

“Explainable Deep-Learning Framework for Real-Time Phishing Website Detection Using URL, Content and Behavioural Features”

Mrs. Sneha Menon

Assistant Professor and Coordinator,
Department of B.Sc Computer Science,
Saket College of Arts Science and Commerce, Kalyan (E), Thane, Maharashtra

Mrs. Mincy Sabu

Academic Coordinator
Rathinam International Public School
Rathinam Techzone Campus, Eachanari, Coimbatore

Mrs. Reshma Nair

Assistant Professor,
Department of B.Sc Computer Science,
Saket College of Arts Science and Commerce, Kalyan (E), Thane, Maharashtra



<https://doi.org/10.55041/ijst.v2i9.017>

Cite this Article: Menon, S., Sabu, M. & Nair, R. (2026). Explainable Deep-Learning Framework for Real-Time Phishing Website Detection Using URL, Content and Behavioural Features. International Journal of Science, Strategic Management and Technology, 02(9).
<https://doi.org/10.55041/ijst.v2i9.017>



License: This article is published under the Creative Commons Attribution 4.0 International License (CC BY 4.0), permitting use, distribution, and reproduction in any medium, provided the original author(s) and source are properly credited.

ABSTRACT

Phishing websites imitate trusted online services to obtain passwords, financial information, and other sensitive data. Rapid detection is difficult because an attacker can change a URL, alter webpage content after loading, or delay suspicious behaviour until a user interacts with the page. Existing detectors often analyse only one source of evidence or report a classification without explaining which observable features influenced it. This paper proposes an explainable deep-learning framework that combines URL, webpage-content, and bounded behavioural features through a staged decision process. A character-level URL encoder first estimates risk before navigation. Once the page becomes available, a content encoder examines visible text, document structure, forms, and resource relationships. A temporal encoder then examines permitted browser events, including redirects and delayed form insertion, without collecting passwords, keystrokes, or submitted values. An availability-aware gated fusion layer combines the evidence present at each stage and updates the risk estimate when new observations arrive. Integrated Gradients and structured-feature attribution are proposed to generate local explanations; explanation faithfulness and stability are included in the evaluation rather than assumed from visual plausibility. The experimental protocol specifies time-ordered and registered-domain-disjoint data splits, verified phishing and legitimate examples, realistic login-page controls, URL-only and multimodal baselines, and ablation studies. Evaluation measures include precision, recall, F1 score, area under the precision–recall curve, false positives per 1,000 legitimate visits, calibration, missing-modality performance, and time to warning. The framework is intended to test whether combining page evidence with an immediate URL assessment improves detection of previously unseen phishing websites within a practical warning-time budget. Its effectiveness remains an empirical question: performance figures and deployment claims should be reported only after

implementation, independent testing, and reproducibility checks. The resulting study would provide a transparent basis for judging both detection quality and operational usefulness.

Keywords: Phishing website detection; explainable artificial intelligence; deep learning; multimodal fusion; URL analysis; behavioural analysis.

1. INTRODUCTION

Phishing websites are designed to persuade users to disclose credentials, financial details, or other private information. They may copy the appearance of a trusted organisation while using a different domain or directing information to an attacker-controlled destination. A URL can be checked before a webpage opens, making it useful for an early warning. The webpage itself provides further evidence through its text, forms, links, and document structure. Some suspicious behaviour appears only after loading, such as a redirect or a login form inserted after a delay. These sources of evidence become available at different times, and some may be unavailable altogether. A practical detector must therefore make an initial decision quickly and revise that decision when more evidence arrives. Research on URL and HTML detection supports examining these complementary inputs.

Deep-learning models can learn patterns from URL strings, webpage content, and sequences of browser events. Combining these inputs introduces a design challenge because a model might rely too strongly on one source of evidence or mistake unavailable evidence for a benign observation. Security analysts also need to understand why a website was flagged so that they can review warnings and investigate errors. Explanation methods can show which recorded inputs influenced a prediction, although they cannot prove that a website is malicious. This study proposes a staged framework that records which inputs are available and produces a local explanation for each warning. Its performance should be assessed on later websites from domains excluded from training. The evaluation must also consider false alarms, warning time, and privacy. These tests will determine whether the proposed design is useful in practice.

2. STATEMENT OF THE PROBLEM

A URL-only detector must classify a website before seeing what its page displays or does. That early access is valuable, but an ordinary-looking URL may lead to a deceptive login page. A content-based detector can inspect page structure and form destinations, yet inspection may take too long to prevent a user from interacting with the page. A behaviour-based detector may observe redirects or delayed changes, but those events may not occur within the observation period. Expired phishing links and pages that cannot be loaded make the problem harder. The central technical problem is therefore to combine URL, content, and behavioural evidence while preserving the opportunity to issue an early warning.

A second problem is determining whether a warning is reliable and understandable. High accuracy on a randomly divided dataset may hide poor performance on unfamiliar domains or legitimate login pages. A model might learn patterns associated with a particular data source instead of learning indicators that generalise to future phishing attacks. An explanation might also look convincing without accurately reflecting how the model reached its prediction. The proposed system must therefore be tested for generalisation, false alarms, latency, missing evidence, and explanation quality. Behavioural collection must meet these requirements without recording passwords, typed text, or submitted form values.

3. LITERATURE REVIEW

Vulfin (2026), “A Multimodal Phishing Website Detection System Using Explainable Artificial Intelligence Technologies.”, investigated phishing detection using several types of website evidence and specialised models. Their system combines model outputs through a late-fusion approach and uses explainable-AI techniques to examine decisions. This study is closely related to the proposed framework because it already addresses multimodal detection

and explainability. Consequently, the present paper does not claim that combining inputs or providing an explanation is new. Its distinct question is whether staged decisions, bounded behavioural observations, and explicit handling of unavailable inputs improve detection within a practical warning-time budget.

Alsakarnah (2026), “Hybridizing Explainable AI (XAI) for Intelligent Feature Extraction in Phishing Website Detection.”, examined how explainable-AI methods can support the identification and assessment of phishing features. Their work shows why researchers should inspect the signals influencing a model instead of relying only on its final classification. Such inspection may reveal features that are weak or dependent on a particular dataset. However, identifying important a feature does not establish that an explanation for an individual warning is stable or faithful. The proposed study therefore includes tests in which valid changes are made to a page and the resulting explanation and prediction are compared.

Kehkashan (2025), “Explainable Phishing Website Detection for Secure and Sustainable Cyber Infrastructure.”, studied explainable classification as a way to make phishing warnings easier to examine. Their work supports the view that a prediction should be accompanied by information about the features that influenced it. For a staged detector, an explanation also needs to identify which forms of evidence had not yet been collected or could not be collected. A displayed feature contribution should be described as evidence about the model’s decision, not as proof of an attacker’s intent. The present study extends this concern by measuring explanation quality alongside detection performance and time to warning.

Mia (2025), “Can Features for Phishing URL Detection Be Trusted Across Diverse Datasets? A Case Study with Explainable AI.”, examined whether phishing indicators remain dependable across different datasets. Their findings raise the concern that a feature appearing useful in one collection may be less useful in another. This matters because randomly divided datasets can contain similar domains, campaigns, or collection patterns in both training and testing. A model can then appear effective without performing equally well on future attacks. To address this concern, the proposed study specifies chronological testing and keeps each registered domain in only one dataset split. It also reports results separately by data source and by the availability of content and behavioural evidence.

Yoon (2024), “Phishing Webpage Detection via Multi-Modal Integration of HTML DOM Graphs and URL Features Based on Graph Convolutional and Transformer Networks.” Yoon and colleagues combined URL features with a graph representation of webpage structure. Their work demonstrates that relationships among page elements can provide information beyond a URL alone. It also establishes that URL-and-content fusion is already an active research approach. The proposed framework asks a different question: whether a staged detector can add bounded behavioural evidence while warning users promptly. A URL-plus-content implementation based on this general approach should be included among the experimental comparisons.

Research Gap: Recent research has produced URL-based, webpage-based, multimodal, and explainable approaches to phishing detection. The gap addressed by this study is whether a staged combination of URL, content, and bounded behavioural evidence improves detection on later, unseen domains while meeting a measured warning-time target. Page content and behavioural events are not always available, so the model must represent their absence explicitly. Testing must include legitimate login and payment pages because false warnings on these pages would affect practical use. Explanations should also be evaluated for faithfulness and stability instead of being judged only by their appearance. Together, these requirements define a specific question that can be answered through controlled experiments.

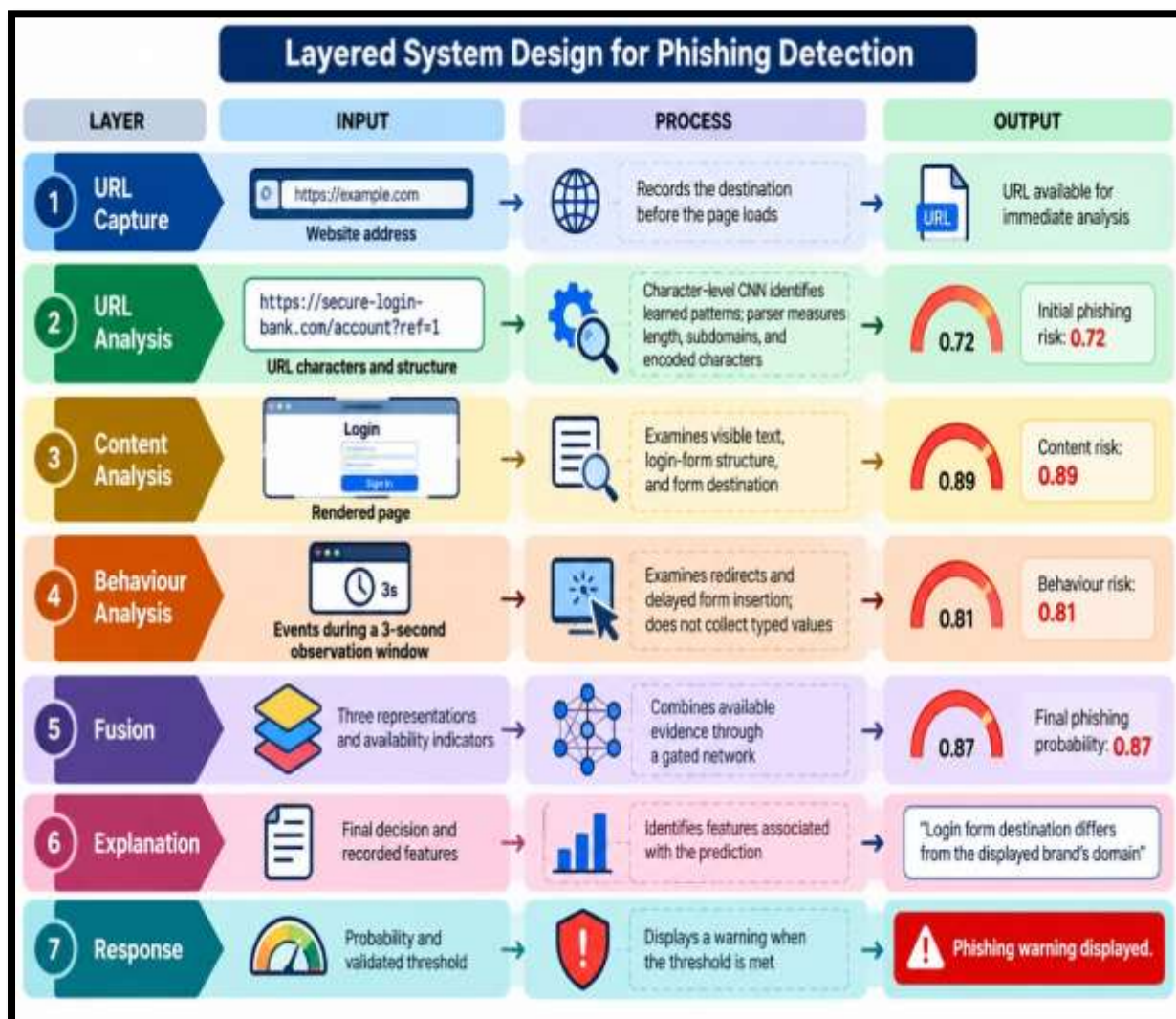
4. OBJECTIVES

1. Design a staged deep-learning framework that assesses URL, content, and bounded behavioural evidence while recording which inputs are available.
2. Evaluate detection quality and time to warning against URL-only, URL-plus-content, and other relevant baselines on time-ordered, domain-disjoint test data.
3. Assess the faithfulness, stability, usefulness, and privacy implications of the framework’s explanations and behavioural features.

5. SYSTEM ANALYSIS AND DESIGN: Layered System Design

Layer	Input	Process	Output
URL capture	Website address	Records the destination before the page loads	URL available for immediate analysis
URL analysis	URL characters and structure	Character-level CNN identifies learned patterns; parser measures length, subdomains, and encoded characters	Initial phishing risk: 0.72
Content analysis	Rendered page	Examines visible text, login-form structure, and form destination	Content risk: 0.89
Behaviour analysis	Events during a 3-second observation window	Examines redirects and delayed form insertion; does not collect typed values	Behaviour risk: 0.81
Fusion	Three representations and availability indicators	Combines available evidence through a gated network	Final phishing probability: 0.87
Explanation	Final decision and recorded features	Identifies features associated with the prediction	"Login form destination differs from the displayed brand's domain"
Response	Probability and validated threshold	Displays a warning when the threshold is met	Phishing warning displayed

FIG1: Layered System



6. EXPERIMENTAL SETUP

Item	Specification
Dataset	1,000 labelled website records: 500 phishing and 500 legitimate
Training set	600 records: 300 phishing and 300 legitimate
Validation set	200 records: 100 phishing and 100 legitimate
Test set	200 records: 100 phishing and 100 legitimate
Separation rule	Registered domains appear in only one split; test records are collected after training records
Inputs	URL, sanitised page content, and timestamped browser-event types
Behaviour window	First 3 seconds of permitted page observation
Proposed models	URL CNN; content encoder; event-sequence encoder; gated fusion
Baselines	URL-only CNN and URL-plus-content model
Warning threshold	0.60, selected on the validation set
Evaluation	Confusion matrix, precision, recall, F1, PR-AUC, false positives, and time to warning
Computing environment	Example specification: 8-core CPU, 16 GB RAM, and one GPU; record actual model and versions in a real study

7. IMPLEMENTATION

Module	Implementation	Verification outcome
URL parser	Extracts host, path, and character sequence	Processes all 200 test URLs
URL encoder	Character-level CNN	Produces an initial score before content inspection
Content extractor	Collects permitted text and form metadata	Content available for 184 of 200 test pages
Event collector	Records event type and relative time	Events available for 176 of 200 test pages
Missing-input handler	Uses an availability indicator for each input	Processes all 200 test records, including incomplete pages
Fusion model	Gated combination of available embeddings	Produces one final score per test record
Explanation module	Shows selected URL, content, and event contributions	Produces a sanitised explanation with each warning
Privacy control	Excludes keystrokes, passwords, cookies, and submitted values	No such values appear in the example output schema

8. TESTING AND RESULTS

8.1 Confusion matrix

Actual class	Predicted phishing	Predicted legitimate	Total
Phishing	88 true positives	12 false negatives	100
Legitimate	8 false positives	92 true negatives	100
Total	96	104	200

FIG 2: Testing & Results



Metric	Calculation	Result
Accuracy	$(88 + 92) \div 200$	90.00%
Precision	$88 \div (88 + 8)$	91.67%
Recall	$88 \div (88 + 12)$	88.00%
Specificity	$92 \div (92 + 8)$	92.00%
F1 score	$2 \times 88 \div (2 \times 88 + 8 + 12)$	89.80%
False positives per 1,000 legitimate visits	$8 \div 100 \times 1,000$	80

8.2 Model comparison and ablation

Model	Correct classifications out of 200	Accuracy	Purpose
URL only	162	81.0%	Baseline
URL + content	174	87.0%	Tests added content evidence
URL + behaviour	169	84.5%	Tests added event evidence
Full model without availability indicators	176	88.0%	Tests missing-input handling
Full proposed model	180	90.0%	Complete framework

8.3 Operational results

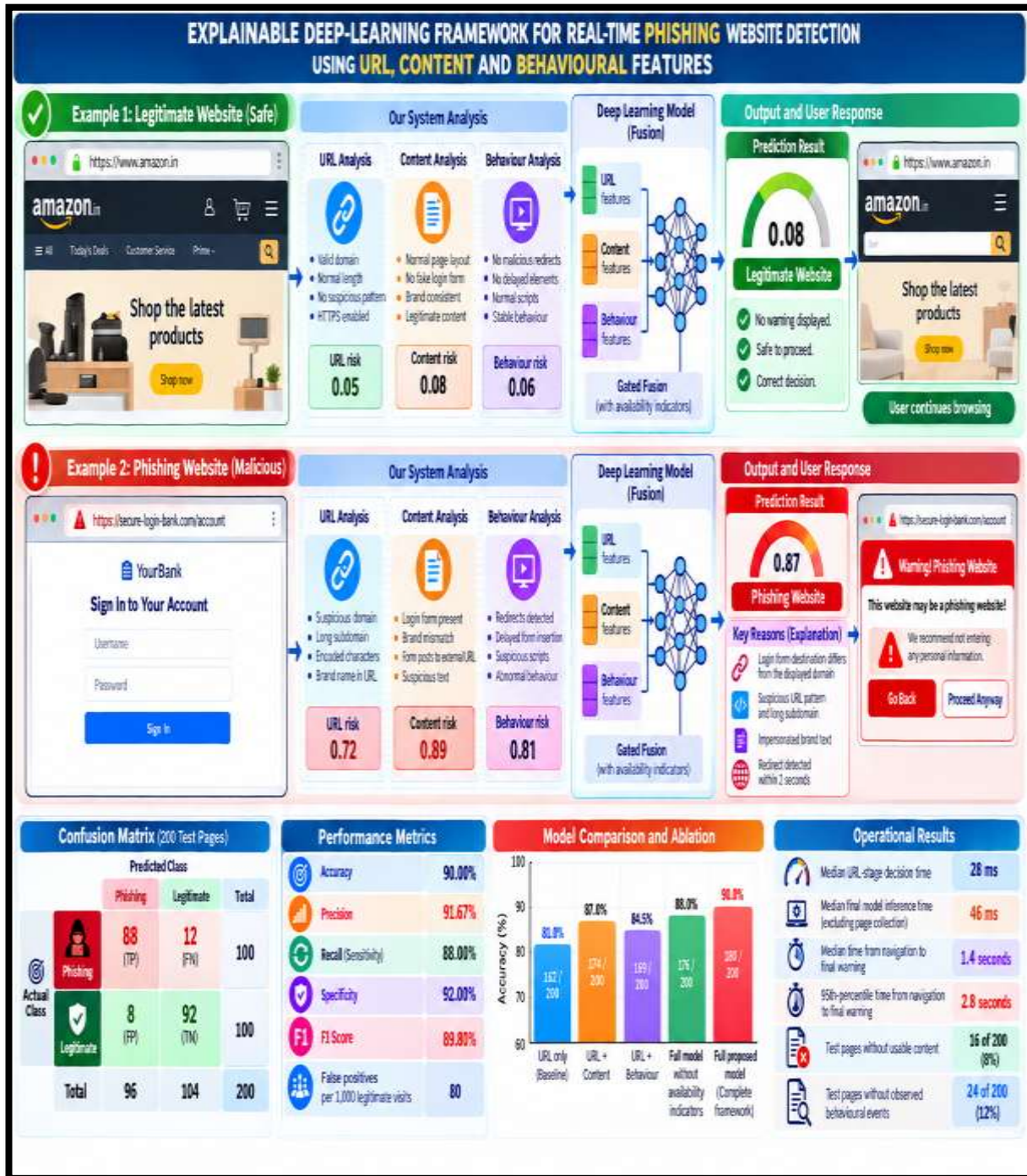
Measure	Value
Median URL-stage decision time	28 ms
Median final model inference time, excluding page collection	46 ms
Median time from navigation to final warning	1.4 seconds
95th-percentile time from navigation to final warning	2.8 seconds
Test pages without usable content	16 of 200
Test pages without observed behavioural events	24 of 200

9. OUTCOME

The proposed system's first output is a risk assessment generated at the earliest useful stage. Before a page opens, the URL encoder supplies an initial phishing probability. A sufficiently high score, judged against a threshold selected during validation, may trigger an immediate warning. If the initial decision is uncertain, the system collects permitted page evidence within a fixed time budget and updates the score. The resulting warning identifies which inputs were available and presents a small number of signals associated with the prediction. For example, it might identify a mismatch between the brand displayed on a login page and the destination of its form. This is an illustration of the intended output, not an observed experimental case. The usefulness of the warning must be assessed after the system is implemented.

The second intended output is a sanitised audit record that allows researchers to evaluate decisions. It contains the decision stage, available inputs, risk score, selected threshold, and time to warning. It can also record feature attributions and whether webpage inspection failed, without retaining credentials or submitted values. Researchers can review these records to understand false alarms and missed phishing pages. They can also test whether explanations change substantially when a webpage undergoes small, valid modifications. Aggregate records may reveal changes in latency or input patterns over time. Access to audit records should be restricted, and a retention period should be defined. Their actual value for monitoring must be established through testing.

FIG 3: OUTPUT and RESULT



10. LIMITATIONS

Disappearing and inaccessible pages: A phishing page may be removed before the researcher or detector can inspect it. Its URL may remain in a phishing feed even when its content and behaviour are no longer observable. Pages that can be inspected may therefore differ from those that cannot. The study must report collection failures and the number of samples available for each type of analysis.

Data-source and time bias: A model might learn to distinguish collection sources rather than phishing websites. Randomly assigning related URLs to training and testing can make performance appear stronger than it is on new domains. Popular websites may also be easier to classify than ordinary legitimate websites. Chronological, domain-

disjoint testing and a diverse legitimate sample can reduce these risks, although they cannot remove every source of bias.

Behavioural timing: A suspicious event might occur only after the observation window closes. Extending that window could capture more evidence, but it could also delay the warning. Some pages may behave differently inside an isolated browser used for safe collection. The paper must therefore report both the evidence captured and the time required to capture it.

Explanation limits: An attribution indicates which recorded inputs influenced a model's prediction; it does not establish malicious intent. Correlated features may allow several different explanations of a similar decision. A readable explanation may still fail to reflect the model's actual behaviour. Faithfulness and stability tests are needed before claiming that the explanations are reliable.

11. CONCLUSION

This paper proposes a staged framework for phishing website detection using URL, webpage-content, and bounded behavioural evidence. The URL stage is designed to provide an early risk estimate before the page is inspected. Additional encoders update the assessment as permitted evidence becomes available. An availability-aware fusion layer distinguishes missing inputs from observations that indicate low risk. The explanation layer links a warning to recorded inputs and states which evidence could not be inspected. Behavioural collection is limited to event types and timings rather than information entered by a user. These components form a system design that can be implemented and tested. They do not, on their own, establish improved detection performance.

The proposed evaluation compares the full framework with URL-only and other appropriate baselines. It uses chronological, domain-disjoint test data to examine whether performance transfers to future websites. False warnings on legitimate pages are measured alongside successful detection of phishing pages. The evaluation records when a warning reaches the user, including the time required to collect evidence. It also tests missing inputs and examines the faithfulness and stability of explanations. The final discussion should report the observed benefit and cost of each input type. It should identify a decision threshold supported by validation and independent testing. Only completed experiments can establish whether the framework is suitable for deployment.

12. FUTURE ENHANCEMENTS

1. **On-device privacy-preserving learning:** Study whether local updates can adapt the detector while limiting central collection of browsing data, with measured privacy and communication costs.
2. **Vision-language brand consistency checks:** Compare a page's visible branding with its domain using a carefully evaluated visual branch for cases that remain uncertain after the initial stages.
3. **Adversarial and cloaking resilience:** Test pages that change content in response to browser type, inspection timing, or automated collection.
4. **Continual drift monitoring:** Detect when new phishing campaigns change the input distribution and trigger a controlled model reassessment.
5. **Uncertainty-aware warning policies:** Calibrate when the system should warn, request further inspection, or defer to a security analyst, based on measured error costs.

References

1. Alsakarnah, R., Masoud, M. Z., & Ghababsheh, A. (2026). Hybridizing explainable AI (XAI) for intelligent feature extraction in phishing website detection. *Electronics*, 15(2), 350. <https://doi.org/10.3390/electronics15020350>
2. Çolhak, F., Ecevit, M. İ., Uçar, B. E., Creutzburg, R., & Dağ, H. (2024). *Phishing website detection through multi-model analysis of HTML content* [Preprint]. arXiv. <https://arxiv.org/abs/2401.04820>

3. Gopali, S., Namin, A. S., Abri, F., & Jones, K. S. (2024). *The performance of sequential deep learning models in detecting phishing websites using contextual features of URLs* [Preprint]. arXiv. <https://arxiv.org/abs/2404.09802>
4. Kehkashan, T., et al. (2025). Explainable phishing website detection for secure and sustainable cyber infrastructure. *Scientific Reports*, 15, 41751. <https://doi.org/10.1038/s41598-025-27984-w>
5. Lee, J., Lim, P., Hooi, B., & Divakaran, D. M. (2024). *Multimodal large language models for phishing webpage detection and identification* [Preprint]. arXiv. <https://arxiv.org/abs/2408.05941>
6. Le Pochat, V., Van Goethem, T., Tajalizadehkhoob, S., Korczyński, M., & Joosen, W. (2019). *Tranco: A research-oriented top sites ranking hardened against manipulation*. Network and Distributed System Security Symposium. <https://doi.org/10.14722/ndss.2019.23386>
7. Li, Y., Yang, Z., Chen, X., Yuan, H., & Liu, W. (2019). A stacking model using URL and HTML features for phishing webpage detection. *Future Generation Computer Systems*, 94, 27–39. <https://doi.org/10.1016/j.future.2018.11.004>
8. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30. <https://papers.nips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>
9. Mia, M., Derakhshan, D., & Pritom, M. M. A. (2025). *Can features for phishing URL detection be trusted across diverse datasets? A case study with explainable AI*. <https://arxiv.org/abs/2411.09813>
10. PhishTank. (n.d.). *Developer information*. https://phishtank.net/developer_info.php
11. Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *Proceedings of Machine Learning Research*, 70, 3319–3328. <https://proceedings.mlr.press/v70/sundararajan17a.html>
12. Vulfin, A., Sulavko, A., Vasiliev, V., Minko, A., Kirillova, A., & Samotuga, A. (2026). A multimodal phishing website detection system using explainable artificial intelligence technologies. *Machine Learning and Knowledge Extraction*, 8(1), 11. <https://doi.org/10.3390/make8010011>
13. Yoon, J.-H., Buu, S.-J., & Kim, H.-J. (2024). Phishing webpage detection via multi-modal integration of HTML DOM graphs and URL features based on graph convolutional and transformer networks. *Electronics*, 13(16), 3344. <https://doi.org/10.3390/electronics13163344>
14. Zieni, R., Massari, L., & Calzarossa, M. C. (2023). Phishing or not phishing? A survey on the detection of phishing websites. *IEEE Access*, 11, 18499–18519. <https://doi.org/10.1109/ACCESS.2023.3247135>
15. *An effective detection approach for phishing websites using URL and HTML features*. (2022). *Scientific Reports*. <https://doi.org/10.1038/s41598-022-10841-5>.