. Shashidhar, and N. Hossain, —Detecting phishing emails the natural language way,|| in *Proc. European Symposium on Research in Computer Security (ESORICS)*, Springer, 2012, pp. 824–841.
- [6] C. Whittaker, B. Ryner, and M. Nazif, —Large-scale automatic classification of phishing pages,|| in *Proc. Network and Distributed System Security Symposium (NDSS)*, 2010.
- [7] S. Marchal, J. François, R. State, and T. Engel, —PhishStorm: Detecting phishing with streaming analytics,|| *IEEE Transactions on Network and Service Management*, vol. 11, no. 4, pp. 458–471, 2014.
- [8] A. K. Jain and B. B. Gupta, —A machine learning based approach for phishing detection using hyperlinks information,|| *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, no. 5, pp. 2015–2028, 2019.
- [9] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, —Machine learning based phishing detection from URLs,|| *Expert Systems with Applications*, vol. 117, pp. 345–357, 2019.
- [10] M. Aassal, S. Baki, A. Das, and R. Verma, —An in-depth benchmarking and evaluation of phishing detection research for security needs,|| *IEEE Access*, vol. 8, pp. 22170–22192, 2020.
- [11] H. Le, Q. Pham, D. Sahoo, and S. C. H. Hoi, —URLNet: Learning a URL representation with deep learning for malicious URL detection,|| *arXiv preprint arXiv:1802.03162*, 2018.
- [12] Y. Huang, Q. Yang, J. Qin, and W. Wen, —Phishing URL detection via CNN and attention-based hierarchical RNN,|| in *Proc. IEEE TrustCom/BigDataSE*, 2019, pp. 112–119.
- [13] G. Vrbančič, I. Fister, and V. Podgorelec, —Datasets for phishing websites detection,|| *Data in Brief*, vol. 33, p. 106438, 2020.

- [14] R. Mohammad, F. Thabtah, and L. McCluskey, —Phishing websites features,|| UCI Machine Learning Repository, 2015.
- [15] A. Tang and N. R. Mabry, —Detecting AI-generated phishing emails with transformer models,|| in Proc. IEEE Int. Conf. on Cyber Security and Resilience (CSR), 2023, pp. 1–7.
- [16] Microsoft, —Cyber Signals: Adversary-in-the-Middle phishing,|| Microsoft Security Insider, Tech. Brief, 2023.
- [17] Proofpoint, —State of the Phish 2024,|| Proofpoint Inc., Sunnyvale, CA, USA, 2024.
- [18] NIST, —Special Publication 800-63B: Digital Identity Guidelines—Authentication and Lifecycle Management,|| National Institute of Standards and Technology, Gaithersburg, MD, 2017.
- [19] Internet Engineering Task Force (IETF), —RFC 7489: Domain-based Message Authentication, Reporting, and Conformance (DMARC),|| 2015.
- [20] FIDO Alliance, —FIDO2: WebAuthn & CTAP,|| 2023. [Online]. Available: <https://fidoalliance.org/fido2/>.
- [21] M. Khonji, Y. Iraqi, and A. Jones, —Phishing detection: A literature survey,|| IEEE Communications Surveys & Tutorials, vol. 15, no. 4, pp. 2091–2121, 2013.
- [22] P. Prakash, M. Kumar, R. R. Kompella, and M. Gupta, —PhishNet: Predictive blacklisting to detect phishing attacks,|| in Proc. IEEE INFOCOM, 2010, pp. 1–5.
- [23] A. K. Jain and B. B. Gupta, —Towards detection of phishing websites on client-side using machine learning based approach,|| Telecommunication Systems, vol. 68, no. 4, pp. 687–700, 2018.
- [24] E. M. Rudd, R. Harang, and J. Saxe, —MEADE: Towards a malicious email attachment detection engine,|| in Proc. IEEE Int. Symp. on Technologies for Homeland Security (HST), 2018, pp. 1–7.
- [25] T. Brown et al., —Language models are few-shot learners,|| in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020, pp. 1877–1901.
- [26] A. Vaswani et al., —Attention is all you need,|| in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017, pp. 5998–6008.
- [27] S. Abu-Nimeh, D. Nappa, X. Wang, and S. Nair, —A comparison of machine learning techniques for phishing detection,|| in Proc. eCrime Researchers Summit, 2007, pp. 60–69.
- [28] I. Fette, N. Sadeh, and A. Tomasic, —Learning to detect phishing emails,|| in Proc. 16th Int. Conf. on World Wide Web (WWW), 2007, pp. 649–656.
- [29] Google Safe Browsing Team, —Safe Browsing transparency report,|| Google, 2024. [Online]. Available: <https://transparencyreport.google.com/safe-browsing/>.
- [30] ENISA, —Threat Landscape 2023,|| European Union Agency for Cybersecurity, 2023.
- [31] M. Kaliappan, S. Vimal, P. Rebecca, S. Senthil, A. B, R. Singh, and G. Dhiman, “Q-learning adaptive security framework using digital twin artificial intelligence for self-healing network defence,” Journal of Animal Environment, vol. 18, no. 4S, 2026. [Online]. Available: <https://www.aejournals.org/index.php/AEJ/article/view/2886>.
- [32] G. Jyothsna Devi and R. Rani Chintala, “Multi-layer intrusion detection in smart IoT systems using voting classifier and explainable AI with Flask integration,” Journal of Animal Environment, vol. 18, no. 4S, pp. 174–180, 2026. [Online]. Available: <https://aejournals.org/index.php/AEJ/article/view/2945>.
- [33] Singh, A. R., & Kaur, A. (2025). AI adoption in agriculture: A managerial framework for sustainable agribusiness development. Journal of Artificial Intelligence in Agriculture and Applied Research, 1(1), 1–12. <https://agrpjournals.com/journals/journal/index.php/jaiaar/article/view/2>
- [34] Pandey, P. (2025). Artificial intelligence for early detection of dental caries from radiographic images: A machine learning. Journal of Dentistry and Clinical Sciences, 1(1), 31–39. <https://agrpjournals.com/journals/journal/index.php/jdacs/article/view/12>