E-ISSN: 3025-0501
Vol. 4, No. 3, May 2026
pp. x-xx
DOI: https://doi.org/10.57185/qkc5mc07

Almakki Health Informatics Journal

Explainable Artificial Intelligence for Phishing Detection and Its Implications for Health Information System Security: A Systematic Literature Review

Habibah Alya
Sekolah Tinggi Manajemen Informatika dan Komputer IKMI, Cirebon, Indonesia
Corresponding Author: habibahalya23@gmail.com

Keywords:
Explainable Artificial Intelligence; Phishing Detection; Health Information Security; SHAP; Systematic Literature Review

This is an open access article under the CC BY-SA license.

Copyright © 2026 by Author. Published by Al-Makki Publisher

ABSTRACT

Phishing attacks continue to grow in scale and sophistication, and health care is a vulnerable target because it holds high-value patient data and depends on available information systems. Deep learning and Transformer-based detectors are accurate but often opaque. This review aimed to synthesize the development of Explainable Artificial Intelligence (XAI) for phishing detection and interpret its implications for health information system security. Following a PRISMA 2020-guided protocol applied to Scopus, 365 records were retrieved, 71 full texts were assessed, and 48 peer-reviewed studies published from 2022 to 2026 were included by a single reviewer, alongside supporting cybersecurity XAI literature. Detection models evolved from traditional machine learning toward ensemble, Transformer, and large language model architectures, with ensemble learning the most frequently reported baseline, while SHAP and LIME dominated explainability. Explainability improved transparency, whereas accuracy gains came mainly from explainability-driven feature selection in a few studies rather than from XAI in general. However, no study integrated explainable feature selection, machine learning, ensemble learning, and deep learning with trustworthy cybersecurity principles in one framework, and none used health care data or settings. Real-time deployment, standardized evaluation of explanation quality, and human-centered validation remain open challenges. A conceptual framework and a research agenda with four directions (explainable AutoML, explainable foundation models, explainable federated learning, and human and AI collaborative security) are proposed toward explainable, adaptive, and privacy-preserving phishing detection. The health care implications are interpretive and require empirical validation in health facilities.

INTRODUCTION

Phishing remains one of the most persistent initial-access vectors in cybersecurity. Its scale and complexity continue to grow as attackers shift from single-channel email lures to multi-channel campaigns that exploit social media, cloud platforms, instant messaging, and generative AI to craft messages that are hard to distinguish from legitimate communication (Salloum et al., 2022). Website- and URL-based phishing has evolved through look-alike domains, obfuscated URLs, and evasion techniques that bypass conventional controls (Zieni et

al., 2023), and recent reviews report machine-learning-aware evasion, dynamic content, and identity spoofing that erode static detection rules (Li et al., 2024). Between 2024 and 2026, phishing further diversified into phishing-as-a-service, QR-code and voice-based attacks, and cloud-authentication abuse (Daoud et al., 2026). Beyond direct financial loss, organizations face incident-response costs, regulatory exposure, reputational harm, and follow-on compromises such as account takeover, data exfiltration, and ransomware (Abrardi et al., 2025; Althobaiti & Alsufyani, 2024; Naqvi et al., 2023), and digital financial fraud is increasingly traced to social-engineering-driven phishing (Puri et al., 2026). Quantitatively, the Anti-Phishing Working Group observed about 3.8 million phishing attacks in 2025 (Anti-Phishing Working Group, 2026), phishing was the initial access vector in 16% of breaches in the 2025 Verizon Data Breach Investigations Report (Verizon, 2025), and Indonesia's national cyber agency recorded about 26.8 million phishing activities among 330.5 million traffic anomalies in 2024 (Badan Siber dan Sandi Negara, 2025).

The health sector faces distinctive risks. Health care facilities store high-value personal health information and depend on information system availability for clinical care. An analysis of 374 ransomware attacks on US hospitals, clinics, and other health care delivery organizations showed that the annual number of attacks more than doubled from 2016 to 2021 and exposed the personal health information of nearly 42 million patients (Neprash et al., 2022). During the COVID-19 pandemic, phishing was among the most prominent attack methods against the health sector, driven by changed working practices, online consultation services, and pandemic-themed lures (He et al., 2021). Human factors also contribute: in a series of simulated phishing experiments over 18 months, about 4% of health care workers in each experiment submitted their credentials to a fraudulent site, although awareness training reduced interaction with phishing emails (Challacombe & McElhiney, 2025). In Indonesia, the mandatory implementation of electronic medical records requires confidentiality, integrity, and availability of data as well as controlled access rights for health workers (Kementerian Kesehatan Republik Indonesia, 2022). Indonesian health institutions have also been affected: the 2021 data breach at BPJS Kesehatan, the national health insurance administrator, exposed weak coordination and the absence of an integrated crisis response protocol among the responsible cyber institutions (Alevrido & Pasha, 2025), and an assessment of primary health care services in nine Indonesian provinces found that hardware was the least mature dimension and that information technology staff and infrastructure were limited (Aisyah et al., 2024). Staff accounts stolen through phishing can become an entry point for violating all three principles, which makes phishing detection part of health information system security governance.

Traditional blacklist- and rule-based defenses struggle against this evolving landscape. They depend on previously observed indicators, cannot recognize zero-day phishing pages before cataloguing, and lose value quickly as attackers modify URL structure, page content, and domain characteristics, while the short lifespan of phishing sites outpaces list-update latency (Abdillah et al., 2022; Salloum et al., 2022; Sánchez-Paniagua et al., 2022; Zieni et al., 2023). This has motivated a sustained shift toward machine learning (ML) and deep learning (DL). Early work relied on handcrafted URL, HTML, and domain features classified with Support Vector Machines, Random Forest, Decision Trees, and gradient-boosted trees (Salloum et al., 2022), complemented by metaheuristic feature selection such as particle swarm

optimization and extreme learning machines tuned with diversity-oriented social network search (Alshahrani et al., 2022; Bacanin et al., 2023). Subsequent research adopted deep neural networks, CNNs, LSTM networks, and hybrid architectures for automatic feature extraction, together with natural language processing, word embeddings, and attention mechanisms for analyzing email and web content (Catal et al., 2022; Kavya & Sumathi, 2025).

As these models grew more accurate, they also grew less transparent. High-performing deep learning and Transformer-based classifiers increasingly operate as black boxes, and several reviews caution that accuracy alone cannot justify deployment in security-critical settings where analysts must audit, validate, and trust automated decisions (Capuano et al., 2022; Ding et al., 2022). Explainable Artificial Intelligence (XAI) has therefore emerged as a central research direction, offering post-hoc and intrinsic mechanisms that expose each feature's contribution to a model's prediction (Charmet et al., 2022; Yan et al., 2022). In phishing detection, XAI is expected to build analyst trust while supporting incident investigation, feature engineering, and regulatory accountability (Fasla & Mehmood, 2026). The same demand is well established in health informatics: a systematic review of 99 articles showed that the black-box nature of AI models reduces trust and limits their use in health care practice, which is why techniques such as SHAP and LIME are widely used to explain model predictions (Loh et al., 2022). For health care facilities, explainable phishing detection would allow information technology staff, who are often few, to understand why a message was blocked and to explain it to clinical staff who are not technical experts.

Despite growing interest, the literature on explainable phishing detection remains fragmented in at least four respects. Prior systematic reviews on feature engineering and XAI for phishing website detection (Alsuqayh et al., 2025) and on interpretable machine learning techniques for phishing attack detection (Chandre et al., 2025) cover adjacent but narrower areas: the former centers on feature-engineering pipelines and the latter on interpretability techniques in isolation, and neither extends to Transformer-, LLM-, or federated-learning-based detectors or proposes an integrative framework.

Four gaps follow from this comparison. First, most studies still prioritize predictive accuracy, precision, recall, and F1-score over interpretability, partly because benchmark datasets make accuracy easy to report and compare (Abdillah et al., 2022; Salloum et al., 2022; Zieni et al., 2023). Second, XAI research has developed along separate tracks of explainable feature selection, machine learning, ensemble learning, and deep learning (Ahmed & EP, 2026; Kundu et al., 2026; Shafin, 2025; Yan et al., 2022). Third, as a preliminary observation to be tested by this review, no study identified in the initial literature scan appears to integrate Explainable Feature Selection (XFS), Explainable Machine Learning (XML), Explainable Ensemble Learning (XEL), Explainable Deep Learning (XDL), and Trustworthy Cybersecurity within a single framework; the cited contributions appear to combine at most two of these components (Ahmed & EP, 2026; Calzarossa et al., 2024; Kundu et al., 2026; Shafin, 2025). Fourth, practical deployment challenges, including computational overhead, explanation latency, the absence of standardized fidelity and faithfulness metrics, and human-in-the-loop integration, remain under-studied compared with model-centric interpretability research (Charmet et al., 2022; Ding et al., 2022; Zafar & Wu, 2026). In this review, XFS, XML, XEL, and XDL are used only as abbreviations for the explainable variants of feature selection,

machine learning, ensemble learning, and deep learning, and XML does not refer to the Extensible Markup Language.

Motivated by these gaps, this review synthesized how AI and XAI jointly evolved for phishing detection between 2022 and 2026, guided by five research questions: (RQ1) How have AI/ML models for phishing detection evolved over time? (RQ2) Which interpretability techniques are most widely adopted? (RQ3) How does explainability contribute to model performance and user trust? (RQ4) What implementation challenges hinder XAI deployment in operational environments? and (RQ5) What research gaps remain in explainable phishing detection? The review contributes a synthesis of model evolution from conventional ML to Transformer- and LLM-based architectures, a comparative analysis of explainability techniques, an evidence-based research-gap analysis, a conceptual framework linking model complexity, explainability, trustworthiness, and operational adoption, and an interpretation of the implications for information system security in health care facilities, which constitutes the novelty of this review. The benefits of this review are threefold. It gives researchers a consolidated map of model evolution, explainability techniques, and unresolved gaps. It gives information technology managers of health facilities a basis for choosing lightweight and explainable detection options and for setting realistic expectations about explanation quality. It gives regulators and policy makers a basis for reading the confidentiality, integrity, and availability requirements for electronic medical records together with current phishing detection practice.

METHOD

This study adopted a systematic literature review (SLR) because it provides a transparent and replicable process for identifying, appraising, and synthesizing scientific evidence on a well-defined research problem, in contrast to narrative reviews (Catal et al., 2022; Page et al., 2021). The protocol followed the three SLR phases of planning, conducting, and reporting, and study selection was guided by the PRISMA 2020 flow of identification, screening, eligibility, and inclusion (Page et al., 2021). The protocol was not prospectively registered in a public registry such as the Open Science Framework, which is acknowledged as a limitation.

Literature was retrieved from Scopus because of its broad indexing of peer-reviewed computer science and cybersecurity venues and its common use as the primary database in phishing detection reviews (Abdillah et al., 2022; Alsuqayh et al., 2025; Catal et al., 2022; Salloum et al., 2022). Scopus was the sole search database, and the counts at each selection stage in Figure 1 reflect this single-database protocol. Using one principal database also keeps every filter count reproducible, and evaluations of academic search systems show that only about half of the systems examined are suitable as principal sources for evidence synthesis (Gusenbauer & Haddaway, 2020). This choice nonetheless excludes records indexed only in IEEE Xplore, ACM Digital Library, or Web of Science, including some conference proceedings, as discussed in the limitations. The search combined three concept groups with the Boolean operator AND: an explainability group ("Explainable Artificial Intelligence" OR XAI OR "Explainable Machine Learning" OR "Interpretable Machine Learning"), a phishing group ("Phishing Detection" OR "Phishing Website" OR "Phishing Email" OR "Malicious URL"), and a modeling group ("Machine Learning" OR "Deep Learning" OR "Ensemble Learning"). For studies addressing generative-AI-driven phishing, the search string was

extended with ("Generative AI" OR "Large Language Model" OR LLM OR ChatGPT) (Sivaneswaran et al., 2026). The extension terms were combined with the three concept groups in the same query, so the 365 records represent the base and extended strings together; the number of records contributed only by the extension is [n, to be confirmed by the author]. Because every concept group was combined with AND, only studies that use explainability terms are retrieved, and Transformer or LLM detectors reported without such terms fall outside the retrieved set. The search was conducted in June 2026 and returned 365 records.

Studies were included if they were peer-reviewed journal articles in English, addressed AI-, ML-, or XAI-based phishing detection, reported an identifiable methodology and empirical results, and were published in 2022–2026. Studies were excluded if they were duplicates, not peer-reviewed (editorials, short papers, or theses), inaccessible in full text, purely secondary reviews when primary evidence was required, or did not report an empirical evaluation of a detection method (Abdillah et al., 2022; Catal et al., 2022; Salloum et al., 2022). The initial 365 records were reduced to 296 after publication-year filtering (2022–2026), 280 after subject-area filtering (Computer Science, Engineering), 236 after document-type filtering (article, conference paper), 235 after language filtering (English), and 71 after source-type filtering (journal). Conference papers retrieved by the document-type filter were removed by the source-type filter because this review is restricted to journal articles; the inclusion criterion was corrected accordingly. The 71 records were then screened by title and abstract, and all 71 were retained for full-text assessment, so this stage excluded no additional records. Full-text eligibility assessment of these 71 records excluded 23 (7 duplicates, 6 not peer-reviewed, 5 with no empirical evaluation, and 5 outside the scope), yielding 48 included studies. Duplicates and records that were not peer reviewed were identified only at the full-text stage because Scopus document-type and source-type labels do not guarantee a unique or fully peer-reviewed record; for example, early access and final versions of the same article can be indexed as separate entries. These studies were analyzed alongside a set of foundational XAI, cybersecurity XAI, and phishing detection surveys and reviews (secondary sources that are not counted among the 48 included studies) used to contextualize interpretability concepts and evaluation criteria. Title, abstract, and full-text screening were performed by the single author without independent second-rater verification or an inter-rater reliability statistic (e.g., Cohen's kappa); this limitation is discussed later. The selection flow is shown in Figure 1.

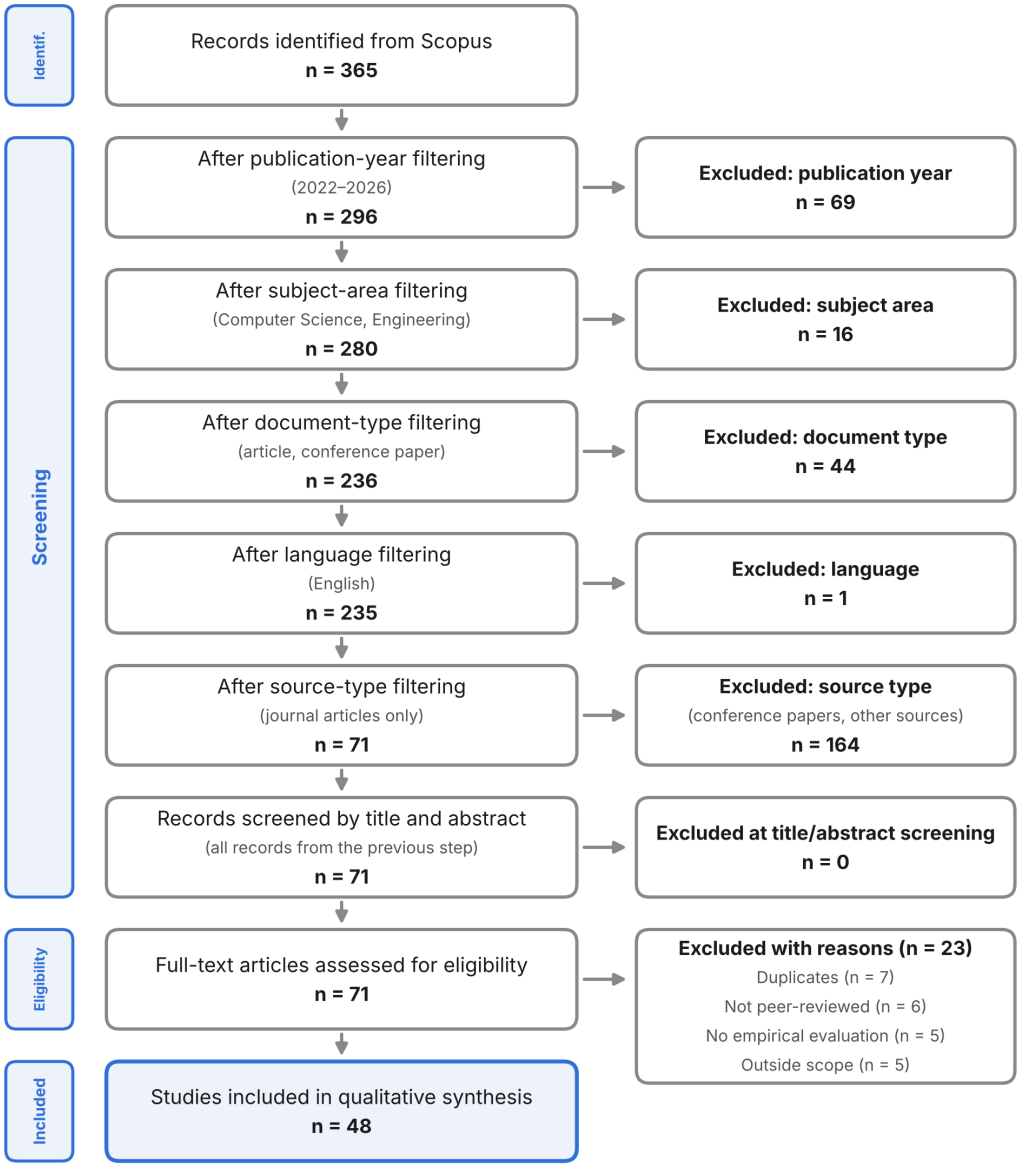


Figure 1. PRISMA 2020 flow diagram of the study selection process (revised to show the title and abstract screening stage)

Beyond the eligibility criteria, each included study was appraised against seven quality assessment (QA) criteria adapted from established SLR quality-appraisal guidance (Catal et al., 2022; Haque et al., 2023) and tailored to the explainability focus of this review (Zafar & Wu, 2026), as listed in Table 1. Each criterion was scored as Yes (1), Partial (0.5), or No (0), for a maximum score of 7. Studies scoring below 3.5 would have been flagged for cautious interpretation or excluded. QA scores ranged from 5.0 to 7.0, with a mean of 6.2, and no study fell below the threshold. QA7 is already satisfied by the eligibility criteria and therefore discriminates little between studies, so QA scores are driven mainly by QA1 to QA6. Dataset quality and evaluation design, such as cross dataset validation and class balance, were not

appraised separately and are noted as a limitation. The QA score of each included study is given in Table A1.

Table 1. Quality assessment criteria

Code	Assessment question
QA1	Are the study's aim and scope clearly stated?
QA2	Is the phishing dataset or data source (URL, website, email, or third-party features) adequately described?
QA3	Is the AI/ML/DL/Transformer/LLM detection method clearly specified and justified?
QA4	Is at least one XAI technique clearly described, including its scope (global/local) and rationale?
QA5	Are evaluation metrics and results (e.g., accuracy, precision, recall, F1-score, AUC) reported in sufficient quantitative detail?
QA6	Are the study's limitations or threats to validity explicitly discussed?
QA7	Is the study published through a peer-reviewed, traceable, and verifiable source (e.g., an indexed venue with a DOI)?

The variables extracted from each study were publication year, dataset and data source, AI/ML/DL algorithm, feature selection technique, XAI method, evaluation metrics (accuracy, precision, recall, F1-score, and area under the curve or AUC), reported performance, and key limitations. Data were synthesized narratively rather than through meta-analysis because datasets, algorithms, and evaluation metrics were highly heterogeneous. Findings were organized thematically into publication trend, evolution of model families, comparison of explainability techniques, research-gap analysis, and trustworthiness and implementation challenges, and were then integrated into a conceptual framework and research agenda, following established practice in framework-oriented XAI reviews (Yan et al., 2022; Zafar & Wu, 2026). To align with the scope of health informatics, each finding was also interpreted against the security needs of health care facility information systems, namely protecting the confidentiality, integrity, and availability of electronic medical records. The extraction matrix of the studies summarized in Table 3 is provided in Table A1.

RESULTS AND DISCUSSION

The distribution of studies by publication year is presented in Table 2. Publication activity was already substantial in 2022, reflecting an established ML/DL phishing detection base, softened in 2023–2024, and accelerated sharply from 2025 onward, with 2026 alone accounting for nearly 40% of the reviewed studies. Because the search was conducted in June 2026, 2026 covers only January to June, so its share of 39.6% understates the acceleration and the true annual rate is likely higher. Early studies (2022) concentrated on deep learning baselines and feature-engineering pipelines (Catal et al., 2022; Kavya & Sumathi, 2025). For context, systematic reviews published in 2023 and 2024, which are secondary sources and not among the 48 included studies, described accuracy as the dominant evaluation criterion (Althobaiti & Alsufyani, 2024; Safi & Singh, 2023; Zieni et al., 2023). From 2025 onward, research intensified around Transformer architectures, LLMs, ensemble learning, and, critically, explainability, with SHAP- and LIME-based components increasingly built into pipelines rather than added afterward (Ahmed & EP, 2026; Alsakarnah et al., 2026; Ghadami & Rahebi, 2025; Mayo et al., 2026; Shafin, 2025; Uddin et al., 2026; Yeboah et al., 2026). The steep rise in 2026 also coincides with emerging federated and privacy-preserving detection

(Almaktoom et al., 2026; Daoud et al., 2026), suggesting that explainability is increasingly co-designed with distributed, privacy-aware architectures.

Table 2. Distribution of selected studies by publication year (2022–2026)

Year	Number of studies	Percentage
2022	9	18.8%
2023	5	10.4%
2024	6	12.5%
2025	9	18.8%
2026	19	39.6%
Total	**48**	**100%**

Table 3 summarizes representative AI models for phishing detection, grouped into ensemble learning (including AutoML-based ensemble pipelines), deep learning, Transformer-based models, and LLMs, together with their dataset context, evaluation metrics, strengths, and limitations. The 14 studies in Table 3 were selected purposively to cover each model family and each explainability technique in Table 4, giving priority to studies with the most complete reporting of datasets and metrics; accuracy values come from different datasets and experimental settings and are not directly comparable across rows. Entries marked NR (Fatima et al., 2025; Rawat et al., 2026; Uddin et al., 2026) did not report numerical values in the sources available for extraction, so they are treated as qualitative evidence and are not used in any quantitative comparison. Beyond the tabulated models, hybrid feature-selection-integrated ensemble pipelines have also been reported as an effective, lower-complexity alternative to deep and Transformer-based detectors (Connolly & Atlam, 2025). No reviewed study proposed a single-algorithm, non-ensemble ML model in isolation; every study combined feature selection, ensembling, or deep/Transformer architectures with an explainability layer. This pattern is most likely a selection effect of the search string, which required explainability terms, and should not be read as a description of phishing detection research as a whole. Importantly for health informatics, the datasets in Table 3 consist of general website, URL, email, online social network, and IoT benchmarks; none originated from the email or information systems of health care facilities.

Table 3. Summary of AI models used in phishing detection

AI model	Category	Dataset	Evaluation metrics	Strengths	Limitations
Ensemble learning					
Random Forest and XGBoost + SLA-FS (SHAP+LIME feature selection) (Shafin, 2025)	Ensemble ML	Public phishing website benchmark	Accuracy: RF 97.41%; XGBoost 97.21%	Fast training; transparent feature ranking; strong baseline	Limited modeling of sequential/semantic patterns
LightGBM + Bat Algorithm feature selection (Almomani et al., 2026)	Gradient-boosted ensemble	ISCX-URL-2016; URL Phishing 2026	Accuracy 99.92% (ISCX-URL-2016); 98.17% (URL Phishing 2026)	Very high accuracy; >60% dimensionality reduction; efficient inference	Explainability added post hoc; optimizer-dependent performance

AI model	Category	Dataset	Evaluation metrics	Strengths	Limitations
Hybrid Explainability Random Forest (HXRF): SHAP+LIME+PDP+PFI (Alsakarnah et al., 2026)	Ensemble ML with multi-technique XAI	Large-scale phishing website dataset	Accuracy 98.2%	Compact, well-validated feature set; multi-view explanation	Heavier computation than single-XAI pipelines
Recursive Backward Elimination (RFE+BE) + ensemble classifier (Alharbi, 2026)	Feature selection + ensemble ML	Phishing URL dataset	Accuracy 90.7%	Reduced dimensionality; improved computational efficiency	Lower accuracy than deep/hybrid approaches
Explainable carbon-aware stacked ensemble (Rawat et al., 2026)	Ensemble ML (stacked)	Online social network phishing/impersonation dataset	NR (not reported in the extracted source)	Stacked design with integrated explainability; resource-aware	Performance and efficiency figures not detailed
Random Forest (best of SVM/RF/DT/LR/KNN) + SHAP feature selection (Kehkashan et al., 2025)	Ensemble ML	Kaggle Phishing Website Detection (11,000+ URLs, 30 features)	Accuracy 97%	Lightweight; SHAP feature ranking; suitable for constrained devices	URL features only; does not address deep/Transformer evasion
XGBoost (best of XGBoost/LightGBM/RF/KNN/Twin SVM/CNN) + LIME (Akhtar et al., 2025)	Ensemble ML	Four public phishing datasets	Accuracy 99.65%	Very high accuracy across datasets; broad benchmarking	No unified latency analysis; interpretability limited to LIME
Random Forest (best of RF/LR/DT) + LIME+SHAP (Fatima et al., 2025)	Ensemble ML	PhishTank	NR (best among compared models; exact values not reported)	IoT/robotic communication context; local and global explanation; interactive interface	Quantitative performance not detailed; generalization unverified
Whale Optimization Algorithm feature selection + H2O AutoML (AQUAPHISH) (Upreti et al., 2025)	AutoML ensemble	10,000 phishing/normal websites; 48 features reduced to 36	R^2 0.9534; RMSE 0.1079; accuracy 98.3%	Regression-based risk scoring; 23.6% faster training and about 18% lower inference latency; SHAP interpretability	Less comparable with classification baselines; explanation quality not an optimization objective
Deep learning					
GAN–CNN–GRU–FMRCNN–Transformer hybrid + BKA optimization (Ghadami & Rahebi, 2025)	Hybrid deep learning	Phishing website/URL dataset	Accuracy 98.67%	Multi-architecture feature learning; SHAP-explained outputs	High architectural complexity and training cost
Transformer-based models					
Ensemble of self-supervised pre-trained models (Yeboah et al., 2026)	Transformer-based ensemble	Multi-domain phishing/web attack dataset	Cross-domain accuracy 94.4%	Strong generalization; reduced labeled-data need	Complexity limits real-time interpretability

AI model	Category	Dataset	Evaluation metrics	Strengths	Limitations
Fine-tuned BERT and RoBERTa (Ibrahim & Elhafiz, 2026)	Transformer (NLP)	Nazario phishing corpus with Enron legitimate emails (19,000 emails)	Accuracy: RoBERTa 94%; BERT 91% (F1 0.96 and 0.94)	Rich contextual representation of email/URL text	Resource-intensive; needs an added XAI layer
Explainable few-shot learning (ModernBERT, XF-PhishBERT) (Tawfik et al., 2025)	Few-shot Transformer	Two public phishing URL datasets (88,647 and 235,795 URLs)	Accuracy 98.5% (1 shot) to 99.9% (10 shots)	Effective for zero-day phishing with scarce labels	Few-shot generalization still maturing
Large language models					
Explainable Transformer/LLM with LITA (Uddin et al., 2026)	Transformer/LLM	Phishing email dataset	NR (qualitative: high accuracy)	Deep semantic context; integrated LIME–Transformer attribution	High inference cost; limited real-time suitability

Table 3 shows that no single model family dominates; model choice reflects a trade-off between predictive strength, computational economy, and explainability. Ensemble methods such as Random Forest, XGBoost, and LightGBM remain the most frequently reported baseline because they combine strong performance on structured URL, HTML, and domain features with low training cost and naturally interpretable feature importance, which explains their pairing with SHAP, LIME, or hybrid feature-selection pipelines rather than their displacement (Alharbi, 2026; Almomani et al., 2026; Alsakarnah et al., 2026; Shafin, 2025). Deep learning and hybrid architectures achieve comparable or higher accuracy by learning representations directly from raw or sequential input, but with higher training complexity and a weaker link to human-interpretable concepts, which is why the reviewed deep models depend almost entirely on post-hoc explanation layers (Ghadami & Rahebi, 2025).

Transformer- and LLM-based classifiers extend this trade-off: their attention-based representations of email and URL text yield stronger cross-domain generalization and better performance in low-label, emerging-attack settings, yet the same scale raises inference cost and constrains real-time explanation (Ibrahim & Elhafiz, 2026; Tawfik et al., 2025; Uddin et al., 2026; Yeboah et al., 2026).

In the studies in Table 3, ensemble learning appears to offer a favorable balance of accuracy, training cost, and interpretability for operational deployment, while Transformer/LLM models suit settings that can absorb higher cost for deeper semantic understanding. This observation is tentative because latency and computational cost were not among the extracted variables and only one study in Table 3 reports inference latency (Upreti et al., 2025). Recent studies favor hybrid pipelines that reserve lightweight ensembles for real-time screening and Transformer/LLM components for deeper, asynchronous analysis (Ghadami & Rahebi, 2025; Sivaneswaran et al., 2026; Uddin et al., 2026).

Table 4 compares the most frequently reported explainability techniques by explanation type, scope, advantages, limitations, computational cost, and typical applications.

Table 4. Comparison of explainability techniques

Technique	Explanation type	Global/local	Advantages	Limitations	Cost	Typical applications
SHAP (Fasla & Mehmood, 2026; Mayo et al., 2026; Shafin, 2025)	Feature attribution (game-theoretic)	Global and local	Theoretically grounded, consistent, model-agnostic, stable	Expensive for large feature sets and deep models	High	Feature ranking; explainable feature selection; auditing
LIME (Ahmed & EP, 2026; Mayo et al., 2026)	Local surrogate modeling	Local	Intuitive; model-agnostic; easy for non-technical users	Sensitive to sampling; less stable than SHAP	Moderate	Per-instance explanation of URL/email classification
Hybrid SHAP–LIME (SLA-FS, HXRF) (Alsakarnah et al., 2026; Shafin, 2025)	Combined global and local attribution	Global and local	Complementary; improves feature selection and accuracy	Higher computational overhead	High	Explainable feature selection frameworks
Partial Dependence Plot (PDP) (Alsakarnah et al., 2026)	Marginal effect visualization	Global	Simple visualization of feature–outcome relationships	Assumes feature independence	Low	Supporting feature selection pipelines
Permutation Feature Importance (PFI) (Alsakarnah et al., 2026)	Performance-drop attribution	Global	Model-agnostic; simple	Biased by correlated features	Low–moderate	Feature ranking alongside SHAP/LIME
LITA (LIME–Transformer Attribution) (Uddin et al., 2026)	Local and attention-based attribution	Local	Consistent explanation for Transformer/LLM classifiers	Limited to Transformers; under-evaluated	High	Explainable Transformer-based classification
Attention visualization	Model-intrinsic weighting	Local	Native to attention models; low added cost	Weights do not always reflect true feature importance	Low	Interpreting attention-based deep models

Table 4 shows an explainability landscape organized around trade-offs between theoretical rigor and computational efficiency and between global and local scope. SHAP's grounding in cooperative game theory provides consistent, theoretically justified attributions across a model, which is why it dominates feature-selection pipelines that require a stable, dataset-wide ranking (Fasla & Mehmood, 2026; Mayo et al., 2026; Shafin, 2025); the same property makes it more expensive as feature dimensionality and model depth increase. LIME, by contrast, is lightweight and produces intuitive per-instance explanations that non-technical stakeholders can read more easily, but its reliance on local surrogate sampling makes it more sensitive to perturbation choices and less stable across runs (Ahmed & EP, 2026; Mayo et al., 2026). This contrast is also reported in methodological work on explanation stability, which shows that LIME explanations can vary across repeated runs because of random sampling and

can be stabilized with statistical stability indices or modified sampling (Visani et al., 2022; Zhou et al., 2021), whereas SHAP values for tree ensembles are computed as exact and consistent attributions (Lundberg et al., 2020). The studies summarized in Table 4 did not report quantitative stability measures, so these characterizations rest on the methodological literature rather than on measurements in the phishing studies. This complementarity, SHAP for consistent global auditing and LIME for fast case-by-case justification, explains why the best-performing frameworks such as SLA-FS and HXRF combine both (Alsakarnah et al., 2026; Shafin, 2025), a pattern that recurs across recent website, IoT, and ML-driven applications (Akhtar et al., 2025; Fatima et al., 2025; Kehkashan et al., 2025).

The local/global distinction mirrors two operational needs. Global methods such as SHAP, PDP, and PFI support model validation, feature engineering, and regulatory auditing, whereas local methods such as LIME and attention visualization support case-specific decisions during incident triage. Because analysts need both, neither scope suffices alone, which is the gap that hybrid SHAP–LIME pipelines and Transformer-specific methods such as LITA attempt to close at a non-trivial computational premium (Alsakarnah et al., 2026; Uddin et al., 2026).

Table 5 synthesizes the research gaps identified from the reviewed literature, linking existing studies and their limitations to recommended research directions.

Table 5. Research gap analysis

Existing studies	Current limitation	Research gap	Recommended direction
Accuracy-oriented ML/DL phishing detection (Abdillah et al., 2022; Catal et al., 2022; Li et al., 2024; Salloum et al., 2022)	Focus on accuracy/precision/recall; interpretability secondary	Explainability rarely embedded from the design stage	Explainability-by-design pipelines evaluated jointly on accuracy and interpretability
Single-technique XAI studies (Ahmed & EP, 2026; Fasla & Mehmood, 2026; Mayo et al., 2026)	Each technique evaluated independently; no systematic comparison	No standardized metrics (faithfulness, fidelity, stability) for XAI quality	Standardized, quantitative XAI evaluation protocols
Partial integration of XFS, XML, XEL, XDL, and Trustworthy Cybersecurity (Ahmed & EP, 2026; Alsakarnah et al., 2026; Calzarossa et al., 2024; Kundu et al., 2026; Shafin, 2025)	Existing frameworks combine at most two of the five components	No unified XFS–XML–XEL–XDL–Trustworthy Cybersecurity framework	Build and validate an integrative framework
Real-time/operational XAI studies (Fasla & Mehmood, 2026; Zafar & Wu, 2026)	High computational cost and latency of post-hoc explanations	Real-time, resource-constrained explainability unresolved	Lightweight/approximate XAI and edge-optimized pipelines
Federated/privacy-preserving detection (Almaktoom et al., 2026; Daoud et al., 2026)	XAI integration with federated learning still nascent	Cross-node explanation consistency under privacy constraints unaddressed	Explainable federated learning with consistency guarantees
LLM/Transformer-based detectors (Ibrahim & Elhafiz, 2026; Sivaneswaran et al., 2026; Uddin et al., 2026)	High accuracy but limited human-centered evaluation of explanations	Usefulness of explanations for analysts rarely measured	Human-centered evaluation with security analysts and end users

Existing studies	Current limitation	Research gap	Recommended direction
All studies in Table 3 (synthesis of this review)	General benchmark datasets; no health care facility data	Performance and usefulness of explanations in health care settings unknown	Validation on health care facility email and information systems involving health workers

The gap analysis in Table 5 confirms the four fragmentation points introduced earlier: an accuracy-first orientation (Abdillah et al., 2022; Catal et al., 2022; Li et al., 2024; Salloum et al., 2022), single-technique XAI evaluation without standardized metrics (Ahmed & EP, 2026; Fasla & Mehmood, 2026; Mayo et al., 2026), the absence of a unified framework (Alsakarnah et al., 2026; Calzarossa et al., 2024; Kundu et al., 2026; Shafin, 2025), and unresolved real-time, federated, and human-centered deployment (Almaktoom et al., 2026; Daoud et al., 2026; Fasla & Mehmood, 2026; Zafar & Wu, 2026).

These gaps reinforce one another: the lack of standardized explanation-quality metrics makes it hard to justify the cost of real-time XAI, discouraging its integration into federated and operational pipelines. They also persist for structural reasons, as publication and benchmarking incentives reward incremental accuracy gains on static datasets, whereas explanation quality, real-time cost, and human usability are harder to quantify, standardize, and compare (Ahmed & EP, 2026; Mayo et al., 2026; Zafar & Wu, 2026). The last row of Table 5 adds a gap specific to health informatics: because no study used data or users from health care facilities, the effectiveness of detection and the usefulness of explanations in that setting cannot yet be inferred from the existing literature.

Beyond model- and technique-level comparisons, the reviewed literature frames explainability as a bridge between technical performance and operational trust, because accuracy alone has proven insufficient to secure analyst and end-user confidence when models operate as black boxes (Capuano et al., 2022; Charmet et al., 2022). Embedding SHAP- and LIME-based explanations lets security teams verify that a URL, email, or website was flagged for relevant reasons, supporting incident investigation, error diagnosis, and consistency checks as attacks evolve (Fasla & Mehmood, 2026; Mayo et al., 2026). Real-time translation remains difficult, however. SHAP and LIME incur non-trivial computational cost, and generating explanations alongside high-throughput classification can introduce latency incompatible with real-time filtering, particularly for deep learning and Transformer models with large parameter counts (Fasla & Mehmood, 2026; Zafar & Wu, 2026), and explanations must be refreshed as attack characteristics shift (Daoud et al., 2026). Few studies evaluate explanation quality directly through faithfulness, fidelity, stability, or robustness; most treat accuracy as a proxy, leaving XAI trustworthiness under-validated (Zafar & Wu, 2026). Integration into security operations center (SOC) workflows, real-time monitoring, and human review remains emerging rather than established (Yan et al., 2022).

Traditional black-box deployment lets operators accept a verdict based on aggregate accuracy alone and treats interpretability as an add-on, whereas an XAI-enabled pipeline shifts the unit of trust from the model as a whole to each decision. This shift is increasingly necessary for security operations, which must justify why a message was blocked, satisfy audit and regulatory requirements, and diagnose the false positives that erode analyst trust and disrupt legitimate communication (Capuano et al., 2022; Charmet et al., 2022; Fasla & Mehmood, 2026). XAI does not remove the accuracy–complexity tension in Table 3, but it converts an

opaque decision into an auditable artifact that analysts can inspect, contest, and learn from, and the reviewed literature increasingly treats explainability as a design requirement for trustworthy cybersecurity systems rather than a reporting feature (Kundu et al., 2026; Yan et al., 2022).

Synthesizing these findings, Figure 2 proposes a conceptual framework linking AI model evolution, explainability techniques, implementation challenges, and research gaps to a research agenda for explainable phishing detection. The framework follows an evolutionary logic consistent with the reviewed literature: increasing model complexity (from traditional ML to Transformer, LLM, and federated architectures) raises the need for explainability techniques (SHAP, LIME, and their hybrids), which are in turn constrained by implementation challenges, namely computational cost, real-time latency, and the absence of standardized explanation-quality metrics. These challenges define the research gaps in Table 5 and point toward four converging directions. Table 6 traces each element of Figure 2 to the reviewed studies, so the framework follows from the extracted evidence and the gaps in Table 5 rather than from expert opinion alone. Elements that had no support in the reviewed studies (counterfactual explanation and adversarial robustness) were removed from the figure.

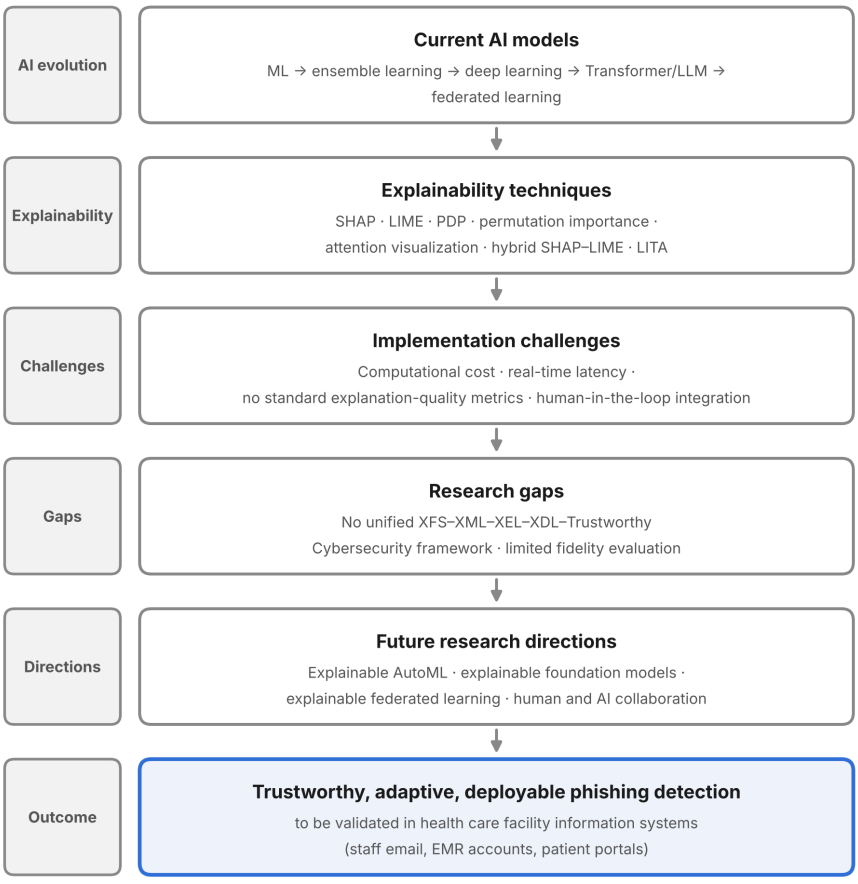


Figure 2. Future research framework for explainable phishing detection

Table 6. Traceability of the framework elements in Figure 2 to the reviewed evidence

Framework element	Evidence in the reviewed literature	Supporting studies
AI model evolution	Models grew more accurate and less transparent; ensembles remain the common baseline; Transformer and LLM classifiers generalize better at higher cost; federated detection is emerging	(Almaktoom et al., 2026; Daoud et al., 2026; Ghadami & Rahebi, 2025; Salloum et al., 2022; Shafin, 2025; Uddin et al., 2026; Yeboah et al., 2026)
Explainability techniques	SHAP, LIME, hybrid SHAP and LIME, PDP, PFI, LITA, and attention visualization (Table 4)	(Ahmed & EP, 2026; Alsakarnah et al., 2026; Mayo et al., 2026; Shafin, 2025; Uddin et al., 2026)
Implementation challenges	Computational cost and latency, no standardized explanation quality metrics, limited integration into security operations	(Charmet et al., 2022; Fasla & Mehmood, 2026; Upreti et al., 2025; Yan et al., 2022; Zafar & Wu, 2026)
Research gaps	Five gaps in Table 5, including the absence of a unified framework and of health care data	(Almaktoom et al., 2026; Calzarossa et al., 2024; Daoud et al., 2026; Kundu et al., 2026; Shafin, 2025)
Explainable AutoML	Ad hoc combination of feature selection, ensembling, and XAI; early metaheuristic AutoML evidence	(Connolly & Atlam, 2025; Shafin, 2025; Upreti et al., 2025)
Explainable foundation models	Transformer and LLM detectors generalize to emerging variants but rely on post hoc explanation	(Tawfik et al., 2025; Uddin et al., 2026)
Explainable federated learning	Privacy benefits of distributed training versus consistent explanations across nodes	(Almaktoom et al., 2026; Daoud et al., 2026)
Human and AI collaborative security	Explanation quality is assessed almost only technically; human centered evaluation is missing	(Yan et al., 2022; Zafar & Wu, 2026)
Sustainable or Green AI	Carbon aware ensemble designs pursue efficiency together with explainability	(Rawat et al., 2026)
Trustworthy Cybersecurity	Accuracy, explainability, robustness, privacy, and efficiency as joint conditions	(Calzarossa et al., 2024; Kundu et al., 2026)

The first direction is Explainable AutoML, which automates feature selection, model selection, and hyperparameter tuning while preserving auditable explanations. It can systematize the currently ad hoc combination of feature selection, ensembling, and XAI seen in Table 3 and Table 5 by searching for pipelines that satisfy accuracy, latency, and explanation-quality constraints together (Connolly & Atlam, 2025; Shafin, 2025); early evidence of metaheuristic-guided AutoML for phishing classification supports its feasibility, although explanation quality was not yet an optimization objective (Upreti et al., 2025). The second direction is explainable foundation models. Transformer- and LLM-based detectors already generalize well to emerging, sparsely labeled phishing variants, so embedding interpretability into pretraining or fine-tuning, rather than adding post-hoc methods such as LITA afterward, could reduce the inference cost that limits real-time use (Tawfik et al., 2025; Uddin et al., 2026).

The third direction is explainable federated learning, which addresses the tension between the privacy benefits of distributed training and the difficulty of producing consistent explanations across nodes that never share raw data, a challenge only beginning to be studied for phishing (Almaktoom et al., 2026; Daoud et al., 2026). The fourth direction is human–AI collaborative security. Explanation quality is still assessed almost exclusively from a technical standpoint, so involving analysts and, in health care, clinical staff as end users in evaluating and refining explanations would close the human-centered evaluation gap (Yan et al., 2022; Zafar & Wu, 2026).

Sustainable or Green AI is an emerging complement: as pipelines combine ensembling, deep architectures, and explanation generation, their cumulative energy footprint becomes a legitimate deployment constraint, and carbon-aware ensemble designs suggest that efficiency can be pursued together with explainability (Rawat et al., 2026). Finally, Trustworthy Cybersecurity functions as the unifying objective, treating accuracy, explainability, robustness, privacy, and efficiency as jointly necessary conditions and supplying the integrative criterion that the fragmented XFS–XML–XEL–XDL literature lacks (Calzarossa et al., 2024; Kundu et al., 2026).

Implications for Health Information System Security

This subsection collects the health care interpretations of the review in one place and keeps them distinct from the evidence in the preceding sections. None of the included studies used health care data, users, or settings, so the points below are interpretations drawn from cybersecurity and health informatics literature, not findings of the review, and they require empirical validation.

Article 29 of the Indonesian regulation on medical records requires electronic medical records to meet the principles of confidentiality, integrity, and availability. Confidentiality is the guarantee that data are protected from interference by internal or external parties without access rights, integrity is the accuracy of the data and information with changes allowed only for people granted editing rights, and availability is the guarantee that data can be accessed and used by people with access rights set by the head of the health facility (Kementerian Kesehatan Republik Indonesia, 2022). Phishing that captures staff credentials threatens all three principles, so detection with auditable explanations supports compliance: global explanations can help information security managers audit these requirements, whereas simple local explanations can be shown to physicians, nurses, or administrative staff when an email is quarantined, which also provides learning material for phishing awareness training that has been shown to be effective (Challacombe & McElhiney, 2025).

As an inference rather than a review finding, lightweight ensembles with SHAP explanations are a more realistic option for deployment at an email gateway or an electronic medical record access portal than LLMs that require GPUs and high inference costs. This inference is consistent with an assessment of primary health care services in nine Indonesian provinces, where hardware was the least mature dimension, some health centers had only two to five computers, and information technology professionals were scarce (Aisyah et al., 2024). Federated learning is also relevant because data across facilities generally cannot be centralized for confidentiality reasons, and hospitals and primary health centers could in principle train detection models jointly without moving emails or access logs that contain patient data (Almaktoom et al., 2026; Daoud et al., 2026).

False positives carry additional consequences in health facilities because a quarantined email may contain laboratory results, referrals, or care coordination, while false negatives can open the way to account takeover and ransomware attacks that disrupt care (Neprash et al., 2022). Decision thresholds and the way explanations are delivered should therefore be adapted to clinical workflows rather than optimized for accuracy alone. As a practical recommendation that remains to be tested, health facilities could quarantine suspicious messages with a notification to the recipient and to the information security team, together with a short local

explanation and a rapid release procedure, rather than delete them, and could apply stricter thresholds to messages that carry attachments or credential requests than to routine administrative mail.

These recommendations were developed partly with generative AI assistance, as disclosed in the Generative AI Statement, and they must be validated empirically in health facilities before they are adopted.

This review has several limitations. The XFS–XML–XEL–XDL–Trustworthy Cybersecurity framework (Figure 2) is a conceptual synthesis that has not been implemented or validated on real phishing datasets, so its feasibility, computational cost, and effect on detection or explanation quality remain to be established empirically. The framework in Figure 2 and the health care implications were partly developed with generative AI assistance, as disclosed in the Generative AI Statement, and both require empirical validation. The synthesis reflects literature indexed in Scopus within the search window; relying on a single database may have omitted studies indexed only in Web of Science or IEEE Xplore. IEEE Xplore, ACM Digital Library, and Web of Science also index much of the phishing detection literature, including major conference proceedings, so adding at least one more database is a priority for future updates of this review. Because the search string required explainability terms, the sample is biased toward pipelines that include XAI and may underrepresent Transformer and LLM detectors reported without such terms. Study selection and quality appraisal were conducted by a single reviewer without second-rater verification, which may introduce some subjectivity despite the explicit, criterion-based process. QA7 added little discrimination, dataset quality and evaluation design were not appraised, and the protocol was not registered, so the strength of the evidence should be read with caution. The rapid pace of LLM-based and federated phishing detection means that newer contributions may not be captured. Heterogeneity in datasets, features, and evaluation metrics limited direct quantitative comparison, and restriction to English-language sources may have introduced language and publication bias. Finally, the implications for health care facilities discussed here are interpretations based on cybersecurity and health informatics literature, not results of testing in health care settings.

CONCLUSION

This review synthesized 48 phishing-focused studies published between 2022 and 2026, alongside supporting cybersecurity XAI literature. For RQ1, phishing detection models evolved from traditional machine learning and early deep learning toward ensemble, Transformer, LLM, and federated architectures, with each generation trading some interpretability and computational economy for higher accuracy and stronger generalization. For RQ2, SHAP and LIME were the dominant techniques, individually and in hybrid configurations such as SLA-FS and HXRF. For RQ3, explainability was not merely a reporting layer: explainability-driven feature selection improved both transparency and accuracy in two studies, one using a public phishing website benchmark (Shafin, 2025) and one using a large-scale phishing website dataset (Alsakarnah et al., 2026), and SHAP/LIME outputs gave analysts a verifiable basis for trusting, auditing, and acting on model decisions. For RQ4, real-time XAI remains constrained by computational cost and latency, and explanation quality is rarely assessed with faithfulness, fidelity, or stability metrics. For RQ5, the main gap is that no

study identified in this review combined explainable feature selection, ensemble learning, deep learning, and trustworthy cybersecurity principles in a single framework, and the most promising directions are Explainable AutoML, explainable foundation models, explainable federated learning, human–AI collaborative security, Sustainable/Green AI, and Trustworthy Cybersecurity. The strength of this evidence is limited by the use of a single database, a single reviewer, and heterogeneous benchmarks that are not directly comparable, so the conclusions are best read as a synthesis of trends rather than definitive effects.

For health informatics, these findings suggest that explainable phishing detection may potentially contribute to protecting the confidentiality, integrity, and availability of electronic medical records, a benefit that remains to be tested because no study used health care settings. Lightweight ensembles with SHAP/LIME explanations appear to be the most realistic option for resource-constrained facilities, as an inference rather than a finding, local explanations can support clinical staff and awareness training, and explainable federated learning could enable collaboration across facilities without moving sensitive data. Because no study used data or users from health care facilities, future research should validate explainable phishing detection directly in health care settings, develop standardized explanation-quality evaluation protocols, and examine how security analysts and health workers use and trust the explanations produced.

FUNDING

The author declared that financial support was not received for this work and/or its publication.

CONFLICT OF INTEREST

The author declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

Habibah Alya: Conceptualization, Methodology, Investigation (literature search, study screening, and quality assessment), Data curation, Formal analysis, Visualization, Writing original draft, Writing review and editing. As the sole author, Habibah Alya has read and approved the final manuscript.

GENERATIVE AI STATEMENT

The author declared that generative AI was used in the creation of this manuscript. Claude (Anthropic) was used to restructure the manuscript according to the journal template, add discussion on health information system security, redraw the diagrams, verify and complete reference metadata, and format the citations. This application was conducted in adherence to established ethical guidelines regarding the use and disclosure of generative AI in scholarly research. The author has provided substantial intellectual contributions to the work, which has been rigorously vetted for accuracy, and assumes full responsibility for the integrity and final content of the manuscript. All statements added during the revision and all reference details, including authors, volumes, article numbers, and DOIs, were checked against the original sources.

APPENDIX

Table A1. Extraction matrix of the studies summarized in Table 3 (n = 14)

Study	Dataset	Model	XAI method	Reported performance	Main limitation	QA
(Shafin, 2025)	Public phishing website benchmark	Random Forest and XGBoost with SLA-FS	SHAP and LIME (hybrid)	Accuracy: RF 97.41%; XGBoost 97.21%	Limited modeling of sequential or semantic patterns	TBC
(Almomani et al., 2026)	ISCX-URL-2016; URL Phishing 2026	LightGBM with Bat Algorithm feature selection	Post hoc	Accuracy 99.92% and 98.17%	Explainability added post hoc; optimizer dependent	TBC
(Alsakarnah et al., 2026)	Large-scale phishing website dataset	HXRF (Random Forest)	SHAP, LIME, PDP, PFI	Accuracy 98.2%	Heavier computation than single XAI pipelines	TBC
(Alharbi, 2026)	Phishing URL dataset	RFE with backward elimination and ensemble classifier	NR	Accuracy 90.7%	Lower accuracy than deep or hybrid approaches	TBC
(Rawat et al., 2026)	Online social network dataset	Carbon-aware stacked ensemble	Integrated explainability	NR	Performance and efficiency not detailed	TBC
(Kehkashan et al., 2025)	Kaggle phishing websites (11,000+ URLs)	Random Forest	SHAP feature selection	Accuracy 97%	URL features only	TBC
(Akhtar et al., 2025)	Four public phishing datasets	XGBoost	LIME	Accuracy 99.65%	No unified latency analysis	TBC
(Fatima et al., 2025)	PhishTank	Random Forest	LIME and SHAP	NR	Quantitative performance not detailed	TBC
(Upreti et al., 2025)	10,000 websites (48 features reduced to 36)	WOA feature selection with H2O AutoML	SHAP	R2 0.9534; RMSE 0.1079; accuracy 98.3%	Explanation quality not an optimization objective	TBC
(Ghadami & Rahebi, 2025)	Phishing website or URL dataset	GAN, CNN, GRU, FMRCNN, Transformer hybrid with BKA	SHAP	Accuracy 98.67%	High architectural complexity and training cost	TBC

Study	Dataset	Model	XAI method	Reported performance	Main limitation	QA
(Yeboah et al., 2026)	Multi-domain phishing and web attack data	Ensemble of self-supervised pre-trained models	NR	Cross-domain accuracy 94.4%	Complexity limits real-time interpretability	TBC
(Ibrahim & Elhafiz, 2026)	Nazario and Enron emails (19,000)	BERT and RoBERTa	None (added layer needed)	Accuracy: RoBERTa 94%; BERT 91%	Resource intensive	TBC
(Tawfik et al., 2025)	Two public URL datasets (88,647 and 235,795)	Few-shot ModernBERT (XF-PhishBERT)	Explainable few-shot design	Accuracy 98.5% to 99.9%	Few-shot generalization still maturing	TBC
(Uddin et al., 2026)	Phishing email dataset	Transformer and LLM	LITA (LIME and Transformer attribution)	NR	High inference cost	TBC

Note. NR = not reported in the extracted source; TBC = QA score to be completed from the appraisal worksheet. The remaining 34 of the 48 included studies are to be added from the extraction sheet of the author, because the reference list combines included studies with background and secondary sources.

REFERENCES

Abdillah, R., Shukur, Z., Mohd, M., & Murah, M. Z. (2022). Phishing classification techniques: A systematic literature review. *IEEE Access, 10*, 41574–41591. https://doi.org/10.1109/ACCESS.2022.3166474

Abrardi, L., Comino, S., & Grassini, S. (2025). The economics of cyber risk: A survey of the literature. *Journal of Industrial and Business Economics*. https://doi.org/10.1007/s40812-025-00370-3

Ahmed, R., & EP, S. (2026). An intelligent phishing email detection system using ensemble methods and explainable AI. *Knowledge-Based Systems, 337*, 115388. https://doi.org/10.1016/j.knosys.2026.115388

Aisyah, D. N., Setiawan, A. H., Lokopessy, A. F., Faradiba, N., Setiaji, S., Manikam, L., & Kozlakidis, Z. (2024). The information and communication technology maturity assessment at primary health care services across 9 provinces in Indonesia: Evaluation study. *JMIR Medical Informatics, 12*, e55959. https://doi.org/10.2196/55959

Akhtar, H. M. U., Nauman, M., Hameed, N. A. M., Hameed, S., & Tareen, M. Z. (2025). Mitigating cyber threats: Machine learning and explainable AI for phishing detection. *VFAST Transactions on Software Engineering, 13*(2). https://doi.org/10.21015/vtse.v13i2.2129

Alevrido, R., & Pasha, J. A. (2025). Implementation of cyber policy in Indonesia: Study of data hacking control by BSSN, Dittipidsiber, and the Ministry of Communication and Information at BPJS Health in 2021. *JISIP (Jurnal Ilmu Sosial dan Pendidikan), 9*(3), 1063–1069. https://doi.org/10.58258/jisip.v9i3.8709

Alharbi, M., et al. (2026). A novel framework for phishing detection based on backward recursive feature selection. *Egyptian Informatics Journal, 33*, 100922. https://doi.org/10.1016/j.eij.2026.100922 [complete author list to be confirmed against the publisher record]

Almaktoom, G., Aladhadh, S., & El Khediri, S. (2026). Privacy-preserving phishing detection: A systematic review of LLMs, federated learning, and blockchain integration. *Computer Modeling in Engineering & Sciences, 147*(2). https://doi.org/10.32604/cmes.2026.078774

Almomani, O., Al-Zyoud, M., Alsarhan, Y., Almomani, A., Khdour, T., Alomari, K. M., & Salloum, S. A. (2026). A metaheuristic feature selection model using bat optimization for malicious URL attack detection. *Scientific Reports, 16*, 20944. https://doi.org/10.1038/s41598-026-51981-2

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

Alshahrani, S. M., Khan, N. A., Almalki, J., & Al Shehri, W. (2022). URL phishing detection using particle swarm optimization and data mining. *Computers, Materials & Continua, 73*(3), 5625–5640. https://doi.org/10.32604/cmc.2022.030982

Alsuqayh, N., Mirza, A., & Alhogail, A. (2025). Exploring feature engineering and explainable AI for phishing website detection: A systematic literature review. *International Journal of Electrical and Computer Engineering, 15*(6), 5863–5878. https://doi.org/10.11591/ijece.v15i6.pp5863-5878

Althobaiti, K., & Alsufyani, N. (2024). A review of organization-oriented phishing research. *PeerJ Computer Science, 10*, e2487. https://doi.org/10.7717/peerj-cs.2487

Anti-Phishing Working Group. (2026). *Phishing activity trends report, 4th quarter 2025*. https://docs.apwg.org/reports/apwg_trends_report_q4_2025.pdf

Bacanin, N., Zivkovic, M., Antonijevic, M., Venkatachalam, K., Lee, J., Nam, Y., & Abouhawwash, M. (2023). Addressing feature selection and extreme learning machine tuning by diversity-oriented social network search: An application for phishing websites detection. *Complex & Intelligent Systems, 9*, 7269–7304. https://doi.org/10.1007/s40747-023-01118-z

Badan Siber dan Sandi Negara. (2025). *Lanskap keamanan siber Indonesia tahun 2024*. Badan Siber dan Sandi Negara.

Calzarossa, M., Giudici, P., & Zieni, R. (2024). Explainable machine learning for phishing feature detection. *Quality and Reliability Engineering International, 40*, 362–373. https://doi.org/10.1002/qre.3411

Capuano, N., Fenza, G., Loia, V., & Stanzione, C. (2022). Explainable artificial intelligence in cybersecurity: A survey. *IEEE Access, 10*, 93575–93600. https://doi.org/10.1109/ACCESS.2022.3204171

Catal, C., Giray, G., Tekinerdogan, B., Kumar, S., & Shukla, S. (2022). Applications of deep learning for phishing detection: A systematic literature review. *Knowledge and Information Systems, 64*(6), 1457–1500. https://doi.org/10.1007/s10115-022-01672-x

Challacombe, D. J., & McElhiney, E. N. (2025). Phishing susceptibility among healthcare workers: The impact of awareness, email type, and location. *International Journal of Advanced Corporate Learning, 18*(1), 4–15. https://doi.org/10.3991/ijac.v18i1.51671

Chandre, P., Bhujbal, P., Jadhav, A., Shendkar, B. D., Wangikar, A., & Sachdeo, R. (2025). A comprehensive review of interpretable machine learning techniques for phishing attack detection. *IAES International Journal of Artificial Intelligence, 14*(4), 3022–3032. https://doi.org/10.11591/ijai.v14.i4.pp3022-3032

Charmet, F., Tanuwidjaja, H. C., Ayoubi, S., Gimenez, P.-F., Han, Y., Jmila, H., Blanc, G., Takahashi, T., & Zhang, Z. (2022). Explainable artificial intelligence for cybersecurity: A literature survey. *Annals of Telecommunications, 77*, 789–812. https://doi.org/10.1007/s12243-022-00926-7

Connolly, A., & Atlam, H. F. (2025). Effective ensemble learning phishing detection system using hybrid feature selection. *Journal of Network and Computer Applications, 242*, 104251. https://doi.org/10.1016/j.jnca.2025.104251

Daoud, E., Garcia-Blas, J., Alawadi, S., & Carretero, J. (2026). Phishing in the age of distributed intelligence: Taxonomies, detection strategies, and the emerging role of federated learning. *Progress in Artificial Intelligence*. https://doi.org/10.1007/s13748-026-00448-6

Ding, W., Abdel-Basset, M., Hawash, H., & Ali, A. M. (2022). Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey. *Information Sciences, 615*, 238–292. https://doi.org/10.1016/j.ins.2022.10.013

Fasla, H., & Mehmood, A. (2026). Explainable artificial intelligence for web application threat detection: A systematic review. *Artificial Intelligence Review*. https://doi.org/10.1007/s10462-026-11581-y

Fatima, Z., Hassan Tanveer, M., Voicu, R. C., Chakravarty, S., Ashfaq, M., Khan, M., Zaib, A., & Desa, H. (2025). Explainable AI for IoT devices and robotic communication phishing detection: A machine learning approach using LIME and SHAP. *Engineering, Technology & Applied Science Research, 15*(5), 26478–26486. https://doi.org/10.48084/etasr.11595

Ghadami, R., & Rahebi, J. (2025). An explainable hybrid deep learning-optimization framework for robust phishing attack detection using GAN and transformer-based feature learning. *Ain Shams Engineering Journal, 16*(12), 103745. https://doi.org/10.1016/j.asej.2025.103745

Gusenbauer, M., & Haddaway, N. R. (2020). Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. *Research Synthesis Methods, 11*(2), 181–217. https://doi.org/10.1002/jrsm.1378

Haque, A. K. M. B., Islam, A. K. M. N., & Mikalef, P. (2023). Explainable artificial intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research. *Technological Forecasting and Social Change, 186*, 122120. https://doi.org/10.1016/j.techfore.2022.122120

He, Y., Aliyu, A., Evans, M., & Luo, C. (2021). Health care cybersecurity challenges and solutions under the climate of COVID-19: Scoping review. *Journal of Medical Internet Research, 23*(4), e21747. https://doi.org/10.2196/21747

Ibrahim, M., & Elhafiz, R. (2026). Phishing email detection using BERT and RoBERTa. *Computation, 14*(2), 46. https://doi.org/10.3390/computation14020046

Kavya, S., & Sumathi, D. (2025). Staying ahead of phishers: A review of recent advances and emerging methodologies in phishing detection. *Artificial Intelligence Review, 58*, 50. https://doi.org/10.1007/s10462-024-11055-z

Kehkashan, T., Abdelhaq, M., Al-Shamayleh, A. S., Huda, N., Yaseen, I. A., Ahmed, A. I. A., & Akhunzada, A. (2025). Explainable phishing website detection for secure and sustainable cyber infrastructure. *Scientific Reports*. https://doi.org/10.1038/s41598-025-27984-w

Kementerian Kesehatan Republik Indonesia. (2022). *Peraturan Menteri Kesehatan Republik Indonesia Nomor 24 Tahun 2022 tentang Rekam Medis* [Regulation of the Minister of Health of the Republic of Indonesia No. 24 of 2022 on medical records]. Kementerian Kesehatan Republik Indonesia. https://jdih.kemkes.go.id/common/dokumen/2022permenkes024.pdf

Kundu, S., Chaturvedi, R. P., Mishra, A., Agrawal, A., Aldossary, S. M., Ayadi, M., Adem, K. B., & Hashmi, A. (2026). A trustworthy cybersecurity model for transparent cyberattack detection using Bald Eagle Search tuned XGBoost and explainable AI. *Journal of Cloud Computing, 15*, 58. https://doi.org/10.1186/s13677-026-00878-6

Li, W., Manickam, S., Chong, Y.-W., Leng, W., & Nanda, P. (2024). A state-of-the-art review on phishing website detection techniques. *IEEE Access, 12*, 187976–188012. https://doi.org/10.1109/ACCESS.2024.3514972

Loh, H. W., Ooi, C. P., Seoni, S., Barua, P. D., Molinari, F., & Acharya, U. R. (2022). Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011-2022). *Computer Methods and Programs in Biomedicine, 226*, 107161. https://doi.org/10.1016/j.cmpb.2022.107161

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Mayo, C., Tchuindjang, M., Brohi, S., & Ersotelos, N. (2026). A temporally dynamic feature-extraction framework for phishing detection with LIME and SHAP explanations. *Future Internet, 18*(2), 101. https://doi.org/10.3390/fi18020101

Naqvi, B., Perova, K., Farooq, A., Makhdoom, I., Oyedeji, S., & Porras, J. (2023). Mitigation strategies against the phishing attacks: A systematic literature review. *Computers & Security, 132*, 103387. https://doi.org/10.1016/j.cose.2023.103387

Neprash, H. T., McGlave, C. C., Cross, D. A., Virnig, B. A., Puskarich, M. A., Huling, J. D., Rozenshtein, A. Z., & Nikpay, S. S. (2022). Trends in ransomware attacks on US hospitals, clinics, and other health care delivery organizations, 2016-2021. *JAMA Health Forum, 3*(12), e224873. https://doi.org/10.1001/jamahealthforum.2022.4873

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: An updated

guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

Puri, L., Singh, R., & Paul, J. (2026). Digital financial fraud: A review and research agenda. *Journal of Economic Surveys*. https://doi.org/10.1111/joes.70103

Rawat, R., Gupta, P., Sankaran, K. S., Dixit, V., Chourasia, A., & Rimal, Y. (2026). An explainable and carbon-aware stacked ensemble model for phishing and impersonation detection in online social networks. *Engineering Reports*. https://doi.org/10.1002/eng2.70792

Safi, A., & Singh, S. (2023). A systematic literature review on phishing website detection techniques. *Journal of King Saud University - Computer and Information Sciences, 35*(2), 590–611. https://doi.org/10.1016/j.jksuci.2023.01.004

Salloum, S. A., Gaber, T., Vadera, S., & Shaalan, K. (2022). A systematic literature review on phishing email detection using natural language processing techniques. *IEEE Access, 10*, 65703–65727. https://doi.org/10.1109/ACCESS.2022.3183083

Sánchez-Paniagua, M., Fernández-Fernández, E., Alegre, E., Al-Nabki, W., & González-Castro, V. (2022). Phishing URL detection: A real-case scenario through login URLs. *IEEE Access, 10*, 42949–42960. https://doi.org/10.1109/ACCESS.2022.3168681

Shafin, S. S. (2025). An explainable feature selection framework for web phishing detection with machine learning. *Data Science and Management, 8*(2), 127–136. https://doi.org/10.1016/j.dsm.2024.08.004

Sivaneswaran, D., Hewage, C. T. E. R., Herath, H. M. K. K. M. B., Rathore, R. S., Singh, V. K., & Jiang, W. (2026). A systematic literature review of large language models in phishing attack generation and detection. *Array, 30*, 100775. https://doi.org/10.1016/j.array.2026.100775

Tawfik, M., Abu-Ein, A. A., Abdelhaliem, A. H., Al-Sharo, Y. M., & Fathi, I. S. (2025). Explainable few-shot learning with Modern BERT for detecting emerging phishing attacks using XF-PhishBERT. *Scientific Reports, 15*, 42821. https://doi.org/10.1038/s41598-025-27500-0

Uddin, M. A., Mahiuddin, M., & Sarker, I. H. (2026). An explainable transformer-based model for phishing email detection: A large language model approach. *Computer Networks, 277*, 112061. https://doi.org/10.1016/j.comnet.2026.112061

Upreti, K., Shankar, U., Upreti, S., Radhakrishnan, G. V., Hundekari, S., & Date, S. (2025). AQUAPHISH: Leveraging metaheuristics and automated machine learning for precision phishing detection. *Journal of Applied Science and Technology Trends*. https://doi.org/10.38094/jastt61238

Verizon. (2025). *2025 data breach investigations report*. Verizon Business. https://www.verizon.com/business/resources/reports/dbir/

Visani, G., Bagli, E., Chesani, F., Poluzzi, A., & Capuzzo, D. (2022). Statistical stability indices for LIME: Obtaining reliable explanations for machine learning models. *Journal of the Operational Research Society, 73*(1), 91–101. https://doi.org/10.1080/01605682.2020.1865846

Yan, F., Wen, S., Nepal, S., Paris, C., & Xiang, Y. (2022). Explainable machine learning in cybersecurity: A survey. *International Journal of Intelligent Systems, 37*(12), 12305–12334. https://doi.org/10.1002/int.23088

Yeboah, P. N., Kayes, A. S. M., Rahayu, W., Pardede, E., & Mahbub, S. (2026). A framework for phishing and web attack detection using ensemble features of self-supervised pre-trained models. *Journal of Network and Computer Applications, 250*, 104474. https://doi.org/10.1016/j.jnca.2026.104474

Zafar, U., & Wu, F. (2026). Methodological challenges in explainable AI for fraud detection: A systematic literature review. *Artificial Intelligence Review, 59*, 115. https://doi.org/10.1007/s10462-026-11516-7

Zhou, Z., Hooker, G., & Wang, F. (2021). S-LIME: Stabilized-LIME for model explanation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 2429–2438). ACM.

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