Privacy-Preserving and Explainable Phishing Website Detection Using Federated Continual Learning

Sonal Kotkar[1], Siddhi Pawar[2], Tanvi Khadse[3], Dhanshri Pansare[4], Prof. Jyoti Sarwade[5]
[1,2,3,4,5]*Department of Computer Science, MAEER's MIT Arts, Commerce and Science College, Pune, India*
doi.org/10.64643/ijirt.208519-459

Abstract*—Phishing websites are continuously modified to imitate legitimate services, evade static blacklists, and exploit user trust. Machine-learning-based detection can identify previously unseen patterns, but conventional centralized training requires sensitive URL, browsing, webpage, and behavioral information to be collected at a central location. This paper proposes FedXPhish, a privacy-preserving and explainable phishing website detection framework that unifies multimodal feature analysis, federated learning, continual learning, differential privacy, secure aggregation, robust aggregation, and explainable artificial intelligence (XAI). Each participating client extracts URL lexical, domain/TLS, HTML/text, visual, and behavioral features locally. A local multimodal classifier is trained without transferring raw data. Protected updates are securely aggregated to obtain a global model. Continual learning enables adaptation to emerging phishing campaigns while replay and proximal regularization reduce catastrophic forgetting. SHAP and LIME are used to provide global and instance-level explanations. The paper defines a time-aware, non-IID evaluation protocol covering detection performance, privacy, continual adaptation, communication efficiency, explanation quality, and robustness to malicious clients. The work is positioned as a conference-ready research framework; numerical results are deliberately not fabricated and should be inserted after implementation and experimentation.

Index Terms*—phishing website detection, federated learning, continual learning, explainable AI, differential privacy, SHAP, LIME, secure aggregation, multimodal learning.

I. INTRODUCTION

Phishing remains a major cybersecurity problem because attackers can rapidly create deceptive websites that mimic banking, e-commerce, social-media, cloud, and institutional services. Traditional blacklist and heuristic methods are useful for known threats but can be slow to react to newly registered domains and rapidly changing campaigns. Machine learning provides a way to learn discriminative patterns from URLs and webpage characteristics; however, a centralized pipeline can require organizations to transfer sensitive browsing or security data to a common server.

Federated learning (FL) offers a privacy-oriented alternative by allowing multiple clients to train a shared model while keeping raw training data locally. Nevertheless, FL is not equivalent to complete privacy: model updates can potentially leak information, and malicious clients can attempt model poisoning. In addition, phishing distributions are naturally heterogeneous across organizations, regions, languages, brands, and time periods. Continual learning is therefore important because a detector trained on historical phishing campaigns may degrade when attack patterns change.

Explainability is a second practical requirement. Security analysts need to understand why a URL is classified as malicious, especially when automated blocking can cause false positives. Recent studies have separately investigated federated learning, continual learning, SHAP/LIME-based explainability, and multimodal phishing detection. The proposed FedXPhish framework synthesizes these directions into one architecture and evaluates their interactions rather than treating privacy, adaptation, and explainability as independent features.

II. RELATED WORK AND RESEARCH GAP

Five recent studies provide the primary foundation for this work. Adeyemo et al. [1] integrate multimodal deep learning, federated learning, and explainable AI for email phishing detection, demonstrating that privacy-preserving learning can remain competitive with centralized training. Joshua et al. [2] combine federated and continual learning for web phishing and

report an attention/residual classifier with continual-learning strategies. Kehkashan et al. [3] use URL-based machine learning with SHAP for explainable phishing website detection. Shafin [4] proposes SLA-FS, using SHAP and aggregated LIME for explainable feature selection, with a reported accuracy of approximately 97.4% for Random Forest after feature reduction. Lim et al. [5] propose EXPLICATE, combining an ML classifier with LIME, SHAP, and an LLM layer for natural-language explanations, reporting approximately 98.4% accuracy on email phishing data. Broader surveys of federated approaches to phishing further motivate this direction [6].

The gap is not that any individual component is absent from the literature. Instead, the gap is the limited joint evaluation of privacy protection, multimodal website analysis, continual adaptation, explainability, and robustness under non-IID federated data. FedXPhish addresses this integration gap and explicitly defines privacy and explanation metrics alongside conventional detection metrics. Beyond integration, the framework's specific technical novelty lies in coupling the replay-and-proximal continual-learning objective (Eq. 5) directly with SHAP/LIME explanation-stability metrics, allowing explanation drift to be measured as phishing campaigns evolve — a joint privacy–adaptation–explainability evaluation that, to the authors' knowledge, no prior federated phishing detector has reported.

III. RESEARCH OBJECTIVES AND QUESTIONS

- Design a federated phishing website detector that does not centralize raw client data.
- Evaluate whether multimodal URL, domain, HTML/text, visual, and behavioral features improve generalization.
- Add differential privacy and secure aggregation to protect model updates.
- Use continual learning to adapt to time-separated phishing campaigns while reducing catastrophic forgetting.
- Generate global and local explanations using SHAP and LIME without exposing sensitive client information.
- Measure the trade-offs among detection utility, privacy, explainability, communication cost, and adversarial robustness.

RQ1: Can federated learning approach centralized detection performance under non-IID phishing data?
RQ2: What is the utility cost of different differential-privacy budgets?
RQ3: Does continual learning improve detection of newly emerging phishing campaigns?
RQ4: Are SHAP/LIME explanations stable and faithful under federated training?
RQ5: Which aggregation method is most robust to heterogeneous or malicious clients?
RQ6: Which feature groups contribute most strongly to phishing decisions?

IV. PROPOSED FEDXPHISH ARCHITECTURE

The complete architecture (Fig. 1) consists of a set of independent clients that perform local feature extraction and local multimodal training; gradients are clipped and perturbed; protected updates are sent through secure aggregation; the server performs robust aggregation; the updated global model is returned to clients; and an XAI service generates explanations from the resulting prediction. Raw URLs, webpage captures, browsing histories, credentials, and cookies remain local.

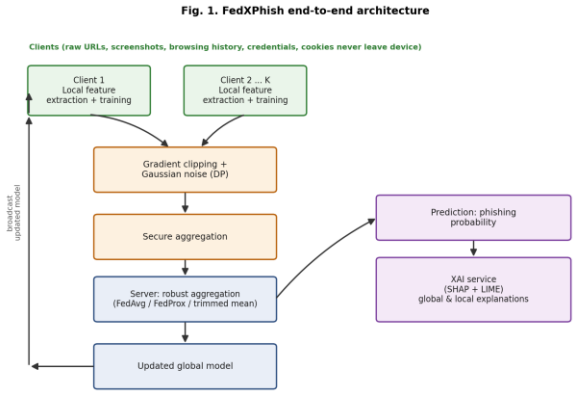

Fig. 1. FedXPhish end-to-end architecture.

TABLE I. CLIENT-SIDE MULTIMODAL FEATURE GROUPS

Feature Group	Representative Features
URL lexical	Length, entropy, subdomain count, special characters, IP-address usage, redirects, suspicious tokens.
Domain / TLS	Domain age, DNS properties, certificate validity and issuer.
HTML / text	Login forms, scripts, hidden elements, external-resource ratio, urgency language.

Feature Group	Representative Features
Visual	Screenshot embeddings, layout similarity, brand-impersonation cues.
Behavioral	Redirect chains, pop-ups, downloads, form-submission destinations.

The five client-side feature groups are summarized in Table I.

V. LOCAL MULTIMODAL DETECTION MODEL

A practical client model contains a multilayer perceptron for structured URL/domain features, a lightweight text or HTML encoder, and an optional compact vision encoder for screenshots. Their representations are concatenated or fused by an attention/fusion layer and passed to a binary classification head producing a phishing probability. The visual branch can be disabled for resource-limited clients. Quantization or knowledge distillation can further reduce client-side computation and communication (Fig. 2).

Fig. 2. Multimodal client-side feature extraction and fusion pipeline

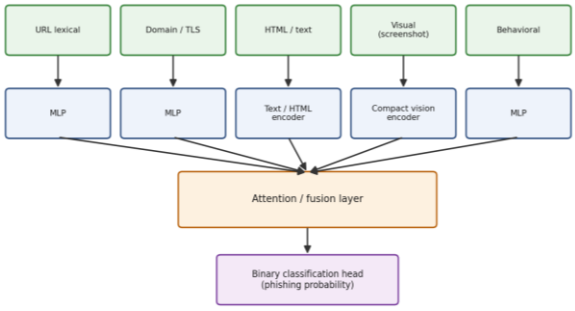

Fig. 2. Multimodal client-side feature extraction and fusion pipeline.

VI. FEDERATED OPTIMIZATION AND PRIVACY

Let K clients participate in a round. Client k owns D_k with n_k samples, and $N = \Sigma_k n_k$. The global objective is a weighted sum of local objectives, and with FedAvg the global model is the sample-weighted average of client updates, as in Eq. (1) and Eq. (2). Because phishing data are expected to be non-IID, FedAvg should be compared with FedProx and robust aggregation such as coordinate-wise median or trimmed mean.

$$F(w) = \Sigma_k (n_k / N) F_k(w) \quad (1)$$

$$w_{(t+1)} = \Sigma_k (n_k / N) w_{(t+1)}^k \quad (2)$$

For differential privacy, a client update g_k is clipped to norm C as in Eq. (3), after which Gaussian noise is added as in Eq. (4). The experiment must report the resulting privacy budget (ε, δ) rather than merely claiming that the system is private. Secure aggregation complements differential privacy by preventing the server from directly inspecting an individual client's update, as illustrated in Fig. 3.

$$g'_k = g_k / \max(1, ||g_k||_2 / C) \quad (3)$$

$$g''_k = g'_k + N(0, \sigma^2 C^2 I) \quad (4)$$

Fig. 3. Differential-privacy protected update and robust aggregation

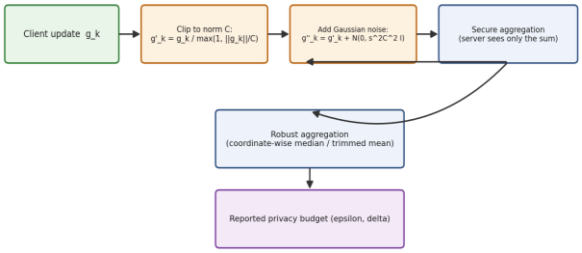

Fig. 3. Differential-privacy protected update, secure aggregation, and robust aggregation.

VII. CONTINUAL LEARNING

Phishing concepts change over time, so the local training process is organized into chronological tasks or windows (Fig. 4). A privacy-controlled replay buffer or latent-memory mechanism retains limited information from earlier tasks. The proposed local objective combines the current loss with a replay term and a proximal term, as shown in Eq. (5). The replay term reduces catastrophic forgetting, while the proximal term constrains excessive divergence from the current global model.

$$L_{total} = L_{current} + \lambda_{replay} L_{replay} + \lambda_{prox} ||w - w_t||^2 \quad (5)$$

Fig. 4. Continual-learning timeline with replay and proximal regularization

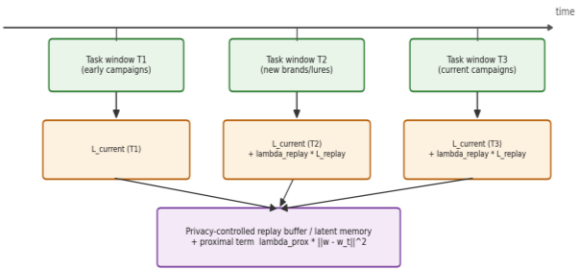

Fig. 4. Continual-learning timeline with replay buffer and proximal regularization.

VIII. EXPLAINABLE AI LAYER

SHAP is used for global feature importance and local feature attribution, while LIME provides a local surrogate explanation for an individual prediction. The explanation interface should show the phishing probability, decision label, confidence/uncertainty, top risk-increasing features, top risk-reducing features, and a concise human-readable summary (Fig. 5). Explanations must be generated from privacy-safe features and must not reveal another client's raw data. Feature attribution should be interpreted as model behavior, not as proof of causality.

Example explanation: High phishing risk because the domain is recently registered, the URL contains an encoded redirect, the TLS hostname is inconsistent with the destination, and the login form submits credentials to a different domain.

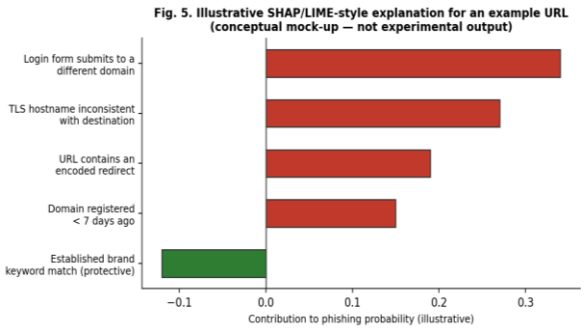

Fig. 5. Illustrative SHAP/LIME-style explanation for an example URL (conceptual mock-up, not experimental output; see Section IX).

IX. EXPERIMENTAL METHODOLOGY

The study should use verified phishing URLs and legitimate websites from reproducible public sources, together with domain metadata and, where legally and safely possible, archived webpage/screenshot information. To reduce leakage, duplicate URLs, related domains, near-identical templates, and campaign variants should be grouped before splitting. A chronological train/validation/test split is preferred over a purely random split because it better represents future phishing detection.

For federated simulation, data can be distributed across 10–50 clients. IID and non-IID conditions should both be evaluated. Non-IID partitions can vary class ratios, brands, languages, regions, and attack campaigns; a Dirichlet distribution can control heterogeneity. Each experiment should record the number of clients, client participation rate, local epochs, optimizer, learning rate, batch size, model size, privacy parameters, aggregation method, and communication rounds.

X. BASELINES AND ABLATION STUDY

TABLE II. BASELINE CONFIGURATIONS

Baseline	Role
Centralized multimodal model	Upper-bound reference.
Local-only model	No cross-client collaboration.
Standard FedAvg	Federated baseline without differential privacy.
DP-FL	FedAvg with gradient clipping and Gaussian noise.
Secure DP-FL	Differential privacy plus secure aggregation.
Fed-Continual	Federated continual learning.
FedXPhish	Complete proposed framework.

Ablation studies should remove visual features, remove domain metadata, remove continual learning, remove differential privacy, and remove robust aggregation, and should contrast SHAP-only versus LIME-only, FedAvg versus FedProx, and different privacy budgets.

XI. EVALUATION METRICS

TABLE III. EVALUATION DIMENSIONS AND METRICS

Dimension	Metrics
Detection	Precision, recall, F1-score, ROC-AUC, PR-AUC, false-positive rate.
Privacy	ε, δ, update leakage, membership-inference resistance.
Continual learning	Average accuracy, forgetting, forward transfer.
Explainability	Fidelity, stability, sparsity, analyst usefulness.
Federation	Communication bytes, rounds, convergence time.
Robustness	Poisoning degradation, client dropout, unreliable-client performance.
Fairness	Per-client and per-language performance variation.

XII. RESULTS REPORTING PLAN

Because this manuscript describes a proposed conference research framework rather than a completed experimental study, numerical results are intentionally not invented. The final experimental section should report mean ± standard deviation over multiple runs, confidence intervals where appropriate, and statistical comparisons between baselines. Results should include a main performance table, a privacy–utility curve for different ε values, a continual-learning curve across time windows, a communication-cost comparison, and SHAP/LIME explanation examples (see Fig. 5). This separation is essential for research integrity.

XIII. EXPECTED CONTRIBUTIONS

- A unified privacy-preserving phishing website detection architecture combining FL, DP, secure aggregation and robust aggregation.
- A multimodal client model integrating URL, domain/TLS, webpage/text, visual and behavioral evidence.
- Continual adaptation for evolving phishing campaigns under non-IID client distributions.
- Dual explainability using SHAP and LIME with privacy-aware explanation generation.
- A multi-dimensional evaluation protocol covering detection, privacy, adaptation, explainability, communication and robustness.

XIV. LIMITATIONS AND ETHICAL CONSIDERATIONS

Differential privacy can reduce utility, particularly for rare phishing classes. Screenshot processing can be expensive for low-resource clients. Domain-registration and certificate features may be unavailable or delayed. Federated learning does not by itself eliminate poisoning or backdoor attacks. XAI explanations may be unstable and should therefore be evaluated rather than assumed to be faithful.

Only minimum necessary data should be processed. Credentials, cookies, browsing histories and personally identifying URL parameters must not be collected for model training. Webpage rendering should occur in isolated sandboxes. Automated blocking should use calibrated thresholds and, for high-impact cases, analyst review.

XV. CONCLUSION

FedXPhish provides a conference-oriented research direction for trustworthy phishing website detection by treating privacy, adaptation, interpretability and robustness as joint requirements. The framework combines multimodal detection with federated continual learning, differential privacy, secure and robust aggregation, and SHAP/LIME explanations. The proposed evaluation is deliberately time-aware and non-IID so that performance is tested under conditions closer to real deployment. The key next step is implementation and empirical validation; only measured results should be inserted into the final conference submission.

REFERENCES

[1] A. Adeyemo, O. Emuoyibofarhe, A. Abandon, J. Adegboye, and S. Ajagbe, “Towards trustworthy email phishing detection: Integrating multi-modal deep learning, federated learning, and explainable AI,” in *Proc. IndabaX Nigeria 2026*, PMLR, vol. 319, pp. 101–117, 2026.

[2] J. J. M. Joshua, R. Adhithya, S. Dananjay, and M. Revathi, “Exploring the efficacy of federated-continual learning nodes with attention-based classifier for robust web phishing detection: An empirical investigation,” arXiv preprint arXiv:2405.03537, 2024.

[3] S. Shevkari, P. S. Bhosale, and G. A. Mule, “Machine learning in cybersecurity: Cutting-edge approaches to threat detection and protection,” International Journal of Research Publication and Reviews, vol. 6, no. 11, pp. 7090–7092, 2025.

[4] S. S. Shafin, “An explainable feature selection framework for web phishing detection with machine learning,” Data Science and Management, vol. 8, no. 2, pp. 127–136, 2025, doi: 10.1016/j.dsm.2024.08.004.

[5] B. Lim, R. Huerta, A. Sotelo, A. Quintela, and P. Kumar, “EXPLICATE: Enhancing phishing detection through explainable AI and LLM-powered interpretability,” arXiv preprint arXiv:2503.20796, 2025.

[6] E. Daoud, J. Garcia-Blas, S. Alawadi, et al., "Phishing in the age of distributed intelligence: Taxonomies, detection strategies, and the emerging role of federated learning," *Progress in Artificial Intelligence*, 2025.