International Journal of Emerging Engineering and Technology (IJEET)
ISSN (e): 2958-3764
Volume: 5, Number: 1-1, Pages: 19 - 25, Year: 2026

Artificial Intelligence for Phishing and Social-Engineering Detection: A Machine Learning Evaluation on a Benchmark URL Dataset

Muhammad Azeem Afzal [1], Usama Ahmad Mughal [1], Malik Hammad Hussain [2], Muhammad Daniyal Qadri [3], Muhammad Daniyal Baig [3,*] and Muhammad Ehsan Qadeer [4]

[1] Department of Cyber Security, NASTP Institute of Information Technology, Lahore, Pakistan
[2] Department of Software Engineering, NASTP Institute of Information Technology, Lahore, Pakistan
[3] Department of Artificial Intelligence, NASTP Institute of Information Technology, Lahore, Pakistan
[4] Department of Mechanical Engineering, University of Lahore, Lahore, Pakistan

[*]Corresponding author: Muhammad Daniyal Baig (Email: daniyalbaig@niit.edu.pk)

Received: 10/06/2026, Revised: 18/08/2026, Accepted: 10/09/2026

***Abstract*—**Phishing remains the leading initial-access vector in modern cyberattacks, and the emergence of large language models (LLMs) has made social-engineering content easier to produce and harder to distinguish from legitimate communication. This paper reviews recent (2023-2026) research on artificial-intelligence-based phishing detection and reports an original empirical evaluation on a public benchmark dataset of 88,647 labelled URLs described by 111 engineered features spanning URL, domain, directory/file, query-parameter, and network/WHOIS-derived attributes. Five supervised learning models logistic regression, linear support vector machine, random forest, gradient boosting, and a multilayer perceptron were trained and evaluated using stratified five-fold cross-validation. The random forest classifier achieved the strongest performance (accuracy = 0.970, F1 = 0.957, ROC-AUC = 0.995), consistent with the ensemble-dominance trend reported in the recent literature. A follow-up robustness simulation shows that although the model is highly resistant to random feature-level noise, this does not imply resistance to the targeted, semantically-aware evasion strategies enabled by generative AI, which recent studies show can bypass commercial filters. The paper concludes with a discussion of the dual-use nature of AI in this domain and directions for more adversarial-robust, content-aware detection systems.

***Index Terms*—**Phishing detection; machine learning; social engineering; large language models; adversarial robustness; ensemble classifiers.

I. INTRODUCTION

Threats to cybersecurity can be broadly divided into two groups: those that take advantage of system flaws and those that exploit human weaknesses. The second category includes phishing and related social-engineering techniques, which consistently rank among the most potent attack vectors at threat actors' disposal because they circumvent technical controls by influencing human judgment instead of cracking cryptography or taking advantage of software flaws. For a fraction of the effort and technical expertise, an attacker can obtain the same amount of access that a complex exploit chain might ordinarily need with just one convincing email, text message, or phone call. One of the most prevalent and financially harmful types of cybercrime is still phishing. With hundreds of thousands of new attacks reported each quarter and credential-related compromise accounting for a significant portion of the human element in breach investigations, industry reporting ranks phishing as one of the most popular initial-access techniques used in breaches [1, 2]. Instead of decreasing, attack volumes have continued to be high: APWG recorded 1,130,393 phishing attacks in the second quarter of 2025, 892,494 in the third quarter, and 853,244 in the fourth.

According to Verizon's 2025 [2] breach data, credential abuse was linked to about one-third of breaches involving a human element. Since many phishing attempts are never reported, these numbers probably underestimate the actual scope of the issue. Additionally, they do not adequately account for the increasing number of attacks that are delivered outside of email, such as SMS, voice calls, and messaging platforms, which are usually tracked separately, if at all. The technical response to phishing has evolved considerably over the past three decades. Early defenses relied on static blacklists of known malicious domains and simple heuristic rules, such as flagging URLs containing an "@" symbol or an unusually long hostname. These approaches proved brittle: attackers could evade a blacklist simply by registering a new domain, and heuristic rules were easy to reverse-engineer once published. This brittleness motivated a long research programme applying supervised machine learning to phishing detection, using engineered features drawn from URLs, page content, and

network metadata to train classifiers capable of generalizing beyond any single known attack [3]. Ensemble methods and, more recently, deep learning architectures have progressively improved detection accuracy on these benchmarks, and machine-learning-based filters are now a standard component of email security and web-browsing protection [4].

However, the widespread availability of big language models has caused another change in the threat landscape. LLMs enable attackers to create fluent, contextually appropriate, and personalized messages at scale, whereas phishing emails were previously identifiable by spelling mistakes and awkward phrasing, cues that both human recipients and early ML classifiers relied on [5,6]. They can even automate portions of the social-engineering pipeline itself, from target reconnaissance to message drafting and iterative refinement. Outside of email, the same creative change is apparent. A claimed 442% increase in vishing events between the first and second half of 2024 was caused by speech-cloning tools that only require a few seconds of reference audio; by 2027, deepfake-enabled voice fraud is expected to cause losses of about $40 billion worldwide [7].

As attackers shift away from increasingly sophisticated email channels, SMS-based smishing and chat-based social engineering are going in similar directions. This gives generative AI in general a dual-use dynamic feature: the same capability that generates more convincing attacks also gives defenders new detection methods, such as transformer-based email classifiers and LLM-assisted content reasoning that can assess a message's intent rather than just its surface features [7, 8]. This change reveals a discrepancy between the issue that phishing-detection research is intended to address and how it is usually carried out. A large portion of the early machine-learning research on phishing was based on a comparatively stable attacker profile that was typified by templated emails, static URL structures, and low linguistic fluency. Generative AI is currently actively undermining this profile on several fronts at once, including content quality, delivery channel, and degree of personalization. According to recent surveys, many detection studies continue to assess models on a single, outdated benchmark dataset, infrequently test cross-dataset or cross-channel generalization, and infrequently assess how a model responds to intentional, adversarial input perturbation as opposed to naturally occurring test data [9]. A detection system that is solely evaluated against old benchmarks runs the danger of providing enterprises that use it with a false sense of security by reporting high numbers on a threat that has already evolved.

This paper is motivated by that gap and structured to address it in two complementary ways. On the review side, it deliberately looks beyond the email/URL focus that dominates much of the existing survey literature to cover the fast-growing multichannel landscape of vishing, smishing, voice deepfakes, and the human-factors and explainability research emerging alongside it, treating phishing detection as part of a broader social-engineering defense problem rather than an isolated email-filtering task. On the empirical side, rather than reporting accuracy on a benchmark dataset in isolation, the study also runs a lightweight adversarial-perturbation simulation to ask a more pointed question: how much does a strong model's performance degrade when its inputs are deliberately altered, even in a simple, non-adaptive way? This does not resolve the generalization gap identified above, but it is offered as a small, concrete step toward the kind of robustness-aware evaluation the literature increasingly calls for. Accordingly, this paper makes three contributions.

First, it synthesizes the current (2023-2026) literature on AI-driven phishing and social-engineering detection, covering classical machine learning, ensemble methods, deep learning, the dual-use role of LLMs, and, going beyond the email/URL focus of most prior surveys, the fast-growing multichannel landscape of vishing, smishing, voice deepfakes, and human-factors/explainability research. Second, it presents an original empirical evaluation on a large, publicly available, peer-reviewed benchmark dataset, comparing five representative model families under identical conditions. Third, it goes beyond standard accuracy metrics by running a lightweight adversarial-perturbation simulation to probe model robustness, a dimension increasingly emphasized in the literature but often omitted from smaller student and applied projects.

II. LITERATURE REVIEW

A. Classical and Ensemble Machine Learning Approaches

In order to feed classifiers like decision trees, naive Bayes, and support vector machines, early phishing detection studies depended on manually created features derived from URLs, page content, and metadata [10]. Because they aggregate weak, uncorrelated decision boundaries and handle the mixed, high-dimensional, and occasionally noisy features typical of phishing data, ensemble methods like random forest, gradient boosting, and extra-trees consistently outperform single learners in comparative surveys [4, 9]. A 2025 comprehensive survey of malicious-URL detection concludes that ensembles combined with modern representation learning now dominate the field and suggests evaluation beyond raw accuracy using metrics like PR-AUC and Matthews correlation coefficient for imbalanced settings. More recent work pushes this further with stacking and voting super-learners that combine classical models with deep and hybrid architectures, reporting improved stability across heterogeneous datasets and imbalanced classes.

B. Deep Learning and Multimodal Detection

Deep-learning approaches like CNNs, LSTMs, and CNN-LSTM hybrids—have been applied to raw URL strings, email bodies, and HTML/DOM structure, often outperforming classical ML when sufficient training data is available [11, 12]. Multimodal systems that fuse HTML DOM graphs with URL features via graph convolutional and transformer networks, or that incorporate visual/screenshot signals, have shown further gains by capturing cues that purely lexical features miss [13]. The PhiUSIIL dataset and framework extended this line of work by combining similarity-index features with incremental learning to keep pace with evolving phishing infrastructure. Transformer-based text classifiers such as BERT and RoBERTa have reported detection accuracies above 98-99% on

merged phishing-email corpora, generally outperforming traditional ML by several percentage points, though at higher computational and data cost.

C. Large Language Models: A Dual-Use Technology

The two-sided role of LLMs is the most important recent development. On the offensive side, researchers have demonstrated that models like GPT-4o can produce phishing emails that are sufficiently contextual and fluent to avoid commercial spam filters, doing away with the grammatical cues that previous detectors depended on [5]. Adversarial generator-critic loops are used by generative frameworks like SpearBot to autonomously revise phishing content, while agent-based frameworks that coordinate LLMs to do reconnaissance and tailor spear-phishing content at scale have also been published [6]. LLMs are increasingly being employed directly as classifiers or reasoning engines on the defensive side: While benchmarking studies compare LLM-based and refined transformer detectors for small and medium businesses [8], ChatSpamDetector asks an LLM to reason about email intent [7]. Some detectors rely on stylometric features to differentiate AI-generated from human-written text because LLM-generated phishing lacks the lexical artifacts of older attacks; gradient-boosted classifiers trained on stylometric feature sets have been reported to separate AI-generated phishing from legitimate email with high accuracy [8], demonstrating a shift toward provenance-based rather than purely content-based detection.

D. Datasets and Open Challenges

In this field, the choice of dataset has a significant impact on reproducibility. The UCI Phishing Websites dataset [14], the PhiUSIIL URL corpus, the Nazario and SpamAssassin email corpora, and [15] feature-engineered URL dataset utilized in this study are examples of frequently used resources. Recent reviews reveal persistent gaps: adversarial robustness against purposefully perturbed or LLM-generated inputs is rarely measured; reported accuracy figures are not always accompanied by precision/recall breakdowns required to evaluate performance on the minority (phishing) class; and many studies evaluate on a single, aging dataset without testing cross-dataset generalization [7]. The experimental design used below is directly motivated by these shortcomings.

E. Multichannel Social Engineering: Vishing, Smishing, and Voice Deepfakes

Although AI-enabled social engineering is becoming more multichannel, phishing research has traditionally focused on email and the web. As attackers shift away from highly filtered email channels, voice phishing, also known as vishing, has increased dramatically. The availability of high-quality voice-cloning tools, some of which only require a few seconds of reference audio, has made impersonating executives, family members, or support staff far more convincing than scripted cold calls. Rather than transferring text-based phishing techniques to audio transcripts, a first comprehensive study of spoken-language social-engineering detection frames vishing as an emergent, understudied research subject and advocates for dedicated detection pipelines [16]. This relates to a larger body of research on deepfake detection: benchmarking studies demonstrate that detectors trained on carefully selected academic datasets lose a significant portion of their accuracy when tested on real-world audio, video, and image deepfakes gathered from actual platforms, with reported AUC drops of approximately 45–50% [17]. This generalization gap is directly comparable to the dataset-staleness issue noted for phishing emails and URLs in Section 2.4, and it implies that detector evaluation protocols in the entire social-engineering space not just phishing text—need to shift toward regularly updated, real-world benchmarks.

Smishing, or SMS-based phishing, exhibits an analogous progression. Similar to the pattern observed for email and URL detection, comparative studies using logistic regression, random forest, SVM, and deep architectures (CNN, LSTM) on SMS corpora reveal that ensemble and hybrid deep-learning approaches once more perform better than single classical models [18]. In order to improve recall on the minority smishing class without inflating false positives, more recent work uses ensemble learning in conjunction with SMOTE-style oversampling to address the severe class imbalance typical of SMS datasets legitimate messages greatly outnumber malicious ones in real deployments [19]. Researchers have also modeled telephone-based social engineering as a multi-turn interaction, going beyond single-message classification. While broader frameworks like SEADer++ generalize social-engineering-attack detection across online chat and messaging environments using supervised learning over conversational features, scam-signature approaches extract recurrent linguistic patterns and conversational stages from real scam-baiting call transcripts to detect attacks as they unfold rather than after the fact [20]. The field is shifting from single-channel, single-message categorization to multichannel, conversation-aware detection, as this body of work collectively shows. This is a direction that the URL-only evaluation in this study does not try to cover but clearly identifies as a shortcoming.

F. Human Factors and Explainable Detection

In a different line of inquiry, the human recipient rather than only the model is considered a component of the detecting system. Regardless of the technical safeguards in place, certain users, particularly older adults, are disproportionately likely to fall for social engineering due to demographic, cognitive, and contextual factors such as time pressure, authority cues, and familiarity with the sender, according to systematic reviews of phishing susceptibility [21]. This has led to the development of explainable AI (XAI) techniques that do more than simply categorize messages; they also provide the user with a rationale for their choice. Participants' own detection accuracy increased from 0.712 to 0.928 after reading the explanations, according to a controlled user study of an SMS-phishing tool (SmishX) that displays AI-generated explanations alongside each verdict. The biggest improvements were seen among older adult participants, who are typically the most frequently targeted by smishing and vishing [22]. In order to expose highlighted features (such as mismatched domain age and suspect redirect

chains) to security analysts instead of treating them as an opaque score, complementary work uses SHAP-based feature attribution to phishing-website classifiers [23-25]. The practical value of a detector also depends on whether its output can be explained to the people it is intended to protect, an evaluation dimension that is largely absent from purely benchmark-driven studies like the current one. These findings suggest that model accuracy figures, such as those reported in Section IV, are necessary but insufficient on their own.

III. DATASET AND METHODOLOGY

A. Dataset

The phishing-URL dataset used in this study was made publicly available through an open GitHub repository and published by [15] in Data in Brief. It contains 88,647 labeled instances (58,000 legitimate, 30,647 phishing), each described by 111 engineered attributes organized into six feature groups: (i) URL-based lexical features (overall length and counts of characters such as dots, hyphens, and digits); (ii) domain-based features (hostname length, presence of an IP address in place of a domain name, and top-level-domain indicators); (iii) directory-based features (path depth and counts of slashes, dots, and special characters in the directory portion); (iv) file-based features (file-name length and extension-related indicators); (v) parameter-based features (query-string length and counts of parameters); and (vi) resolving/external features derived from WHOIS and DNS lookups, including time-to-live (TTL), domain activation (registration) time, and autonomous system number (ASN).

No additional feature-selection or dimensionality-reduction procedure was applied prior to model training: all 111 provided attributes were retained for every model, and the only pre-processing applied was the standardization. This dataset was chosen because it is structurally similar to the feature sets used in the reviewed literature [14], peer-reviewed, large enough for stable cross-validated estimates, and free of the class-leakage problems reported for some older toy datasets.

B. Experimental Setup

To maintain class balance, stratified sampling was used to divide the data into training (75%) and held-out test (25%) sets, with the test set set aside before any further processing and touched only for final evaluation. For models sensitive to feature scale (logistic regression, linear SVM, and the multilayer perceptron), a standardization transform (zero mean, unit variance) was fit exclusively on the training data and then applied to the held-out test set; raw feature values were used for tree-based models. To prevent information leakage during internal validation, the stratified five-fold cross-validation described below was performed entirely within the training set, and for scale-sensitive models the standardization scaler was re-fit on each training fold and applied to the corresponding validation fold, rather than being fit once on the full training set; the held-out test set played no part in this process. Logistic regression, linear support vector machine, random forest, gradient boosting, and a multilayer perceptron were the five models trained, representing the major families utilized in the literature. Hyperparameters were kept close to widely used, literature-informed defaults and adjusted through limited manual tuning guided by five-fold cross-validation performance on the training set only, rather than an exhaustive automated search, consistent with the comparative, baseline-oriented scope of this study [4, 9].

The main settings used were: logistic regression (L2 penalty, C = 1.0, lbfgs solver); linear SVM (L2 penalty, C = 1.0, hinge loss); random forest (200 trees, Gini impurity criterion, no maximum depth restriction, minimum samples per leaf = 1); gradient boosting (100 estimators, learning rate = 0.1, maximum tree depth = 3); and the multilayer perceptron (two hidden layers of 64 and 32 units, ReLU activation, Adam optimizer, default learning rate, trained for up to 200 iterations with early stopping on a validation split). Accuracy, precision, recall, F1-score, and ROC-AUC were used to evaluate the model's quality on the held-out test set.

C. Robustness Simulation

A feature-space evasion simulation was conducted on the top-performing model (random forest) to investigate robustness beyond aggregate accuracy, a gap mentioned in Section 2.4. The evaluation subset for this simulation was restricted, by design, to the test-set phishing URLs that the random forest already classified correctly at zero perturbation (i.e., true positives); this restriction is what defines the 0%-perturbation "baseline" detection rate reported in Table II as 100%, and this baseline should be read as a reference point internal to the simulation rather than as the model's overall recall on the full test set, which is reported separately in Table 1 (0.9593).

The five features with the highest random-forest importance (directory length, domain activation time, URL slash count, directory slash count, and URL length) were perturbed one URL at a time. For each selected URL and each target intensity $m \in \{10\%, 25\%, 50\%\}$, every one of the five features x was replaced with $x' = x \times (1 + \varepsilon)$, where ε was drawn independently from a uniform distribution on $[-m, m]$. Perturbed values were then post-processed to remain within realistic ranges for each feature: count-based and length-based features (e.g., slash counts, directory length, URL length) were rounded to the nearest non-negative integer, and domain activation time was clipped at zero, so that no perturbation produced a negative length, a negative count, or a negative age.

The perturbed samples were then re-classified by the unmodified, already-trained random forest model, and the evasion rate at each intensity was computed as the proportion of these previously-correctly-classified phishing URLs that were reclassified as legitimate. This procedure is different from the gradient-based or LLM-crafted hostile content and mimics a straightforward, non-adaptive evasion effort rather than an optimized attack.

IV. RESULTS

Table I reports held-out test performance for all five models.

TABLE I. Test-set performance of five classifiers on the phishing-URL benchmark dataset.

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	0.9320	0.8900	0.9167	0.9032	0.9791
Linear SVM	0.9302	0.8823	0.9210	0.9013	0.9789
Random Forest	0.9699	0.9539	0.9593	0.9566	0.9951
Gradient Boosting	0.9537	0.9311	0.9354	0.9333	0.9894
MLP (Neural Network)	0.9607	0.9399	0.9468	0.9433	0.9914

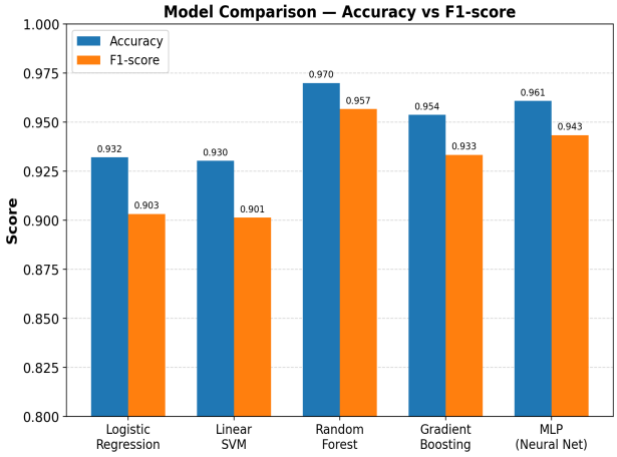

Fig. 1. Accuracy and F1-score comparison across the five evaluated models.

The random forest classifier outperformed the multilayer perceptron and gradient boosting in all metrics (accuracy = 0.970, F1 = 0.957, ROC-AUC = 0.995) as shown in Fig. 1. Despite having similar overall accuracy, the two linear models (logistic regression and linear SVM) fared considerably worse on recall and F1, indicating a higher number of missed phishing cases. This serves as a reminder—repeated throughout the literature—that accuracy alone can conceal weaker minority-class identification. The ranking was validated by five-fold cross-validation, and the random forest demonstrated stable, non-overfit performance with the greatest mean F1 (0.956) and the lowest variance between folds (SD = 0.0005) as shown in Fig. 2.

Feature-importance analysis (Fig. 3) shows that URL directory length, domain activation time, and path/slash-related structural features dominate the model's decisions, in line with prior findings that phishing infrastructure tends to use longer, more convoluted paths and recently registered or recently activated domains. Table II summarizes the robustness simulation.

Even at 50% random perturbation of its five most important features, the random forest model's detection rate on this already-correctly-classified subset dropped by less than one percentage point (from a 100% starting point, by definition, to 99.22%), indicating strong resilience to unstructured, non-adaptive noise on this feature set, which describes behavior on previously-detected phishing URLs and should not be read as the model's overall recall on the full test set.

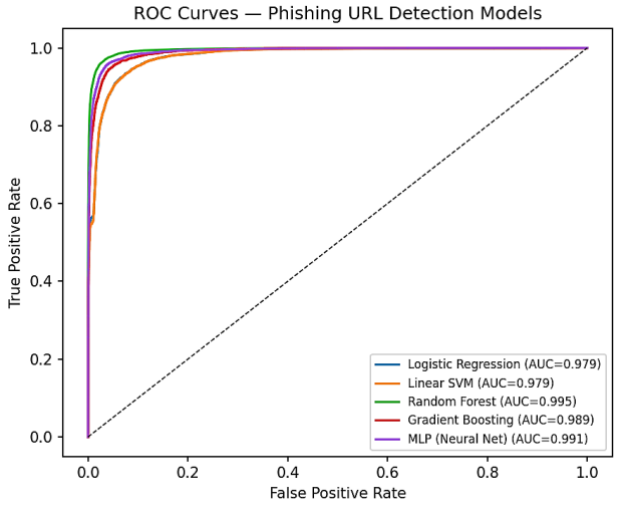

Fig. 2. ROC curves for all five classifiers on the held-out test set.

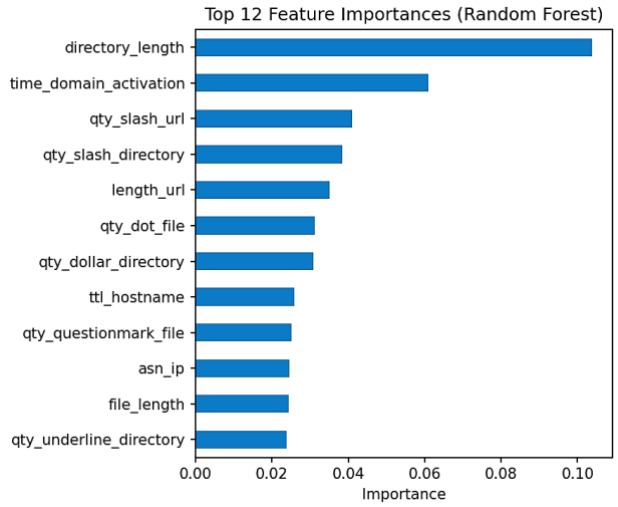

Fig. 3. Top twelve feature importances from the random forest model.

TABLE II. Evasion rate of correctly-detected phishing URLs under random perturbation of top features.

Perturbation magnitude	Evasion rate	Post-perturbation detection rate
0% (baseline)	0.00%	100.00%
10% random noise	0.22%	99.78%
25% random noise	0.52%	99.48%
50% random noise	0.78%	99.22%

V. DISCUSSION

The experimental findings support the larger body of literature: ensemble tree-based techniques continue to be the most effective general-purpose classifiers for structured phishing-URL data, and the particular characteristics they rely on path structure, domain age, and network metadata—align with the discriminative signals found in all of the reviewed studies. Although reported benchmark results on similarly engineered feature sets are consistent with the near-perfect ROC-AUC (0.995) obtained here, such figures should be

interpreted cautiously because strong performance on a single, static benchmark does not guarantee equivalent performance on live traffic, which is constantly reshaped by attackers. It is important to keep these two notions of performance distinct. Benchmark performance, as reported in Table I, measures how well a model separates phishing from legitimate URLs drawn from the same fixed, historical distribution used for training and testing; it is informative about relative model quality under controlled, like-for-like conditions but says nothing about calibration drift, concept drift, or exposure to attack patterns invented after the dataset was collected. Real-world deployment performance, by contrast, depends on the live traffic distribution a system actually encounters, including phishing infrastructure and URL-obfuscation techniques that postdate this benchmark, feedback effects from attackers adapting to a deployed filter, and integration factors such as latency and false-positive tolerance in an operational pipeline. A model that tops Table I is not thereby guaranteed to retain the same ranking, or even the same absolute performance, once deployed, that move beyond a single static benchmark.

The robustness simulation highlights a crucial detail that is simple to exaggerate in either way. It is not appropriate to interpret the model's resistance to random feature noise as resistance to actual adversarial pressure. The risks outlined in gradient-guided feature modification and, more importantly, social-engineering content created by LLMs with the express purpose of eliminating the lexical and structural cues that a detector relies on—are only weakly represented by random perturbation. Previous research demonstrates that GPT-4o-generated phishing emails can already evade commercial filters by removing the very artifacts (grammar errors, awkward phrasing) that many detectors were trained to identify, and that specific stylometric or provenance-based features rather than the URL-structure features examined here are increasingly needed to distinguish AI-generated from human-written text. This highlights a structural flaw in both the current study and a large portion of the URL-only literature: URL-based models do not account for the persuasiveness or fluency of an accompanying phishing email or chat message, which is exactly where LLM-driven social engineering is developing at the fastest rate.

This has practical ramifications. Businesses that only use URL- or metadata-based filters should combine them with content-level, provenance-aware, or LLM-assisted reasoning detectors of the type covered in Section 2.3 to treat them as one layer of a defense-in-depth approach rather than a comprehensive solution. To stay up with generative attack tools, future detection systems will probably need to integrate structural features (as employed here), content/stylometric signals, and behavioral or contextual signals (sender history, organizational ties).

A. Limitations

There are a number of limitations to this study. Conclusions about content-level detection are derived from the literature rather than from original experiments because the dataset, albeit big and peer-reviewed, does not include LLM-generated email or message content and instead represents phishing infrastructure tendencies as of its publication. The provided robustness numbers indicate a lower bound on the threat rather than an upper bound, because the adversarial simulation uses random, non-adaptive perturbation instead of an optimized, gradient-based, or LLM-crafted attack. Lastly, instead of evaluating the entire spectrum of transformer-based and graph-based architectures, the study assesses five widely used classical and ensemble model families within the scope of a focused comparative evaluation.

VI. CONCLUSION

This work integrated an original empirical assessment on a sizable public benchmark dataset with a targeted study of recent (2023–2026) research on AI for phishing and social engineering detection. Among the five models tested, a random forest classifier produced the best results (accuracy 0.970, F1 0.957, ROC-AUC 0.995), which is in line with the ensemble-dominance trend documented in the literature and demonstrated resilience to random feature perturbation. However, the review and discussion show that resilience against the generative, content-level social-engineering techniques now made possible by big language models is not the same as structural, URL-based robustness. The primary open problem addressed by this work and the larger literature it relies on is closing that gap using multimodal, content-aware, and adversarial-tested detection systems.

FUNDING STATEMENT

The author(s) received no specific funding for this study.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest to report regarding the present study.

AUTHOR CONTRIBUTIONS

Conceptualization, M.A.F, M.H.H., and U.A.M.; methodology, M.A.F, M.H.H., M.D and U.A.M.; software, M.A.F, M.H.H., M.H.D. and M.D.Q.; validation, M.A.F, M.D.Q., M.H.H., M.E.Q., and M.D.B.; writing—original draft preparation, M.A.F, M.D.B., M.E.Q. and U.A.M.; writing—review and editing, M.H.D., M.D.B., and M.E.Q.

INSTITUTIONAL REVIEW BOARD STATEMENT

Not applicable.

INFORMED CONSENT STATEMENT

Not applicable.

DATA AVAILABILITY STATEMENT

Data is available on reasonable request.

REFERENCES

[1] Anti-Phishing Working Group. Phishing activity trends report. APWG, 2025.

[2] Verizon. Data breach investigations report. Verizon Business. 2025.

[3] Zieni, R., Massari, L., & Calzarossa, M. C. “Phishing or not phishing? A survey on the detection of phishing websites.” *IEEE Access*, vol. 11, p. 18499-18519, 2023.

[4] Sharma, A., & Gupta, N. “A comprehensive survey on machine learning-based phishing detection.” *ACM Computing Surveys*, vol. 56, no. 4, p.1-28, 2024.

[5] Heiding, F., Schneier, B., Vishwanath, A., Bernstein, J., & Park, P. S. “Devising and detecting phishing emails using large language models.” *IEEE Access*, vol. 12, p. 42131-42146, 2024.

[6] Afane, K., Wei, W., Mao, Y., Farooq, J., & Chen, J. “Next-generation phishing: How LLM agents empower cyber attackers.” *In 2024 IEEE International Conference on Big Data (BigData), 2024,* (pp. 2558-2567). IEEE.

[7] Koide, T., Fukushi, N., Nakano, H., & Chiba, D. “Detecting phishing sites using ChatGPT.” 2023. arXiv preprint arXiv:2306.05816.

[8] Zhang, J., Wu, P., London, J., & Tenney, D. “Benchmarking and evaluating large language models in phishing detection for small and midsize enterprises: A comprehensive analysis.” *IEEE Access*, vol. 13, p. 28335-28352, 2025.

[9] Kytidou E, Tsikriki T, Drosatos G, Rantos K. “Machine learning techniques for phishing detection: A review of methods, challenges, and future directions.” *Intelligent Decision Technologies*, vol. 19, no. 6, p. 4356-79, 2025.

[10] Omari, K. “Comparative study of machine learning algorithms for phishing website detection.” *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 9, 2023.

[11] N. Fawad, M. N. Khan, A. Addas, Q. Awais, and N. Ayub. "Game mechanics and artificial intelligence personalization: A framework for adaptive learning systems." *Education Sciences* 15, no. 3, p. 301, 2025

[12] Alshingiti, Z., Alaqel, R., Al-Muhtadi, J., Ul Haq, Q. E., Saleem, K., & Faheem, M. H. “A deep learning-based phishing detection system using CNN, LSTM, and LSTM-CNN.” *Electronics*, vol. 12, no. 1, p. 232, 2023.

[13] Tamal, M. A., Islam, M. K., Bhuiyan, T., Sattar, A., & Prince, N. U. “Unveiling suspicious phishing attacks: Enhancing detection with an optimal feature vectorization algorithm and supervised machine learning.” *Frontiers in Computer Science*, vol. 6, p, 1428013, 2024.

[14] Yoon, J. H., Buu, S. J., & Kim, H. J. “Phishing webpage detection via multi-modal integration of HTML DOM graphs and URL features based on graph convolutional and transformer networks.” *Electronics*, vol.13, no.17, p.3344, 2024.

[15] Mohammad, R., & McCluskey, L. Phishing websites [Dataset]. UCI Machine Learning Repository, 2015. https://doi.org/10.24432/C51W2X

[16] Vrbančič, G., Fister, I., & Podgorelec, V. “Datasets for phishing websites detection.” *Data in Brief*, vol.33, p.106438, 2020.

[17] Triantafyllopoulos, A., Spiesberger, A. A., Tsangko, I., Jing, X., Distler, V., Dietz, F., Alt, F., & Schuller, B. W. Vishing: Detecting social engineering in spoken communication — a first survey & urgent roadmap to address an emerging societal challenge. Computer Speech & Language, 94(C), 2025.

[18] Layton, S., Tucker, T., Olszewski, D., Warren, K., Butler, K., & Traynor, P. “SoK: The good, the bad, and the unbalanced: Measuring structural limitations of deepfake media datasets.” *In Proceedings of the 33rd USENIX Security Symposium*, 2024.

[19] Prajapati, P., Bhagat, D., & Shah, P. “Smishing: A SMS phishing detection using various machine learning algorithms.” *In Communication and Intelligent Systems (ICCIS 2023), Lecture Notes in Networks and Systems,* vol. 968, 2024.

[20] Xu, H., Qadir, A., & Sadiq, S. “Malicious SMS detection using ensemble learning and SMOTE to improve mobile cybersecurity.” *Computers & Security*, vol.154, p.104443, 2025.

[21] Derakhshan, A., Harris, I. G., & Behzadi, M. “Detecting telephone-based social engineering attacks using scam signatures.” *In Proceedings of the 2021 ACM Workshop on Security and Privacy Analytics*, 2021, (pp. 67-73).

[22] Desolda, G., Ferro, L. S., Marrella, A., Catarci, T., & Costabile, M. F. “Human factors in phishing attacks: A systematic literature review.” *ACM Computing Surveys*, vol.54,no.8, p.1-35, 2021.

[23] Wang Y, Zhai H, Wang C, Hao Q, Cohen NA, Foulger R, Handler JA, Wang G.. “Can you walk me through it? Explainable SMS phishing detection for improved user comprehension and trust.” *In USENIX Symposium on Usable Privacy and Security (SOUPS 2025)*, 2025.

[24] S.R. Babary, and R. Khalid. "Leveraging Transformer-Based Natural Language Processing for Adaptive Language Learning in Digital Humanities Environments." *Journal of Artificial Intelligence and Computing* 4, no. 1, p. 36-43, 2026.

[25] Alsuqayh, N., Mirza, A., & Alhogail, A. “A phishing website detection system based on hybrid feature engineering with SHAP explainable artificial intelligence technique.” *In Web Information Systems Engineering — WISE 2024 PhD Symposium, Demos and Workshops*, 2025, (pp. 3-17).