

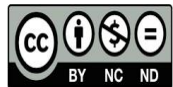
# Multimodal AI for Detecting Evolving Phishing Attacks and Social Engineering

Mohd Shahwez<sup>1</sup>, Vaibhav Kumar<sup>2</sup>

<sup>1</sup>M.Tech Student, Department of Computer Science Engineering & Application, Bhagwnat Institute of Technology, Muzaffarnagar, Uttar Pradesh, India.

<sup>2</sup>Assistant Professor, Department of Computer Science Engineering & Application, Bhagwnat Institute of Technology, Muzaffarnagar, Uttar Pradesh, India.

**To Cite this Article:** Mohd Shahwez<sup>1</sup>, Vaibhav Kumar<sup>2</sup>, “Multimodal AI for Detecting Evolving Phishing Attacks and Social Engineering”, Indian Journal of Computer Science and Technology, Volume 05, Issue 03 (September-December 2026), PP: 408-412.



Copyright: ©2026 This is an open access journal, and articles are distributed under the terms of the [Creative Commons Attribution License](#); Which Permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Abstract:** Phishing and social engineering combine technical impersonation with psychological manipulation, exploiting urgency, authority, trust, fear, and curiosity to induce unsafe actions. Detection systems based on a single signal—such as a URL blacklist, lexical URL features, or message text—may fail when attackers rotate domains, paraphrase lures, clone legitimate interfaces, or move between communication channels. This paper presents a research framework for multimodal phishing detection that integrates four evidence streams: message text and persuasion cues; URL and domain structure; webpage HTML and visual appearance; and contextual technical metadata. Each modality is processed by a specialized encoder, while early, late, and confidence-aware gated fusion are compared under a shared evaluation protocol. The proposed protocol prioritizes temporal and domain-separated testing, duplicate control, calibration, explainability, missing-modality resilience, and inference cost. Candidate sources include public phishing webpage benchmarks and threat feeds, subject to licensing and label-quality checks. The framework is a research proposal rather than a report of completed experiments; therefore, no performance results are asserted. Its intended contribution is a reproducible way to test when multimodal evidence improves detection of evolving campaigns and when additional complexity fails to justify its cost.

**Key Word:** phishing detection, social engineering, multimodal learning, cybersecurity, explainable AI, temporal evaluation, URL analysis.

## I. INTRODUCTION

Phishing is a deceptive attempt to make a person disclose credentials, transfer funds, reveal sensitive information, or open malicious content by impersonating a trusted individual or service. Social engineering is the broader manipulation process behind many such attacks: the lure may invoke urgency, authority, fear of account suspension, a promised reward, secrecy, or a familiar brand. Modern campaigns can reach targets through email, SMS, direct messages, social networks, QR codes, and cloned webpages. Their technical and psychological components are related but not identical; a persuasive message can point to a malicious destination, while a visually convincing page can hide a suspicious domain [12], [17], [24], [25].

Traditional controls—including blocklists, reputation feeds, hand-crafted rules, and single-modality classifiers—remain useful, but they have different failure modes. Blocklists may not yet contain newly registered or short-lived domains. URL-only models cannot directly assess brand impersonation on a page. Text-only models may miss destination inconsistencies, and screenshot-only models may overlook sender identity, redirect behaviour, or hidden form destinations. The challenge is compounded by changing attacker infrastructure and by evaluation protocols that can overestimate generalization when related domains or near-duplicate pages appear in both training and test sets [7], [10], [16], [18].

Recent research already demonstrates multimodal phishing analysis, including combinations of message content, persuasion strategies, URL structure, webpage appearance, and technical evidence [1]–[6]. Accordingly, multimodality itself is not claimed as a new idea here. The research problem is more specific: how should complementary signals be fused and evaluated so that any benefit remains visible on later campaigns, unseen domains, and incomplete observations, while keeping false alarms, latency, and privacy risk within practical limits?

### The paper makes three intended contributions:

- A modular four-branch architecture that preserves the provenance and availability of each evidence stream.
- A controlled evaluation protocol comparing strong unimodal baselines with early, late, and gated fusion under temporal, domain-held-out, and missing-modality conditions.
- An operational evaluation plan covering calibration, evidence-based explanations, abstention, inference cost, and safe handling of malicious content.

### 1.1 Threat model and research questions

The assumed adversary seeks to persuade a target to follow a malicious link, disclose credentials, approve a payment, or trust an impersonated service. The adversary may register short-lived domains, alter URL tokens, use redirects or shared hosting,

paraphrase message text, imitate a brand's visual identity, and distribute the same lure across multiple channels. The defender is assumed to have access only to lawfully obtained messages, URLs, page artifacts, and metadata available at the time of analysis. The study does not require interaction with attacker infrastructure or deployment of live phishing content. This bounded threat model makes the experiments repeatable while acknowledging that real campaigns may use additional signals unavailable to the prototype.

Four questions guide the evaluation: (RQ1) Does fusion improve macro-F1 or PR-AUC over the strongest unimodal baseline on temporally separated data? (RQ2) Which combinations of message, URL, visual, and contextual evidence provide the greatest incremental value? (RQ3) How much performance is lost when a modality is absent, noisy, or perturbed? (RQ4) Can calibrated confidence and evidence summaries support safer triage without producing an unacceptable false-positive rate or excessive review workload? These questions are deliberately empirical; no outcome is assumed in advance.

## II. RELATED WORK

Early machine-learning studies established email- and webpage-based phishing detection as classification problems. Fette et al. examined learning-based phishing email detection [14], Abu-Nimeh et al. compared machine-learning techniques [15], and CANTINA+ combined webpage features with filtering rules [13]. URL-oriented methods later used lexical and language-derived features [16], while URLNet learned character- and word-level representations directly from URL strings [11]. These approaches provide essential baselines, but a URL representation alone cannot establish whether a rendered page visually impersonates a legitimate service.

Visual methods address this limitation by comparing webpage appearance and brand-related features. VisualPhishNet uses learned visual similarity for zero-day phishing pages [12], while subsequent work investigates global and local visual cues [9]. However, visual similarity can be evaded through logo manipulation, page-layout changes, or partial branding; it can also produce false positives when legitimate services share templates [10]. Multimodal language models have also been studied for webpage identification and brand/domain consistency [8].

The newest literature moves beyond single-modality classification. Bustio-Martínez et al. combine textual, URL, and persuasion-related signals with explanations [1]; PhishAgent uses multimodal information retrieval for webpage detection [2]; and a systematic review considers multimodal phishing on social-networking sites [3]. Other recent work studies behaviour-aware multimodal detection [4], cross-modal semantic and visual cues [5], and explainable multimodal webpage analysis [6]. PhreshPhish provides a large-scale phishing/benign HTML–URL benchmark designed to reduce label-quality and leakage problems [7]. These developments make a carefully delimited contribution essential: the present framework focuses on a common comparison protocol, temporal generalization, missing evidence, and calibrated decision support rather than claiming that multimodal phishing detection is unprecedented.

## III. PROPOSED MULTIMODAL FRAMEWORK

Let each observation be represented as  $x = \{m, u, v, c\}$ , where  $m$  is message text,  $u$  is URL/domain evidence,  $v$  is webpage HTML or visual evidence, and  $c$  is contextual technical metadata. A binary label  $y$  distinguishes phishing from legitimate content. Because some channels will be unavailable, the input also contains an availability mask  $a$ . The model estimates a calibrated probability  $p(y = \text{phishing} \mid x, a)$  and returns a risk class, a concise evidence summary, and an abstention or human-review recommendation when the available signals are insufficient or contradictory.

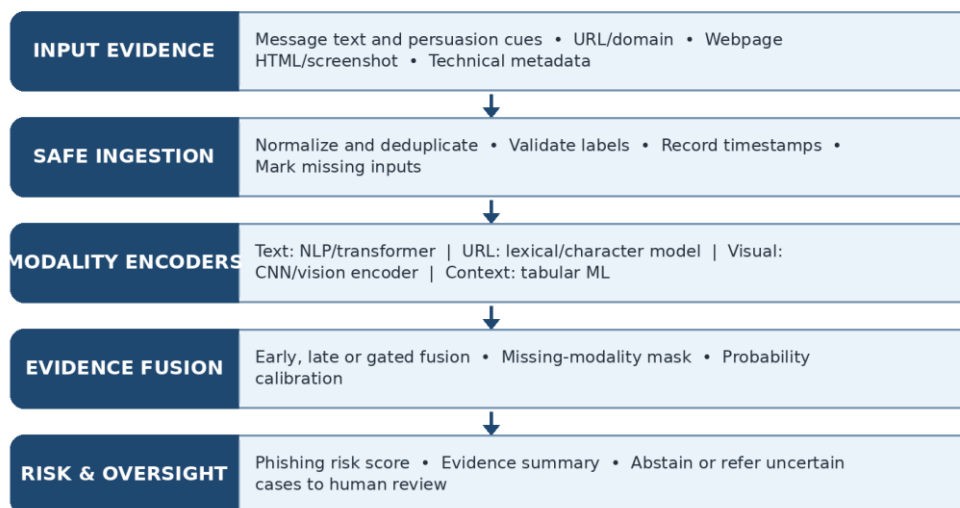


Fig. 1. Proposed evidence-fusion architecture for evolving phishing detection.

### 3.1 Modality-specific evidence

Text branch. Message subject, body, and available conversational context will be normalized and represented using TF–IDF with Logistic Regression or a linear SVM as transparent baselines. A compact transformer encoder may be evaluated where data volume and compute permit. Additional features can represent urgency, authority claims, threats, rewards, requests for

secrecy, credential requests, and unusual payment instructions. These are probabilistic cues—not proof of malicious intent—because legitimate messages can also be urgent [1], [17].

**URL/domain branch.** Features may include URL length, subdomain depth, character distribution, encoded strings, punycode, IP-address hosts, suspicious tokens, redirect count, and domain-reputation attributes available at prediction time. Random Forest, Gradient Boosting, and Logistic Regression will be compared with a character-level URL encoder inspired by URLNet [11], [16]. The model will not use post-event information or label-derived features unavailable in deployment.

**Visual and page-content branch.** For accessible HTML or screenshots, a pretrained CNN or vision encoder can represent page layout, logo placement, forms, colour composition, and visual similarity. HTML/DOM features may capture form destinations, hidden elements, resource patterns, and inconsistencies between a claimed brand and the actual domain. This branch is intended to identify impersonation that is difficult to infer from text or URL strings alone [8]–[10], [12].

**Context branch.** Structured metadata may include sender/domain relationships, available certificate or TLS attributes, redirect behaviour, hosting context, and threat-feed reputation. Shared hosting and content-delivery networks complicate attribution; therefore, an infrastructure feature must not be treated as inherently malicious without page- or tenant-level evidence [24], [25]. Each feature will have a documented definition, availability condition, extraction cost, and leakage risk.

| Modality                  | Representative evidence                     | Main limitation                                       |
|---------------------------|---------------------------------------------|-------------------------------------------------------|
| Text / social engineering | Semantics, urgency, authority, requests     | Paraphrase, multilingual variation, benign overlap    |
| URL / domain              | Lexical tokens, redirects, domain signals   | Domain churn, obfuscation, shared infrastructure      |
| Visual / HTML             | Brand similarity, layout, form and DOM cues | Screenshot drift, logo manipulation, capture failures |
| Technical context         | Sender relation, TLS/DNS, reputation        | Missing or stale metadata; attribution ambiguity      |

*Table 1. Modalities, representative evidence, and principal limitations*

### 3.2 Fusion and decision support

Three fusion strategies will be compared. Early fusion concatenates standardized features or embeddings before classification and includes explicit missing-modality indicators. Late fusion combines calibrated branch-level probabilities through validation-set weights or a small meta-classifier. Gated fusion learns sample-dependent weights so that a weak, absent, or contradictory branch contributes less. All variants will be compared on identical data splits. Calibration will be assessed before thresholds are selected; low-confidence cases may be abstained from and sent for human review rather than automatically blocked.

Explanations will be matched to model type: feature attributions such as SHAP for structured models, highlighted text spans for text models, and saliency or matched visual regions where the selected encoder supports them. Such outputs describe model evidence and must not be presented as causal proof. The framework follows a human-oversight approach consistent with risk-management guidance for AI systems [22], while adversarial robustness will be considered in light of established taxonomy and terminology [23].

## IV. RESEARCH METHODOLOGY

### 4.1 Data sources and preparation

Candidate sources include PhreshPhish for webpage HTML–URL pairs [7], OpenPhish and PhishTank for threat-indicator enrichment or validation [27], [28], and the UCI Phishing Websites collection for a classical structured-feature baseline [26]. Dataset access, redistribution conditions, timestamp coverage, label provenance, and modality availability will be checked before inclusion. The UCI collection is an older feature-engineered benchmark and will not be treated as a substitute for contemporary temporal testing. A benign email corpus such as Enron, if used, will be labelled only as a source of benign language and not represented as a phishing-labelled dataset.

Exact duplicates will be removed before splitting, and near-duplicates will be identified using URL, content-hash, and page-similarity checks. The primary split will be chronological, reserving the latest feasible interval for testing; a domain- or campaign-held-out split will be added where identifiers permit. Tokenizers, scalers, feature selection, class reweighting, and resampling will be fitted on training data only. Each source will be documented in a dataset card containing labels, exclusions, time coverage, licensing, missing modalities, and known bias. These controls respond to documented benchmark-leakage and base-rate concerns [7].

### 4.2 Models and experimental comparisons

The baseline set will include TF–IDF plus Logistic Regression or linear SVM for text, structured classifiers for URL/context features, a character-level URL model, and a pretrained visual encoder where usable screenshots exist. The main experiments are: (E1) unimodal baselines; (E2) early versus late versus gated fusion; (E3) random-split diagnostic versus temporal and domain-held-out evaluation; (E4) deliberate removal of each modality; (E5) controlled URL obfuscation, wording shifts, and visual/logo perturbations; (E6) calibration, abstention, and explanation review; and (E7) latency and memory measurement. Hyperparameters will be selected on validation data; the final test set will remain untouched until evaluation.

### 4.3 Metrics and analysis

Accuracy will be reported but will not be the sole criterion. Precision, recall, macro-F1, PR-AUC, and false-positive rate will capture detection performance and class imbalance. Brier score or expected calibration error will assess probability quality. The study will record average and percentile latency, model size or peak memory, and the proportion routed to human review. Results will be reported by channel and modality availability when sample sizes support interpretation. Repeated runs and bootstrap confidence intervals will be used where feasible, with paired comparisons on the same test examples. No numerical result is assumed in advance.

### 4.4 Ethics, privacy, and safe handling

Only lawfully accessible datasets and permitted research feeds will be used. The study will not collect real passwords, reuse authentication tokens, send deceptive messages to unsuspecting users, or interact with credential-capture forms. Potentially malicious pages will be acquired only in an isolated, restricted environment; automatic downloads and unnecessary outbound access will be disabled where feasible. Personal data will be minimized or redacted, raw artifacts access-controlled, and retention decisions documented. Public reporting will emphasize defensive findings and avoid operational details that could facilitate phishing.

### 4.5 Result reporting and reproducibility

The final report will separate development-set decisions from held-out test findings. For every model, it will identify the dataset version, feature availability, split dates, deduplication procedure, hyperparameter-selection method, random seeds, and decision threshold. A primary comparison table will report precision, recall, macro-F1, PR-AUC, false-positive rate, and calibration alongside uncertainty intervals where feasible. A second analysis will report relative changes between the best unimodal baseline and each fusion variant, while ablation tables will show performance with one evidence stream removed. Channel- or language-specific results will be reported only where sample sizes support stable interpretation.

Reproducibility materials should include data-processing scripts, model configurations, evaluation code, and a data card describing licensing and known biases. Where raw pages or URLs cannot be redistributed safely or lawfully, the paper should provide source identifiers and access instructions rather than republishing restricted artifacts. Error analysis will classify false positives and false negatives by observable cause—for example, benign urgency language, shared hosting, missing screenshots, or visual impersonation—without exposing functioning credential-harvesting links. These reporting requirements are part of the research contribution because they allow other researchers to distinguish genuine temporal generalization from dataset-specific effects.

## V. DISCUSSION AND LIMITATIONS

The expected scientific value lies in establishing when additional modalities provide complementary information under realistic testing, rather than assuming that a larger model is always better. Fusion may help when a message, destination, and rendered page reveal mutually reinforcing inconsistencies; it may offer little benefit when one branch is noisy, missing, stale, or computationally expensive. A simple text–URL model may therefore remain preferable for low-latency channels, while screenshot analysis may be justified for high-risk webpage investigations. The evaluation should report positive, negative, and conditional findings.

Threats to validity include incomplete labels, uneven language and brand coverage, changes in webpage rendering, domain churn, dataset selection bias, and limited access to private enterprise traffic. A model can also learn shortcuts such as host reputation or template artifacts that fail on new campaigns. Time-separated tests reduce but do not eliminate these risks. Explanations can be unstable, and offline benchmark success does not establish production readiness. Before claiming novelty or deployment value, the completed study must compare against current methods, report actual experimental outcomes, and disclose reproducibility artifacts.

## VI. CONCLUSION

This paper defines a modular multimodal research framework for detecting evolving phishing and social-engineering attempts through text, URL/domain, visual webpage, and contextual technical evidence. Its central methodological emphasis is a fair comparison of unimodal baselines and fusion strategies using temporal and domain-separated evaluation, explicit missing-modality tests, calibration, explanations, and operational-cost measurements. The framework is designed to produce a defensible empirical contribution without presuming that multimodal fusion will always improve performance. The next step is implementation and evaluation on permitted datasets; only those results can establish the size, stability, and practical value of any observed benefit.

## REFERENCES

1. L. Bustio-Martínez et al., “Enhanced phishing detection using multimodal data,” *Knowledge-Based Systems*, vol. 334, art. 115105, 2026, doi: 10.1016/j.knosys.2025.115105.
2. T. Cao, C. Huang, Y. Li, H. Wang, A. He, N. Oo, and B. Hooi, “PhishAgent: A Robust Multimodal Agent for Phishing Webpage Detection,” *Proc. AAAI Conf. Artificial Intelligence*, vol. 39, no. 27, pp. 27869–27877, 2025, doi: 10.1609/aaai.v39i27.35003.
3. T. Wangchuk and T. Gonsalves, “Multimodal Phishing Detection on Social Networking Sites: A Systematic Review,” *IEEE Access*, vol. 13, pp. 103405–103416, 2025, doi: 10.1109/ACCESS.2025.3579584.
4. M. Murhej and G. Nallasivan, “Multimodal framework for phishing attack detection and mitigation through behavior analysis using EM-BERT and SPCA-BASED EAI-SC-LSTM,” *Frontiers in Communications and Networks*, vol. 6, art. 1587654, 2025, doi:

- 10.3389/frcmn.2025.1587654.
5. "CrossPhire: Benefiting Multimodality for Robust Phishing Web Page Identification," *Applied Sciences*, vol. 16, no. 2, art. 751, 2026, doi: 10.3390/app16020751.
6. "A Multimodal Phishing Website Detection System Using Explainable Artificial Intelligence Technologies," *Machine Learning and Knowledge Extraction*, vol. 8, no. 1, art. 11, 2026, doi: 10.3390/make8010011.
7. T. Dalton et al., "PhreshPhish: A Real-World, High-Quality, Large-Scale Phishing Website Dataset and Benchmark," arXiv:2507.10854, 2025. [Online]. Available: <https://arxiv.org/abs/2507.10854>
8. J. Lee, P. Lim, B. Hooi, and D. M. Divakaran, "Multimodal Large Language Models for Phishing Webpage Detection and Identification," arXiv:2408.05941, 2024. [Online]. Available: <https://arxiv.org/abs/2408.05941>
9. M. Wang, L. Song, L. Li, Y. Zhu, and J. Li, "Phishing webpage detection based on global and local visual similarity," *Expert Systems with Applications*, vol. 252, art. 124120, 2024, doi: 10.1016/j.eswa.2024.124120.
10. F. Ji et al., "Evaluating the Effectiveness and Robustness of Visual Similarity-based Phishing Detection Models," arXiv:2405.19598, 2024. [Online]. Available: <https://arxiv.org/abs/2405.19598>
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:1802.03162, 2018. [Online]. Available: <https://arxiv.org/abs/1802.03162>
12. S. Abdelnabi, K. Krombholz, and M. Fritz, "VisualPhishNet: Zero-Day Phishing Website Detection by Visual Similarity," arXiv:1909.00300, 2019. [Online]. Available: <https://arxiv.org/abs/1909.00300>
13. G. Xiang, J. Hong, C. P. Rose, and L. Cranor, "CANTINA+: A Feature-Rich Machine Learning Framework for Detecting Phishing Web Sites," *ACM Transactions on Information and System Security*, vol. 14, no. 2, art. 21, 2011, doi: 10.1145/2019599.2019606.
14. I. Fette, N. Sadeh, and A. Tomasic, "Learning to detect phishing emails," in *Proc. 16th International Conference on World Wide Web*, 2007, pp. 649–656, doi: 10.1145/1242572.1242660.
15. 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, doi: 10.1145/1299015.1299021.
16. 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, doi: 10.1016/j.eswa.2018.09.029.
17. R. Dhamija, J. D. Tygar, and M. A. Hearst, "Why phishing works," in *Proc. SIGCHI Conference on Human Factors in Computing Systems*, 2006, doi: 10.1145/1124772.1124861.
18. D. Sahoo, C. Liu, and S. C. H. Hoi, "Malicious URL Detection using Machine Learning: A Survey," arXiv:1701.07179, 2017. [Online]. Available: <https://arxiv.org/abs/1701.07179>
19. S. Kavya and D. Sumathi, "Staying ahead of phishers: a review of recent advances and emerging methodologies in phishing detection," *Artificial Intelligence Review*, vol. 58, art. 50, 2025, doi: 10.1007/s10462-024-11055-z.
20. E. Kytidou, T. Tsirikiki, G. Drosatos, and K. Rantos, "Machine learning techniques for phishing detection: A review of methods, challenges, and future directions," 2025, doi: 10.1177/18724981251366763.
21. "Phishing or Not Phishing? A Survey on the Detection of Phishing Websites," *IEEE Access*, vol. 11, pp. 18499–18519, 2023, doi: 10.1109/ACCESS.2023.3247135.
22. National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1, 2023, doi: 10.6028/NIST.AI.100-1.
23. National Institute of Standards and Technology, *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations*, NIST AI 100-2e2025, 2025. [Online]. Available: <https://csrc.nist.gov/pubs/ai/100/2/e2025/final>
24. Cybersecurity and Infrastructure Security Agency, "Avoiding Social Engineering and Phishing Attacks." [Online]. Available: <https://www.cisa.gov/news-events/news/avoiding-social-engineering-and-phishing-attacks>
25. MITRE, "Phishing (T1566), MITRE ATT&CK Enterprise Techniques." [Online]. Available: <https://attack.mitre.org/techniques/T1566/>
26. UCI Machine Learning Repository, "Phishing Websites Data Set," dataset 327. [Online]. Available: <https://archive.ics.uci.edu/dataset/327/phishing+websites>
27. OpenPhish, "OpenPhish Phishing Database." [Online]. Available: [https://www.openphish.com/phishing\\_database.html](https://www.openphish.com/phishing_database.html)
28. PhishTank, "PhishTank: Join the fight against phishing." [Online]. Available: <https://phishtank.org/>
29. H. M. Le and V. T. Nguyen, "A secure and explainable phishing detection framework using deep learning and LLM-based reasoning," *Discover Informatics*, vol. 1, art. 20, 2026, doi: 10.1007/s44564-026-00020-3.
30. C. Mayo, M. Tchuindjang, S. Brohi, and N. Ersotelos, "A Temporally Dynamic Feature-Extraction Framework for Phishing Detection with LIME and SHAP Explanations," *Future Internet*, vol. 18, no. 2, art. 101, 2026, doi: 10.3390/fi18020101.