© SEP 2026 | IRE Journals | Volume 10 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I3-1723012

Machine Learning-Based Phishing Detection Techniques: A Systematic Literature Review

EKPO PRECIOUS EMMANUEL[1], SOLOMON AHIBA[2], VIVIAN NWAOCHA[3], OJIMA GABRIELLA OB'LAMA[4], ERU AKWUMA NATHANIEL[5]

[1, 2, 4, 5]*Department of Information Technology, Faculty of Computing, Nile University of Nigeria, Abuja, Nigeria.*

[3]*Assoc. Prof., Department of Information Technology, Faculty of Computing, Nile University of Nigeria, Abuja, Nigeria.*

Abstract- ***Phishing remains a persistent cybersecurity threat because attackers combine social engineering with rapidly changing emails, websites, URLs, text messages, QR codes, and mobile channels. Static blacklists and manually defined rules are increasingly inadequate for previously unseen campaigns. This study systematically reviews machine learning-based phishing detection research published between 2020 and 2026. A PRISMA-guided protocol was used to identify, screen, appraise, and synthesise 50 studies. The review examined the techniques employed, their reported effectiveness, the datasets and evaluation metrics used, and the principal limitations and emerging research directions. Five technique families were identified: traditional machine learning, ensemble learning, deep learning, transformer and large-language-model approaches, and explainable or hybrid intelligent systems. Traditional classifiers remain attractive for efficiency and interpretability, while ensemble, deep, transformer, multimodal, and continual-learning methods increasingly address complex features, contextual language, concept drift, and cross-channel attacks. PhishTank, OpenPhish, the UCI Phishing Websites dataset, ISCX-URL2016, SpamAssassin, Enron, and related repositories were commonly used. Accuracy was reported in 48 of the 50 studies, followed by precision (43), recall (42), and F1-score (41), whereas ROC-AUC, Matthews correlation coefficient, and error-rate measures were less frequent. Dataset ageing, class imbalance, weak cross-dataset generalisation, adversarial manipulation, computational cost, and limited explainability remain major barriers. Practical progress depends on adaptive, lightweight, explainable, and robust models evaluated with current datasets and standardised multi-metric protocols.***

Index Terms— ***Cybersecurity, deep learning, explainable artificial intelligence, large language models, machine learning, phishing detection, systematic literature review, transformer models***

I. INTRODUCTION

The rapid growth of digital banking, e-commerce, cloud services, social media, and mobile communication has given phishing a larger scale and reach. Maltreated patrons are made to believe that they are dealing with credible organizations or individuals to convince them to divulge their credentials, financial information, or other sensitive data. Phishing that is contemporaneous goes beyond conventional email to include spear phishing, fake websites, malicious URLs, smishing, vishing, QR-code lures, and attacks embedded in the cloud and Internet-of-Things environments. [1], [2].

Blacklists, signatures, and handcrafted rules, which are conventional means of protection, and manual inspection are still helping to detect known indicators; however, their effectiveness decreases when attackers change domains, hide links, rewrite page contents, or customize messages. It has been shown that when it comes to tested performance, the composition of the dataset, the design of the features, and the validation practices have a major influence. [3], [4], [5]. The concept of machine learning provides an auxiliary method wherein it obtains statistical regularities in the URL syntax, host and domain attributes, HTML structure, message content, visual appearance, and user or network behavior.

The field of study has moved forward from the basic classifiers like Random Forest, Support Vector Machine, Decision Tree, Logistic Regression, Naive Bayes, and k-Nearest Neighbour to the sophisticated alternatives like voting and boosting ensembles, convolutional and recurrent neural networks,

transformer models, large language models, multimodal fusion, continual learning, and explainable artificial intelligence [6]-[16]. The aforementioned advances have noticeably elevating the feature learning and contextual analysis, but at the same time they come out with new demands for the labeled data, computation resources, model governance, and interpretable decisions.

The research findings are still scattered, although there is a rapid increase. The research materials include attributes that are somewhat different among phishing channels in public and proprietary datasets, class distributions, train-test protocols, and metrics. This is due to the fact that a single benchmark can yield high accuracy in specific samples which may not be a sign of resistance to the new campaigns or the dependable operational performance [3], [17]-[20]. Current changes in the field of explainable AI, transformer architectures, large language models, multimodal learning, and incremental adaptation raise the complexity of comparing the present with the earlier methods.

This article consolidates the project findings through four questions: (1) which machine learning techniques are used for phishing detection; (2) how effective are the reported approaches; (3) which datasets, features, and evaluation metrics are most common; and (4) which limitations and emerging directions should guide future work. Its contribution is a compact taxonomy of methods, a synthesis of dataset and metric practice, and deployment-oriented recommendations for research and cybersecurity practice.

II. LITERATURE REVIEW

A. Phishing Data and Feature Spaces

Detecting phishing through machine learning is initiated by the interpretation of the attack's modulus. URL-oriented systems are designed to analyze the lexical features like length, character patterns, tokens, subdomains, redirections, and obfuscation. On the other hand, the website detectors include HTML, hyperlink, form, script, and certificate as well as host-based attributes, while email and messaging systems utilize headers, sender identity, subject and body text, attachments, embedded links, and semantic cues [6], [21]. The visual and behavioral signals are being used more and more often to identify the pages that look like trusted brands or campaigns that distribute the weak indicators through several channels.

Model selection and operational costs are influenced by the features' source. For fast classical classifiers, Structured URL and site properties are the main contributors, while raw text, long sequences, screenshots, and heterogeneous evidence facilitate the use of deep or multimodal architectures. Feature engineering is a promising remedy for both discrimination and computational efficiency but it should be noted that features that work well in one dataset might fail to do so in another, provided the domains, labeling rules, and attack periods are different [17], [18].

B. Machine Learning Technique Families

Supervised classifiers are common can because they do not necessarily require much cost for training and deployment. Random Forest is very likely to withstand overfitting and it manages mixed features, while Support Vector Machine signifies efficiency in spaces that have become high dimensions. Decision Tree and Logistic Regression are able to give unambiguous decision logic and probabilities, however, Naive Bayes and k-Nearest Neighbour were commonly used text or baseline classifiers [7]. The main restriction they face is the manual feature engineering requirement and the training distributions that are too static.

Ensemble learning is a method used to bring together many learners through the techniques of bagging, boosting, voting, or stacking. The reviewed studies generally correlate the use of ensembles with better stability and a lower variance in comparison to that of a single classifier, especially in URL and website classification [8], [9]. Deep learning has the benefit of less dependence on manual feature design by learning hierarchical representations from raw URLs, HTML, text, or images. CNN, LSTM, GRU, and CNN-LSTM models are the most notable; however, their data and compute requirements may hinder real-time or resource-limited deployment [10], [11].

© SEP 2026 | IRE Journals | Volume 10 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I3-1723012

Phishing mail, message and URL string transformation models are no longer just for contextual language sense. With regard to long-range dependencies and delicate semantic manipulation, BERT and similar attention-based architectures can capture what ordinary bag-of-words or recurrent models may omit [12]. Explainable AI methods like SHAP and LIME are being used in conjunction with top-notch classifiers to identify which features have affected a decision [13], [14]. Both transformers and the large language models framework introduce adaptability and increase the richness of semantic reasoning, but their issues such as latency, transparency, cost, misuse, and governance are always debated [15], [16].

C. Evaluation and Operational Relevance

A phishing detector is generally judged not only by its accuracy. Precision quantifies the share of predicted phishing incidents that are really malicious, recall points out the possibility of finding the real attacks, F1-score integrates both functions. When classes are imbalanced ROC-AUC, false-positive and false-negative rates, specificity, and Matthews correlation coefficient are some other metrics that can help to gain insights. Besides, factors such as inference time, memory, alert volume, explanation usefulness, model drift, and cost of wrong decisions are crucial operationally.

The analytical chain that was located in the reviewed literature is shown in the figure 1. The phishing evidence is translated into lexical, structural, semantic, visual, or behavioral features; learning models generate classifications or risk scores; and evaluation has to relate the technical metrics to deployability, explainability, and adaptation. Quality of dataset, privacy, adversarial robustness, drift monitoring, and human oversight are issues transcending all steps.

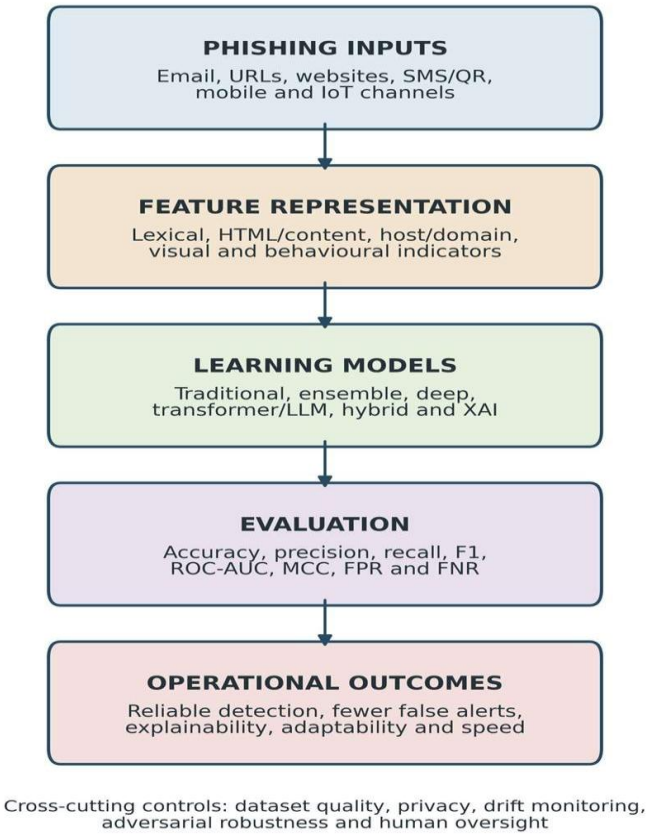

Figure 1: Conceptual framework for machine learning-based phishing detection.

III. METHODOLOGY

A. Review Design and Questions

In this study, we have used a systematic literature review approach to carry out the identification, evaluation, and integration of the evidence but not to create a new experimental dataset. The report is consistent with the PRISMA 2020 guidelines [22], while the protocol design, search, screening, extraction, and synthesis were all based on software engineering review guidance [23]. The protocol was constructed based on the four research questions highlighted in the Introduction.

Table 1: Review objectives and evidence extraction

RQ	Evidence extracted	Synthesis output
RQ1	Algorithms, architectures and learning paradigms	Technique taxonomy
RQ2	Reported performance, strengths and constraints	Effectiveness comparison
RQ3	Datasets, features, metrics and validation	Evaluation profile
RQ4	Challenges, limitations and proposed future work	Research agenda

B. Search Strategy and Eligibility

The databases accessed were IEEE Xplore, ACM Digital Library, SpringerLink, ScienceDirect, Wiley Online Library, and MDPI. The search phrases mixed phishing terminologies with machine learning terminologies through the use of Boolean operators. A typical statement used was ("phishing detection" OR "phishing website" OR "phishing email" OR smishing) AND ("machine learning" OR "deep learning" OR transformer OR "explainable AI"). Besides this, other terms such as web addresses, large language models, multimodal learning, and incremental learning were also employed.

The only articles under consideration were those written in English, presented on the internet of the institution of the researcher, or showcased in the conference samples fully published from 2020 to 2026. They should be the ones dealing with cutting-edge technologies like machine learning or artificial intelligence used for phishing detection should give enough methodological or evaluative details to imply extraction. An equivalent copy, a non-scholarly article, an editorial, documents that are out of the period, and an article that is not about phishing detection were not considered for the research study.

C. Study Selection, Quality Appraisal, and Extraction

By searching the database and employing the reference tracking techniques, the project picked a total of 155 items. After going through the processes of handling duplicates, screening titles and abstracts, assessing full texts for eligibility, and evaluating quality, 50 studies were selected for qualitative synthesis. The flowchart in Figure 2 illustrates the selection process through documentation while showing both the initial and final reconciled counts. The quality appraisal checking involved aspects like relevance, the learning method's clarity, description of data and features, evaluation practice, practical testing, limitation, reproducibility, and the contribution to the study.

© SEP 2026 | IRE Journals | Volume 10 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I3-1723012

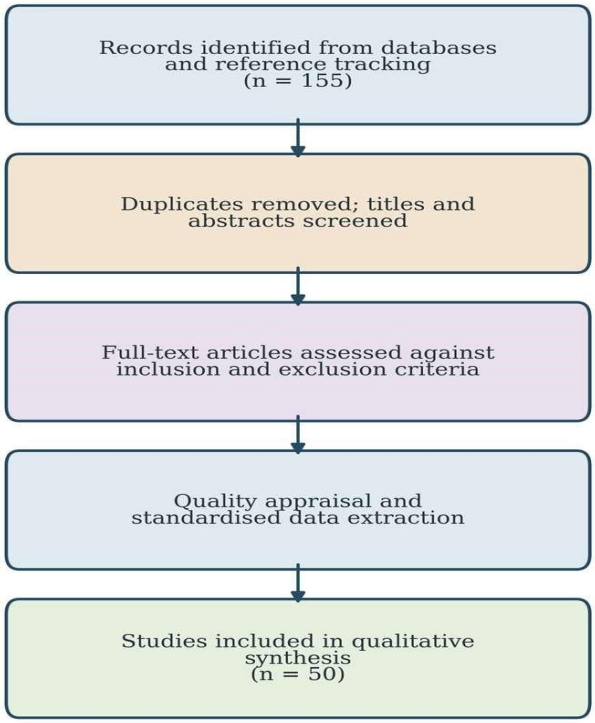

Figure 2: PRISMA-guided study-selection workflow.

An extraction form that is standard for literature analysis gathered necessary information such as author name and year of publication, research area, method of study, machine learning techniques, dataset, source of features, newly introduced evaluation metric, critical finding, mentioned limitation, and research gap. Technical and metric frequencies were condensed using the descriptive synthesis; thematic synthesis was employed to draw comparisons between advantages, data quality, adaptation, and interpretability as well as other variables like computing needs and new directions. A statistical meta-analysis was not conducted since the items utilized diverse characteristics, features, and validation parameters.

IV. RESULTS AND DISCUSSION

A. Technique Families and Reported Effectiveness

The 50-study evidence base illustrates a logical flow from regular supervised classification to the group, deep, transformer-based, explainable, and hybrid systems. As for the structured URL and website data, the traditional machine learning method is still the most uncomplicated baseline. Random Forest, Support Vector Machine, Decision Tree, Logistic Regression, Naive Bayes, and k-Nearest Neighbour are algorithms that top the list due to their quick training time, multipurpose capabilities, and proven performance even with little pre-processing or feature engineering [7], [24].

Multiple methods combine the decision boundaries that complement and thereby attain better resilience in the case of phishing samples having different varieties. The main features that are effective with voting, boosting, bagging and stacking are the reduction of such decision errors Physicist White calls them "fallacies" - that might have been produced by only one weak classifier [8], [9], [25]. Yet, they still need representative labels and careful adjustments, and their seemingly better performance may vanish when tested on a different or newer dataset.

The major application area for deep learning is the email, URL, and content-based detection. CNNs discover the local patterns, recurrent models yield the order, and the combination of both functions form hybrid CNN-LSTM architectures [10], [11]. The biggest benefit is the automatic extraction of features, but the main drawbacks are the need for substantial training data, a high computational load, and the lack of transparency. Transformer models, which provide a further contextual understanding and achieve favorable results for dealing with advanced phishing vocabularies and semantic matching, are [12], [26].

The most prioritized technical direction of recent publications is the deployment of explainable and hybrid techniques. These are the most directed time on the practical problem-solving of the issues at hand. The SHAP, LIME, and attention visualization can also facilitate the task of the analyst by determining the significant URL, content, or host attributes [13], [14], [27], [28]. The systems using multimodal technology, on the other hand, integrate sensor, image, text, and human behavior inputs, while the continual-learning approaches reset their knowledge base and adapt as per the project needs [29], [30], [31], [32]. The systems with LLM's assistance are also promising for message comprehension and messages description, while they need to be subjected to more stringent validation in the areas of security, consistency, latency, and governance [16], [33].

No single family is uniformly best. Reported accuracy is shaped by the task, dataset, feature construction, class balance, optimisation, and validation protocol. Classical and ensemble models may be preferable for low-latency URL screening; deep and transformer models are more suitable for raw, sequential, or semantic data; and hybrid or explainable systems are attractive where multiple channels and analyst trust matter. Table 2 summarises these trade-offs.

Table 2: Suitability of machine learning approaches for phishing detection

Approach	Best-suited role / strength	Principal limitation
Traditional ML	Fast URL or website classification with engineered features	Feature dependence and weak adaptation
Ensemble	Robust classification and reduced variance	Tuning and label dependence
Deep learning	Automatic learning from text, URL, HTML or image data	Data, compute and black-box behaviour
Transformers / LLMs	Contextual email, message and semantic analysis	Cost, latency, governance and explainability
Explainable / hybrid / continual	Cross-modal, adaptive and analyst-facing detection	Limited cross-dataset and live validation

B. Datasets and Feature Sources

Frequent mentions were found for the PhishTank, OpenPhish, the UCI Phishing Websites dataset, ISCX-URL2016, URLhaus, SpamAssassin, Enron, and research-specific repositories. PhishTank and OpenPhish are the platforms that have active malicious URLs but mostly they need a separate source of legitimate samples as a must. UCI has structured benchmarking support but may miss the contemporary attack protocols. The really important SpamAssassin and Enron databases can be used for the email experiments while the new smishing and IoT studies are utilizing the data from the specific channels more [34]-[37].

© SEP 2026 | IRE Journals | Volume 10 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I3-1723012

Feature representations include lexical URL attributes, HTML and page-content indicators, domain and host data, certificate information, email metadata and text, visual similarity, and behavioural signals. Combining several feature sources usually provides richer coverage than a single modality, but also increases collection cost, latency, missing-data risk, and privacy requirements [6], [29], [21].

Dataset quality was one of the strongest recurring concerns. Duplicate samples, inconsistent labels, domain overrepresentation, class imbalance, and ageing data can inflate within-dataset results and reduce performance on unseen campaigns [17]-[20]. Cross-dataset evaluation is therefore more informative than a single random split. Recent work on generalised malicious-URL detection and multimodal BERT models similarly emphasises temporal and source diversity [38], [32].

C. Evaluation Metrics and Validation Practice

Accuracy dominated reporting: 48 studies (96%) used it. Precision appeared in 43 studies (86%), recall in 42 (84%), and F1-score in 41 (82%). ROC-AUC was used in 16 studies (32%), specificity in nine (18%), false-positive rate in seven (14%), Matthews correlation coefficient in six (12%), false-negative rate in five (10%), detection rate in four (8%), and Cohen's kappa in two (4%). Figure 3 shows the distribution.

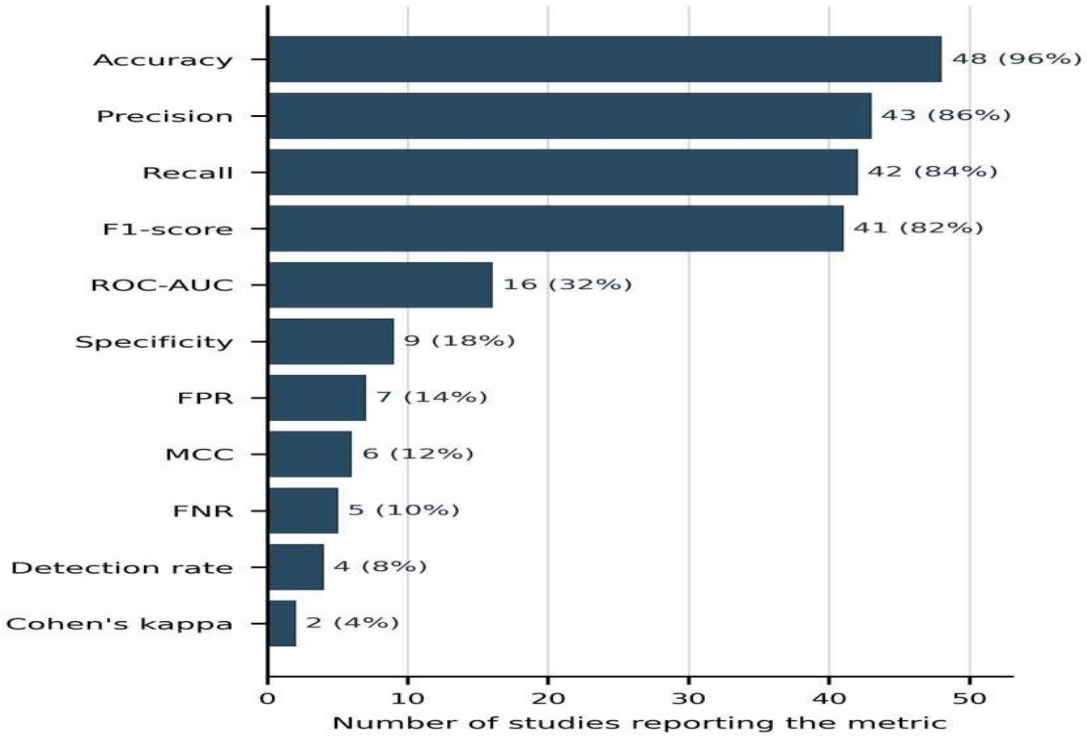

Figure 3: Frequency of evaluation metrics in the reviewed studies (n = 50).

Accuracy alone is quite misleading but in the case of phishing and legitimate classes as being imbalanced it is not at all good. The situations where precision and false-positive rate are important arise when excessive blocking or warnings undermine trust; whereas recall and false-negative rate are far more critical than those since missed attacks lead to direct security exposure, the F1-score value at is what you can call a balance.

while MCC uses all four confusion-matrix outcomes and is particularly informative for imbalanced binary classification [39]. A trustworthy evaluation must, therefore, consist of various complementary

measures, give an account of the class distributions, make proper use of cross-validation, and include cross-dataset or temporally separated testing whenever it is feasible.

D. Challenges and Emerging Research Directions

Concept drift is one of the main difficulties encountered in the operation. Areas, models, companies, language, and concealment techniques continuously change, hence a model concocted from past data could deteriorate even without a clear failure. Learning that is ongoing and incremental can cut back the need for full retraining, but they have to deal with issues like catastrophic forgetting, label noise, and feedback bias [15], [30]. Thus, regular drift monitoring and time-aware evaluation are the necessary parts of deployment.

The issues of adversarial manipulation and weak generalization are still open. For instance, tiny adjustments to URLs or features of a webpage can trick models based on fragile correlations, and the excellent benchmark scores could have plummeted due to distribution shift [19]. Robustness testing has to encompass obfuscation, adversarial perturbation, and unseen domains, as well as data collected post model training. Benchmarking must also register duplicates, label provenance, collection dates, and sampling decisions.

Trust, auditing, and incident handling necessitate the need for explainability; however, adherence to explanation fidelity and usability is not necessarily the result of simply appending an XAI tool. A proper explanation must point out the proofs of a security analyst's validation and also should be assessed for stability, accuracy, and actionability [27], [28]. In similar context, lightweight architectures, distillation, feature selection, and client-side models are equally conducive, particularly while they confront constraints Latency, bandwidth, memory, or privacy when deploying [40]. The best and most outstanding future direction is not a one-point decision but a complex integration: The current and various datasets; multimodal or hybrid features; adaptive learning; explainable outputs; adversarial and cross-dataset testing; and clear human oversight. Transformer and LLM systems have to be set against efficient baselines and not be taken for granted as better, and their simultaneous ability to detect or create convincing phishing content requires true governance [16], [26]. Privacy-preserving and federated learning, multilingual evaluation, and standardized benchmark suites are also significant areas of research that deserve to be examined thoroughly.

E. Review Limitations

The combination of these studies is qualitative since the studies vary in terminologies, exercises, data, functionalities and class distribution as well as measures for the software validation. The review included only English-language journals published over the selected time period and that some new issues may have fewer studies than standard URL or website classification. The results must therefore be seen as recommendations for the layout and writing and not as a list of algorithms ranging from best to worst.

V. CONCLUSION AND RECOMMENDATIONS

This literature review is a clear indication that machine learning has completely taken over phishing detection across various platforms, modes, and even the forthcoming multimodal paragraph. We still have traditional classifiers that keep on bringing a lot of speed, simplicity, and interpretability; ensembles are the ones ensuring stability; deep and transformer networks are the ones doing feature learning automatically and contextually; and also explainable, hybrid, multimodal, and continual systems tackle trust and dynamic attack behavior. The pieces of evidence, however, do not endorse any single universally excellent methodology. Efficiency is contingent on the interplay between the model, feature representation, dataset quality, class distribution, validation protocol, and operational constraints.

The necessity of diverse and regularly updated datasets, documentation of the data collection and labeling procedures, and cross-dataset or temporal validation should be the primary tasks of the researchers. Besides, the accuracy of the

classification results should be bolstered through precision, recall, F1-score, ROC-AUC, MCC, and error-rate metrics. Program designers should base their choice of models on the type of phishing channel as well as the deployment environment, use rule-based controls jointly with machine learning where it makes sense, and further incorporate explanations/drift monitoring/adversarial testing/human review to the system design.

Lightweight and privacy-conscious transformer or multimodal models, continual learning that adapts safely to new campaigns, multilingual and cross-channel benchmarks, and explanation methods that are useful to analysts rather than merely visually persuasive, should be the future work. Collaboration between researchers, industry, phishing-data providers, and cybersecurity operations groups are the foundation for the transition of the high experimental performance into the reproducible and reliable real-world protection.

REFERENCES

[1] Z. Alkhalil, C. Hewage, L. Nawaf, and I. Khan, "Phishing attacks: A recent comprehensive study and a new anatomy," *Frontiers in Computer Science*, vol. 3, 2021. Crossref

[2] S. Kavya and D. Sumathi, "Staying ahead of phishers: A review of recent advances and emerging methodologies in phishing detection," *Artificial Intelligence Review*, vol. 58, no. 2, 2024. Crossref

[3] A. E. Aassal, S. Baki, A. Das, and R. Verma, "An in-depth benchmarking and evaluation of phishing detection research for security needs," *IEEE Access*, vol. 8, pp. 22170–22192, 2020. Crossref

[4] R. Abdillah, Z. Shukur, M. Mohd, and M. Z. Murah, "Phishing classification techniques: A systematic literature review," *IEEE Access*, vol. 10, pp. 41574–41591, 2022. Crossref

[5] K. Subashini and V. Narmatha, "Detecting phishing websites using recent techniques: A systematic literature review," *ITM Web of Conferences*, vol. 57, art. no. 01008, 2023. Crossref

[6] A. Aljofey et al., "An effective detection approach for phishing websites using URL and HTML features," *Scientific Reports*, vol. 12, art. no. 8842, 2022. Crossref

[7] Y. Al-Tamimi and M. Shkoukani, "Employing cluster-based class decomposition approach to detect phishing websites using machine learning classifiers," *International Journal of Data and Network Science*, vol. 7, no. 1, pp. 313–328, 2023. Crossref

[8] Y. Wei and Y. Sekiya, "Sufficiency of ensemble machine learning methods for phishing websites detection," *IEEE Access*, vol. 10, pp. 124103–124113, 2022. Crossref

[9] N. Innab et al., "Phishing attacks detection using ensemble machine learning algorithms," *Computers, Materials & Continua*, vol. 80, no. 1, pp. 1325–1345, 2024. Crossref

[10] Z. Alshingiti, R. Alaqel, J. Al-Muhtadi, E.-H. Qazi, K. Saleem, and M. H. Faheem, "A deep learning-based phishing detection system using CNN, LSTM, and LSTM-CNN," *Electronics*, vol. 12, no. 1, art. no. 232, 2023. Crossref

[11] U. A. Butt, R. Amin, H. Aldabbas, S. Mohan, B. Alouffi, and A. Ahmadian, "Cloud-based email phishing attack using machine and deep learning algorithm," *Complex & Intelligent Systems*, vol. 9, no. 3, pp. 3043–3070, 2023. Crossref

[12] R. Melendez, M. Ptaszynski, and F. Masui, "Comparative investigation of traditional machine-learning models and transformer models for phishing email detection," *Electronics*, vol. 13, no. 24, art. no. 4877, 2024. Crossref

[13] 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. Crossref

[14] T. Kehkashan et al., "Explainable phishing website detection for secure and sustainable

cyber infrastructure," *Scientific Reports*, vol. 15, art. no. 41751, 2025. Crossref

[15] M. N. Suhaimee et al., "Real-time incremental transformer with continual learning for adaptive phishing email detection," pp. 1–6, 2025.

[16] D. Sivaneswaran, C. T. E. R. Hewage, H. M. K. K. M. B. Herath, R. S. Rathore, V. K. Singh, and W. Jiang, "A systematic literature review of large language models in phishing attack generation and detection," *Array*, vol. 30, art. no. 100775, 2026. Crossref

[17] A. Hannousse and S. Yahiouche, "Towards benchmark datasets for machine learning based website phishing detection: An experimental study," *Engineering Applications of Artificial Intelligence*, vol. 104, art. no. 104347, 2021. Crossref

[18] P. Afonso, E. Maia, I. Amorim, and I. Praca, "Rethinking phishing detection: How dataset quality affects model generalization," pp. 542–547, 2025.

[19] B. Sabir, M. A. Babar, R. Gaire, and A. Abuadbba, "Reliability and robustness analysis of machine learning based phishing URL detectors," *IEEE Transactions on Dependable and Secure Computing*, pp. 1–18, 2022. Crossref

[20] I. Skula and M. Kvet, "A framework for preparing a balanced and comprehensive phishing dataset," *IEEE Access*, vol. 12, pp. 53610–53622, 2024. Crossref

[21] C. Opara, Y. Chen, and B. Wei, "Look before you leap: Detecting phishing web pages by exploiting raw URL and HTML characteristics," *Expert Systems with Applications*, vol. 236, art. no. 121183, 2024. Crossref

[22] M. J. Page et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *BMJ*, vol. 372, art. no. n71, 2021. Crossref

[23] B. Kitchenham and S. Charters, *Guidelines for Performing Systematic Literature Reviews in Software Engineering*, EBSE Technical Report EBSE-2007-01, Keele University and Durham University, 2007.

[24] A. Awasthi and N. Goel, "Phishing website prediction using base and ensemble classifier techniques with cross-validation," *Cybersecurity*, vol. 5, art. no. 22, 2022. Crossref

[25] R. S. Rao, C. Kondaiah, A. R. Pais, and B. Lee, "A hybrid super learner ensemble for phishing detection on mobile devices," *Scientific Reports*, vol. 15, art. no. 16839, 2025. Crossref

[26] M. A. Uddin and I. H. Sarker, "An explainable transformer-based model for phishing email detection: A large language model approach," *SSRN Electronic Journal*, 2024. Crossref

[27] Z. Fatima et al., "Explainable AI for IoT devices and robotic communication phishing detection," *Engineering, Technology & Applied Science Research*, vol. 15, no. 5, pp. 26478–26486, 2025. Crossref

[28] M. C. Calzarossa, P. Giudici, and R. Zieni, "An assessment framework for explainable AI with applications to cybersecurity," *Artificial Intelligence Review*, vol. 58, no. 5, 2025. Crossref

[29] Y. Sun, G. Liu, X. Han, W. Zuo, and W. Liu, "FusionNet: An effective network phishing website detection framework based on multi-modal fusion," pp. 474–481, 2023. Crossref

[30] A. Ejaz, A. N. Mian, and S. Manzoor, "Life-long phishing attack detection using continual learning," *Scientific Reports*, vol. 13, art. no. 11488, 2023. Crossref

[31] A. Vulfin, A. E. Sulavko, V. Vasiliev, A. Minko, A. Kirillova, and A. Samotuga, "A multimodal phishing website detection system using explainable artificial intelligence technologies," *Machine Learning and Knowledge Extraction*, vol. 8, no. 1, art. no. 11, 2026. Crossref

[32] Y. Wei, M. Nakayama, and Y. Sekiya, "Enhancing generalization in phishing URL detection via a fine-tuned BERT-based multimodal approach," *IEEE Access*, vol. 13, pp. 131197–131216, 2025. Crossref

[33] A. Alhuzali, Q. Al-Qahtani, A. Niyazi, L. Alshehri, and F. Alharbi, "PhishNet: A real-time, scalable ensemble framework for smishing attack detection using transformers and LLMs," *Computers, Materials & Continua*, vol. 86, no. 1, pp. 1–19, 2025. Crossref

[34] S. Ariyadasa, S. Fernando, and S. Fernando, "PhishRepo: A seamless collection of phishing data to fill a research gap in the phishing domain," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 5, 2022.

[35] J. C. G. de Barros et al., "Piracema: A phishing snapshot database for building dataset features," *Scientific Reports*, vol. 12, art. no. 15149, 2022. Crossref

[36] F. Janez-Martino, R. Alaiz-Rodriguez, V. Gonzalez-Castro, E. Fidalgo, and E. Alegre, "A review of spam email detection: Analysis of spammer strategies and the dataset shift problem," *Artificial Intelligence Review*, vol. 56, no. 2, pp. 1145–1173, 2023. Crossref

[37] D. Timko and M. L. Rahman, "Smishing Dataset I: Phishing SMS dataset from Smishtank.com," pp. 289–294, 2024. Crossref

[38] Y.-D. Tsai, C. Liow, Y. S. Siang, and S.-D. Lin, "Toward more generalized malicious URL detection models," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 19, pp. 21628–21636, 2024. Crossref

[39] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, art. no. 6, 2020. Crossref

[40] G. D. Bispo et al., "PHILDER: Lightweight framework for intelligent phishing detection on resource-limited devices," *IEEE Access*, vol. 13, pp. 131967–131979, 2025. Crossref