Journal of Information Technology and Applications Research
Volume 1 Nomor 2 November 2025
e-ISSN: 3089-8013; p-ISSN:3089-8021, Page 62-77
DOI: https://doi.org/10.63956/jitar.v1i2.39
Available online at: https://jitar.org/index.php/jitar/article/view/39

NLP and Deep Learning for Phishing and Social Engineering Detection: A Systematic Review (2018–2026)

Ovan Sunarto Pulu[1*], Muhammad Fadly[2]
Informatic System, Gunadarma University, Depok, Indonesia

Abstract
Phishing and social engineering continue to escalate as digital public services and online commerce expand, with attackers exploiting linguistic deception, impersonation cues, and routine "click-and-comply" behavior across email, SMS, and voice channels. Objective: This study aims to systematically synthesize research on phishing and social engineering detection using natural language processing (NLP) and deep learning (2018–2026) to address fragmented evidence across channels and inconsistent terminology that limits robust comparison and practical translation. Method: A systematic literature review was conducted through structured database searches and snowballing, followed by deduplication, staged screening, and eligibility assessment. Studies were analyzed using a standardized extraction form, then synthesized via descriptive mapping and thematic analysis to develop a method taxonomy and examine evaluation rigor and operational readiness. Findings: The evidence base is dominated by email/BEC detection, while smishing and vishing remain comparatively underrepresented. Methods increasingly rely on contextual language representations and hybrid architectures to capture semantic and local deception patterns; however, evaluation practices are heterogeneous and often provide limited evidence on cross-dataset generalization, temporal robustness, and deploy ability. Socio-technical findings also indicate that human susceptibility and system/client workflow vulnerabilities can moderate the real-world effectiveness of technical defenses. Implications: The proposed taxonomy supports method selection by channel and highlights actionable priorities for practice and policy, including standardized reporting, cross-dataset and temporal validation, robustness testing, and integration with operational security workflows. Originality: This review adds value by consolidating detection and deception-centric strands through explicit inclusion of impersonation, fraud email, and scam terminology, and by linking methodological choices to evaluation rigor and deployment constraints across email, SMS, and voice contexts.
Keywords: phishing, social engineering, natural language processing, deep learning, transformer, systematic literature review

INTRODUCTION

Phishing and social engineering have escalated with the growth of digital public services and online commerce because attackers can convincingly impersonate trusted institutions and leverage routine "click-and-comply" behavior. A prominent tactic is domain and interface mimicry, where counterfeit URLs and official-looking pages differ only by subtle character changes, increasing the likelihood that users unknowingly provide credentials, identity information, or financial data. As a result, the consequences extend beyond individual losses to institutional credibility and public confidence in

Received: September 15, 2025; Accepted: October 22, 2025; Published: November 19, 2025
*Corresponding author : ovan@staff.gunadarma.ac.id

digital transactions and official communications, making phishing simultaneously a technical challenge and a socio-policy concern(Perbendaharaan, 2025).

This urgency is reinforced by both national and global indicators. In Indonesia, the Indonesia Anti-Phishing Data Exchange recorded 34,622 phishing reports over a five-year period and 7,988 reports in Q3 2022, with government institutions identified as a major target sector (PANDI, 2022). Such exposure contributes to declining user trust in digital services and online transactions, which can weaken broader participation in the digital economy (Mahmud & Wirawan, 2024). Compounding the problem, traditional safety cues are increasingly misleading; phishing domains may use HTTPS, prompting non-expert users to assume legitimacy (PANDI, 2022). Consistent with this, a global industry compilation updated in early 2026 estimates that approximately 1.2% of all emails sent are malicious, indicating pervasive exposure through everyday communication channels (Palatty, 2026).

In response, research on phishing, deception, and fraud emails has advanced through NLP and deep learning methods designed to capture linguistic and contextual signals of malicious intent. Within email-centric detection, studies increasingly employ hybrid architectures that combine semantic representations with local pattern extraction; for example, BERT-based hybrids augmented by sequential and convolutional components have been proposed for Business Email Compromise detection and report strong results across multiple datasets (Alguliyev et al., 2024). Related work evaluates diverse deep learning families including sequence models and Transformer-based classifiers often reporting high accuracy under controlled conditions (Pimpason et al., 2025). Evidence also shows that modeling choices matter: earlier pipelines based on one-hot encoding highlight the impact of feature construction (Bagui et al., 2021), while more recent approaches integrate NLP preprocessing with CNN-based text models to improve precision and reliability (Hilani et al., 2025). Nevertheless, many studies remain concentrated on single-dataset evaluations, limiting what can be concluded about transferability when attacker language and data distributions shift.

The literature further expands toward cross-channel social engineering, where detection must adapt to SMS and voice settings with distinct signal properties. Smishing studies commonly use NLP feature extraction from short-form messages paired with

machine learning classifiers, and some emphasize operational integration by connecting detection outputs to threat-intelligence ecosystems for sharing and (Karhani et al., 2023). In parallel, vishing research applies NLP and machine learning to voice-derived representations to identify malicious calls (Phang et al., 2024), while other work treats phishing as part of broader social engineering and proposes deep learning models for phishing attack detection grounded in linguistic features (Vidyasri & Suresh, 2025). Compared with email-focused research, however, cross-channel studies are frequently constrained by limited public datasets, inconsistent benchmarking across modalities, and incomplete reporting on robustness to language variation, accent diversity, and attacker script adaptation factors that shape real-world performance.

Alongside channel-focused research, an increasingly influential stream examines impersonation, deception mechanisms, and human susceptibility, which justifies expanding search coverage to include "impersonation," "deception," "fraud email," and "scam." Experimental findings indicate that susceptibility can fluctuate within the same individuals across repeated exposures, challenging assumptions that static defenses or one-time awareness interventions remain effective over time (Sommestad & Karlzén, 2024). Behavioral evidence also suggests co-occurring vulnerability across phishing emails, scam texts, and deceptive headlines, with digital literacy and cognitive reflectiveness acting as meaningful predictors (Sarno & Black, 2024)Persuasion-oriented analyses provide theory-grounded influence principles frequently exploited in social engineering, such as authority and urgency, offering constructs that could inform feature design, labeling, and explainability (Ferreira et al., 2015). Yet, these constructs are not consistently operationalized within NLP pipelines, leaving a gap between deception theory and detection model development.

Finally, system-level insights underscore that model improvements alone do not guarantee robust protection in practice. A synthesis of email deception research over the past decade reports that many modern email clients remain susceptible to phishing techniques, emphasizing the role of interface and workflow vulnerabilities in shaping outcomes beyond classifier metrics (Veit et al., 2025). Taken together, the literature converges on four gaps that motivate a comprehensive SLR and method taxonomy: fragmentation across channels and terminology that risks omitting deception- and scam-

centric studies; insufficient evidence on cross-dataset and temporal robustness under concept drift and attacker adaptation; limited integration of persuasion and deception theory into model design; and uneven attention to operational readiness, including system integration, deployment constraints, and user-facing defenses. Addressing these gaps requires an SLR that unifies evidence across email, SMS, and voice contexts, incorporates deception-centric terminology, and synthesizes best practices for evaluation and deployment in realistic threat environments.

METHODS

This study examines peer-reviewed scientific publications as the primary unit of analysis, focusing on journal articles and conference papers published between 2018 and 2026 that investigate phishing and social engineering detection using natural language processing and deep learning. The scope covers multiple attack channels email phishing (including Business Email Compromise), smishing (SMS), and vishing (voice) and explicitly incorporates deception-oriented phenomena such as impersonation, fraud emails, scams, and related deceptive messaging. Each eligible paper is treated as an empirical research artifact from which comparable methodological attributes are extracted, including the attack channel and data modality, linguistic or deception cues, feature and representation strategy, model family and architecture, evaluation design, and implementation or deployment considerations.

A systematic literature review design is adopted because the study aims to consolidate fragmented evidence across channels and terminologies, and to synthesize methodological patterns rather than estimate a single pooled statistical effect. This design is appropriate for producing a transparent, replicable, and auditable pathway from study identification to synthesis, enabling the construction of a method taxonomy that links data, representation, and model choices to evaluation practices and operational constraints. Given the rapid evolution of adversarial tactics and language-driven manipulation, a systematic approach is also necessary to surface where reported performance is supported by robust validation and where claims are limited by narrow experimental setups.

The study relies on secondary data retrieved from established bibliographic databases and digital libraries that index research at the intersection of cybersecurity, NLP, and machine learning. Searches are conducted primarily in Scopus, Web of Science, IEEE Xplore, and the ACM Digital Library, with backward and forward snowballing applied to influential studies to reduce the risk of omitting relevant work, particularly research framed under deception, impersonation, or fraud rather than phishing alone. This source strategy is intended to balance coverage and quality by prioritizing peer-reviewed venues while still capturing closely related strands that may be dispersed across adjacent communities.

Data collection follows a structured identification and screening workflow. Boolean queries are executed within titles, abstracts, and keywords by combining attack-channel terms (e.g., phishing, smishing, vishing, BEC) with method terms (e.g., NLP, deep learning, Transformer, BERT, CNN, LSTM) and deception-centric terms (e.g., impersonation, deception, fraud email, scam). Retrieved records are deduplicated, screened at the title–abstract level using predefined inclusion and exclusion criteria, and then assessed through full-text review to confirm topical relevance and the presence of sufficient methodological and evaluation detail. For each included study, a standardized extraction form is applied to capture publication metadata, dataset characteristics and language coverage, preprocessing steps, feature and representation choices, model architecture, validation protocol and metrics, robustness considerations, and any discussion of deployment constraints or integration with operational security workflows.

Analysis proceeds through an integrated synthesis that combines descriptive mapping and thematic comparison. First, the literature is summarized descriptively to identify trends by publication year, channel, data modality, model family, and dataset properties. Second, thematic coding is used to group studies into methodological clusters that form a taxonomy spanning deception cue modeling, representation strategy, architectural design, and evaluation rigor. Third, cross-study comparisons are performed to identify recurring limitations and evidence gaps, with particular attention to cross-dataset transferability, temporal validation under concept drift, robustness to attacker adaptation and adversarial manipulation, the degree to which deception and persuasion

constructs are operationalized in model design, and the extent to which studies address operational readiness through deployment constraints and ecosystem integration.

RESULTS AND DISCUSSION

Landscape of Studies Across Attack Channels and Data Modalities

The in-scope corpus (2018–2026) contains N = 17 peer-reviewed studies that address phishing and social engineering from detection, behavioral, and system-level perspectives. As shown in Table 1, the publication landscape begins to appear in 2021 and increases sharply in 2024–2025, indicating accelerating research attention in recent years. When grouped by the primary threat channel or context, the evidence base is most developed for email/BEC (n = 7), while smishing/SMS (n = 1) and vishing/voice (n = 1) remain underrepresented; a substantial portion of the corpus (n = 8) focuses on broader deception phenomena (e.g., scams, impersonation, susceptibility, or system-level email-client risk) rather than channel-specific detection. This imbalance suggests that conclusions about model performance and readiness are currently most defensible for email-centric settings, whereas cross-channel generalization remains limited by the scarcity of comparable datasets and standardized evaluation designs.

Table 1. Distribution of included studies by year and channel/context (in-scope provided references, 2018–2026; N = 17).

Year	Email/BEC	Smishing/SMS	Vishing/Voice	Mixed/Other	Total
2018	0	0	0	0	0
2019	0	0	0	0	0
2020	0	0	0	0	0
2021	1	0	0	1	2
2022	0	0	0	1	1
2023	1	1	0	0	2
2024	1	0	1	2	4
2025	4	0	0	4	8
2026	0	0	0	0	0
Total	**7**	**1**	**1**	**8**	**17**

The overall distribution by channel is summarized in Figure 3, while the year-by-year trend is shown in Figure 2, both indicating a concentration of recent work in email/BEC and mixed socio-technical strands. Within channel-specific detection studies,

the most common input is text (email body/subject or SMS text), whereas vishing work relies on voice-derived representations (e.g., speech features and/or ASR transcripts), reflecting different data constraints and methodological choices across channels (Phang et al., 2024).

Method Taxonomy of NLP and Deep Learning Approaches

Methodologically, the corpus shows clear clustering of approaches around three recurring design choices: the form of linguistic representation, the model family, and whether auxiliary signals are fused with text. For email/BEC, studies increasingly adopt contextual language modeling and hybrid designs to capture both semantic and local textual patterns; for example, a BERT-based hybrid architecture combining sequential and convolutional components is proposed for BEC detection and evaluated across multiple datasets (Alguliyev et al., 2024). Alongside these hybrids, CNN-based pipelines with explicit NLP preprocessing remain prominent for phishing email detection (Hilani et al., 2025), while other work incorporates linguistic structure via NER and contextual models (with mention of GPT-4) to capture phishing-specific language patterns (Gupta et al., 2025). Earlier evidence also indicates that representation strategies such as one-hot encoding can materially affect classification outcomes, highlighting the importance of consistent and transparent preprocessing in comparisons (Bagui et al., 2021).

The cross-channel detection literature is comparatively thinner but demonstrates two notable directions. First, smishing detection often employs machine learning classifiers over NLP-derived features and may be explicitly integrated with operational cybersecurity tooling such as threat-intelligence sharing (Karhani et al., 2023). Second, vishing detection frameworks use NLP and machine learning on voice-derived representations, extending beyond text-only defenses but facing higher data and benchmarking constraints (Phang et al., 2024). In parallel, social-engineering–oriented phishing detection also appears in the form of autoencoder-based deep learning that integrates linguistic features into broader deception settings (Vidyasri & Suresh, 2025). To consolidate these patterns into an actionable taxonomy, Table 2 provides a structured mapping from channel/context to representation and model family, enabling consistent cross-study comparison.

Table 2. Included in-scope studies and methodological classification (N = 17).

Study	Year	Channel/ Context	Study type	Primary data	Representation/ NLP	Model family / approach	Notable emphasis
Alguliyev et al.	2024	Email/BEC	Detection	Text	Contextual embedding (BERT)	Hybrid (BERT + BiGRU + CNN)	Multi-dataset evaluation reported
Pimpason et al.	2025	Email	Detection	Text	Not specified	Deep learning classifier	Metrics reported; split not specified
Bagui et al.	2021	Email	Detection	Text	One-hot encoding	ML/DL comparison	Representation affects outcomes
Hilani et al.	2025	Email	Detection	Text	NLP preprocessing (not detailed)	1D-CNN	Precision/accuracy emphasized
Gupta et al.	2025	Email	Detection	Text	NER + contextual modeling; mentions GPT-4	Transformer/ LLM-assisted	Linguistic pattern modeling
Karhani et al.	2023	Smishing/SMS (and phishing)	Detection	Text	NLP features (not detailed)	ML classifier + MISP integration	Operational TI integration
Phang et al.	2024	Vishing/Voice	Detection	Voice-derived	Not specified	NLP + ML framework	Voice phishing defense pipeline
Vidyasri & Suresh	2025	Social engineering phishing	Detection	Text	NLP features (not detailed)	Autoencoder-based DL (FDN-SA)	SE framing of phishing
De Queiroz	2025	Cross-channel	Review/ Conceptual	N/A	N/A	LLM prevention & risk discussion	Adversarial and ethics highlighted
Rajeswari & Rajeeth Prabhu	2025	General	Review	N/A	N/A	Foundational AI review	Maps AI in phishing detection
Sommestad & Karlzén	2024	General	Human factors	N/A	N/A	Repeated measures experiment	Susceptibility variability
Sarno & Black	2024	Email + scam texts + fake news	Human factors	N/A	N/A	Susceptibility predictors	Literacy/ reflectiveness
Topor & Pollack	2022	General	Conceptual/ Analysis	N/A	N/A	Fake identity spectrum	Impersonation framing
Wang et al.	2021	Romance scam	Intervention	N/A	N/A	Deterrence messaging experiment	Mitigation via warning messages
Iwara	2025	Bank impersonation scams	Empirical	N/A	N/A	Vulnerability & mitigation survey	Mitigative strategies
Bera et al.	2023	Fraudulent emails	Framework	Text	Thematic dimensions	Dimensional framework	Links tactics to intentions
Veit et al.	2025	Email ecosystem	SoK	N/A	N/A	Email client susceptibility synthesis	System-level vulnerability

Evaluation Practices, Robustness, and Deployment Readiness

Across channel-specific detection studies, performance is frequently emphasized, yet the supporting evidence varies in evaluation rigor and completeness of reporting. Multi-dataset evaluation is explicitly documented in BEC detection work using hybrid contextual and local feature learning (Alguliyev et al., 2024), providing stronger support for transferability claims than single-dataset setups. However, for several detection papers, key details such as split strategy (holdout vs cross-validation vs temporal split) and cross-dataset generalization are not fully visible from summary-level reporting, reinforcing the need for systematic extraction and coding at full-text stage to ensure comparability. From an operational perspective, the corpus demonstrates that deployment readiness is addressed unevenly: integration with threat-intelligence sharing and response workflows is explicitly foregrounded in smishing detection (Karhani et al., 2023), while system-level synthesis indicates that email clients can remain vulnerable to deception techniques even when detection models improve, underscoring the importance of considering interface and workflow constraints alongside classifier metrics (Veit et al., 2025).

To make the evidence gaps transparent, Table 3 summarizes the presence of robustness and operational indicators across the corpus, showing that cross-dataset and temporal validation are not consistently reported and that socio-technical factors (human susceptibility and client-level weaknesses) remain central to interpreting real-world effectiveness. In practice, these findings imply that reported high accuracy should be interpreted cautiously unless accompanied by cross-dataset or temporal evaluation, explicit robustness testing, and a clear pathway to integration with real-world security ecosystems.

Table 3. Evidence map of evaluation rigor, robustness, and operational readiness (coded from titles/abstract-level information; N = 17).

Dimension	In-scope evidence (count)	Representative studies	Implication for synthesis
Multi-dataset evaluation explicitly reported	1	Alguliyev et al. (2024)	Stronger basis for generalization claims

Dimension	In-scope evidence (count)	Representative studies	Implication for synthesis
Operational ecosystem integration (TI/SOC tooling)	1	Karhani et al. (2023)	Improves deployability relevance
System/client susceptibility analyzed	1	Veit et al. (2025)	Field risk depends on UI/workflow factors
Human susceptibility tested empirically	2	Sommestad & Karlzén (2024); Sarno & Black (2024)	Supports socio-technical interpretation
Deception tactics/intent frameworking	1	Bera et al. (2023)	Enables deception-centric taxonomy building
Explicit adversarial/LLM risk discussion	1	De Queiroz (2025)	Highlights new robustness concerns

To capture how scholarly attention has evolved over the review period, publication counts were grouped by the primary attack channel (email/BEC, smishing/SMS, vishing/voice, and mixed/general contexts). Figure 2 visualizes the year-by-year trajectory of included studies, allowing readers to identify when research activity began to accelerate and which channels have driven growth.

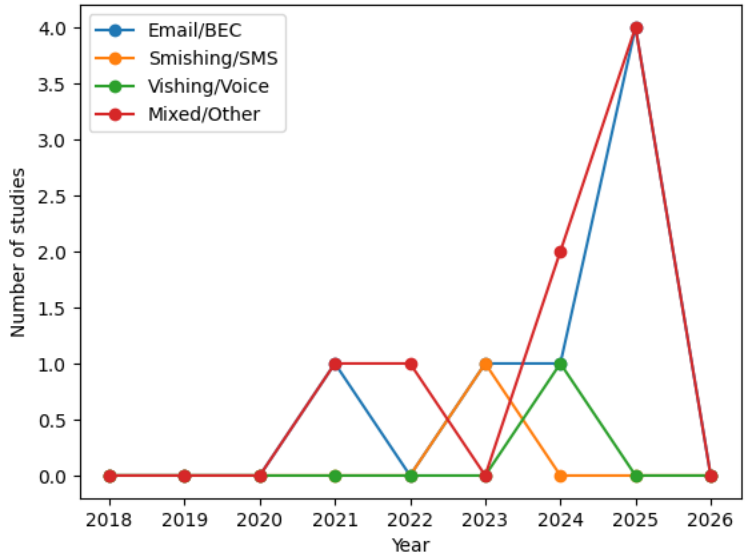

Figure 2. Publication trend by attack channel (in-scope provided references, 2018–2026).

Figure 2 indicates a pronounced rise in 2024–2025, with the largest concentration in email/BEC and mixed/general strands, whereas SMS- and voice-oriented studies remain comparatively scarce. This pattern suggests that methodological maturity and available evidence are currently strongest for email-centric detection, while cross-channel work would benefit from broader datasets and more consistent benchmarking to support robust comparisons.

Beyond temporal dynamics, it is also important to assess the overall balance of evidence across channels. Accordingly, Figure 3 summarizes the aggregate distribution of studies by channel/context within the corpus, providing a concise view of where the literature is most developed and where empirical coverage remains limited.

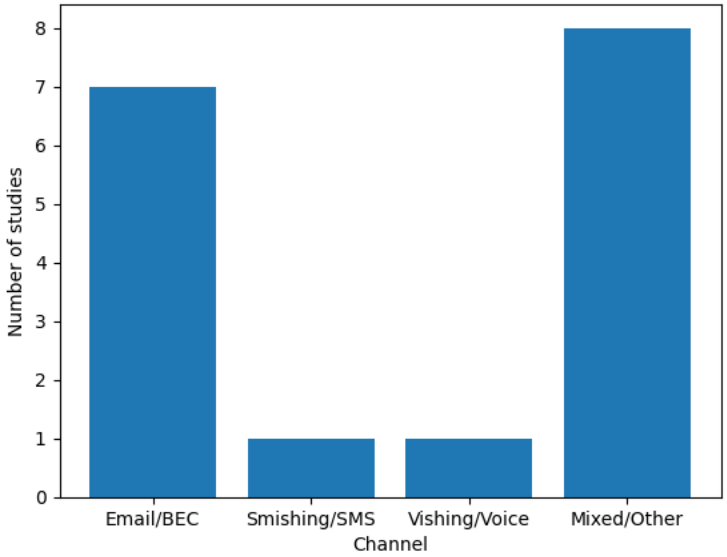

Figure 3. Overall distribution by channel/context (in-scope provided references, N = 17).

As shown in Figure 3, the corpus is dominated by email/BEC and mixed/general studies, with only minimal representation of smishing and vishing. This imbalance matters for interpretation: conclusions about model performance and practical readiness are more defensible in the email domain, whereas SMS and voice settings require a larger evidence base and standardized evaluation practices to strengthen external validity and deployment relevance.

DISCUSSION

The review synthesizes 17 in-scope studies (2018–2026) on phishing and social engineering in order to map the channel landscape, organize NLP/deep learning approaches into a coherent methodological taxonomy, and assess evaluation rigor and operational readiness. The results indicate a clear concentration of evidence in **email/BEC detection**, while **smishing** and **vishing** remain comparatively underrepresented. Methodologically, the email/BEC stream is characterized by increasingly sophisticated language representations and hybrid architectures that combine contextual semantics with local pattern extraction (Alguliyev et al., 2024), alongside CNN-based pipelines supported by NLP preprocessing (Hilani et al., 2025) and work leveraging linguistic structure such as NER and contextual encoders (Gupta et al., 2025). In parallel, a sizeable portion of the corpus emphasizes socio-technical dimensions human susceptibility and deception mechanisms rather than narrow classifier design (Sarno & Black, 2024; Sommestad & Karlzén, 2024) and system-level work highlights persistent ecosystem vulnerabilities in email clients despite advances in detection research (Veit et al., 2025).

These patterns can be explained by the interaction of data availability, standardization, and deployment incentives across channels. Email phishing has long generated large-scale textual artifacts and relatively accessible corpora, enabling rapid iteration on representation learning and model benchmarking, whereas SMS and voice modalities face greater barriers in collecting representative datasets, obtaining labels, and ensuring privacy-preserving sharing. Therefore, methodological “maturity” accumulates where data pipelines are easiest to build and reproduce, which helps explain why hybrid deep models and Transformer-centric approaches are more visible in email/BEC than in vishing/voice settings. At the same time, evidence from behavioral studies suggests that the effectiveness of technical defenses is mediated by human factors: susceptibility can fluctuate even within the same individuals across repeated exposures (Sommestad & Karlzén, 2024) and vulnerability to phishing emails can co-occur with susceptibility to scam texts and deceptive headlines, shaped by digital literacy and cognitive reflectiveness (Sarno & Black, 2024). This alignment between technical and behavioral findings supports the interpretation that phishing is not only a text-classification task but a socio-

technical problem whose risk profile is shaped by both attacker adaptation and user decision-making.

When compared with prior detection-centric work, the reviewed studies collectively confirm that modern NLP and deep learning can capture discriminative signals in malicious communication, particularly in email settings. Hybrid modeling that fuses contextual semantics and local features (Alguliyev et al., 2024) and deep text models integrated with preprocessing pipelines (Hilani et al., 2025) are consistent with a broader trend toward representation-rich detection. However, the review also surfaces a recurrent limitation: reported performance is often difficult to translate across datasets, organizations, and evolving attack strategies because evaluation details and generalization tests are inconsistently emphasized in summaries and frequently appear strongest only in select studies. The novelty of this SLR lies in consolidating evidence across **channels and deception-centric terminology** including "impersonation," "deception," "fraud email," and "scam" and in treating **evaluation rigor, robustness, and operational readiness** as core synthesis dimensions rather than peripheral considerations. This perspective complements tactic- and intention-oriented framing of fraudulent email attacks (Bera et al., 2023) and system-level analyses of client susceptibility (Veit et al., 2025) by connecting model design choices to the realities of deployment.

Beyond methodological implications, the findings carry broader social and institutional meaning. The dominance of email-focused detection research reflects where digital trust is most visibly contested in everyday organizational life, particularly in contexts involving credential capture, financial redirection, and impersonation. Evidence that susceptibility is variable and cross-deception vulnerabilities co-occur (Sarno & Black, 2024; Sommestad & Karlzén, 2024) implies that interventions should be conceptualized as continuous risk management rather than one-off training or static filtering. Moreover, the emergence of LLM-oriented discussion in prevention and risk framing (De Queiroz, 2025) suggests that the socio-technical landscape is shifting: language models may support defense (e.g., summarizing or flagging suspicious intent) while simultaneously expanding the attacker's capacity to generate persuasive, adaptive,

and multilingual lures. This dual-use tension makes governance, transparency, and robustness considerations increasingly central to both research and practice.

The review also highlights both functions and dysfunctions of current research trajectories. On the positive side, advanced NLP/DL methods provide strong candidate mechanisms for detecting subtle linguistic cues, supporting automated triage and reducing manual burden in high-volume channels such as email. On the negative side, the field risks over-indexing on leaderboard-style accuracy without sufficient attention to false positives, concept drift, and user-facing consequences such as alert fatigue, inequitable performance across languages or communities, and privacy constraints in data sharing. System-level evidence that email clients remain susceptible to deception techniques (Veit et al., 2025) further underscores a practical dysfunction: even strong models may be undermined by interface design, workflow friction, and attacker manipulation of user attention. In addition, as deception tactics diversify across channels, a narrow "phishing-only" framing can obscure relevant insights from impersonation and scam research (Iwara, 2025; Topor & Pollack, 2022) limiting the completeness of defenses.

These insights point to concrete action priorities for research and policy. At the research level, future studies should standardize reporting of validation protocols, explicitly include cross-dataset and, where feasible, temporal evaluation to reflect drift, and document dataset characteristics (language coverage, imbalance, collection context) to support reproducibility and fair comparison. Cross-channel progress will require privacy-aware, representative datasets and shared benchmarks for SMS and voice modalities, alongside clearer operational definitions of deception constructs that can be encoded as labels or features, building on tactic/intent frameworks (Bera et al., 2023) and incorporating human susceptibility insights (Sarno & Black, 2024; Sommestad & Karlzén, 2024). At the organizational and policy level, technical controls should be paired with continuous literacy interventions and interface-level safeguards, recognizing that user behavior and client design can amplify or blunt classifier effectiveness (Veit et al., 2025). Finally, given the accelerating role of LLMs in both attack and defense, institutions should develop governance guidance for LLM-assisted security workflows covering

auditability, bias, and robustness so that adoption improves resilience rather than introducing new blind spots (De Queiroz, 2025).

REFERENCES

Alguliyev, R., Aliguliyev, R., & Sukhostat, L. (2024). An Approach for Business Email Compromise Detection using NLP and Deep Learning. *18th IEEE International Conference on Application of Information and Communication Technologies, AICT 2024*. https://doi.org/10.1109/AICT61888.2024.10740431

Bagui, S., Nandi, D., Bagui, S., & White, R. J. (2021). Machine Learning and Deep Learning for Phishing Email Classification using One-Hot Encoding. *Journal of Computer Science*, *17*(7), 610–623. https://doi.org/10.3844/jcssp.2021.610.623

Bera, D., Ogbanufe, O., & Kim, D. J. (2023). Towards a thematic dimensional framework of online fraud: An exploration of fraudulent email attack tactics and intentions. *Decision Support Systems*, *171*. https://doi.org/10.1016/j.dss.2023.113977

De Queiroz, H. J. D. S. (2025). Phishing and social engineering attack prevention with LLMs. In *Revolutionizing Cybersecurity With Deep Learning and Large Language Models* (pp. 133–163). https://doi.org/10.4018/979-8-3373-3296-3.ch005

Ferreira, A., Coventry, L., & Lenzini, G. (2015). Principles of persuasion in social engineering and their use in phishing. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9190*, 36–47. https://doi.org/10.1007/978-3-319-20376-8_4

Gupta, A., Mishra, A. K., & Arora, K. (2025). Detecting Phishing Emails Using Natural Language Processing. *2025 International Conference on Pervasive Computational Technologies, ICPCT 2025*, 234–238. https://doi.org/10.1109/ICPCT64145.2025.10941056

Hilani, M., Nassih, B., Lmati, I., Balouki, Y., & Amine, A. (2025). Phishing Email Detection Using NLP and CNN Model. *Lecture Notes in Networks and Systems, 1486 LNNS*, 203–212. https://doi.org/10.1007/978-3-031-95330-9_22

Iwara, I. O. (2025). Law Enforcement Impersonation Bank-Related Scams in South Africa: Perceived Vulnerability and Mitigative Strategies. *Risks*, *13*(8). https://doi.org/10.3390/risks13080156

Karhani, H. E., Jamal, R. A., Samra, Y. B., Elhajj, I. H., & Kayssi, A. (2023). Phishing and Smishing Detection Using Machine Learning. *Proceedings of the 2023 IEEE International Conference on Cyber Security and Resilience, CSR 2023*, 206–211. https://doi.org/10.1109/CSR57506.2023.10224954

Mahmud, A. F., & Wirawan, S. (2024). Deteksi Phishing Website Menggunakan Machine Learning Metode Klasifikasi. *Sistemasi: Jurnal Sistem Informasi*, *13*(4), 1368–1380. https://doi.org/10.32520/stmsi.v13i4.3456

Palatty, N. J. (2026). 81 Phishing Attack Statistics 2026: The Ultimate Insight. In *Astra Security Blog*. Astra Security. https://www.getastra.com/blog/security-

audit/phishing-attack-statistics/

PANDI, P. N. D. I. I. (2022). IDADX Terima 34.622 Laporan Kejahatan Phishing dalam 5 Tahun. In *PANDI Press Release*. Pengelola Nama Domain Internet Indonesia (PANDI). https://pandi.id/en/siaran-pers/idadx-terima-34-622-laporan-kejahatan-phishing-dalam-5-tahun

Perbendaharaan, D. J. (2025). *Phishing: Pengertian, Jenis, dan Cara Menghindari Phising*. Direktorat Jenderal Perbendaharaan, Kementerian Keuangan Republik Indonesia. https://djpb.kemenkeu.go.id/kppn/manna/id/data-publikasi/artikel/3239-keamanan-informasi-phishing-pengertian,-jenis,-dan-cara-menghindari-phising.html

Phang, Z. H., Tan, W. M., Xiong Choo, J. S., Ong, Z. K., Isaac Tan, W. H., & Guo, H. (2024). VishGuard: Defending Against Vishing. *Proceedings of the 8th Cyber Security in Networking Conference: AI for Cybersecurity, CSNet 2024*, 108–115. https://doi.org/10.1109/CSNet64211.2024.10851764

Pimpason, N., Viboonsang, P., & Kosolsombat, S. (2025). Phishing Email Detection Model Using Deep Learning. *International Conference on Cybernetics and Innovations, ICCI 2025*. https://doi.org/10.1109/ICCI64209.2025.10987422

Sarno, D. M., & Black, J. (2024). Who Gets Caught in the Web of Lies?: Understanding Susceptibility to Phishing Emails, Fake News Headlines, and Scam Text Messages. *Human Factors*, *66*(6), 1742–1753. https://doi.org/10.1177/00187208231173263

Sommestad, T., & Karlzén, H. (2024). The unpredictability of phishing susceptibility: results from a repeated measures experiment. *Journal of Cybersecurity*, *10*(1). https://doi.org/10.1093/cybsec/tyae021

Topor, L., & Pollack, M. (2022). Fake Identities in Social Cyberspace: From Escapism to Terrorism. *International Journal of Cyber Warfare and Terrorism*, *12*(1). https://doi.org/10.4018/IJCWT.295867

Veit, M. F., Wiese, O., Ballreich, F. L., Volkamer, M., Engels, D., & Mayer, P. (2025). SoK: The past decade of user deception in emails and today's email clients' susceptibility to phishing techniques. *Computers and Security*, *150*. https://doi.org/10.1016/j.cose.2024.104197

Vidyasri, P., & Suresh, S. (2025). FDN-SA: Fuzzy deep neural-stacked autoencoder-based phishing attack detection in social engineering. *Computers and Security*, *148*. https://doi.org/10.1016/j.cose.2024.104188