International Journal of Emerging Engineering and Technology (IJEET)
ISSN (e): 2958-3764
Volume: 4, Number: 2, Pages: 61- 70, Year: 2025

Phishing Attack in the Age of Artificial Intelligence: A Systematic Survey

Amara Nazir[1,*], Irshad Ahmed Sumra[2], and Fahad Ali[3]

[1]Department of Artificial Intelligence, University of Management & Technology, Lahore, 54000, Pakistan.
[2,3] Department of Informatics and Systems, University of Management & Technology, Lahore, 54000, Pakistan.
[3] School of Systems and Technology, University of Management & Technology, Lahore, 54000, Pakistan.
[*]Corresponding author: Amara Nazir (Email: f2023436011@umt.edu.pk)

Received: 11/08/2025, Revised: 20/10/2025, Accepted: 18/11/2025

***Abstract*— One of the most dangerous frauds from today's techno-psychological era is phishing that plays on the vulnerabilities of technology and human psychology through email, SMS, telephone or social media and other new channels. Phishing attacks have evolved significantly with the advent of Artificial Intelligence, changing their nature, scope, and stealthiness, and continuing the arms race with defenders. The results reveal a number of points that need to be addressed, such as the growing prevalence of detections, the continued vulnerabilities within the ML systems and the necessity for multi-layered defense mechanisms that combine automated AI answers with controlled, human monitoring. Critical gaps include explainable AI, non-email vector coverage, and adversarial robustness benchmarks.**

***Index Terms*— Adversarial Attacks, Artificial Intelligence, Cybersecurity, Deep Learning, Machine Learning, Phishing Detection.**

INTRODUCTION

In the world we live in today, the phishing attack is a serious and costly problem. While traditional cyber exploits focus on software vulnerabilities and the technical aspects of hacking, phishing attacks focus much more on the human element of security and the use of simple human psychological techniques such as trust, urgency, authority and curiosity, to gain the victim's attention and make them provide sensitive information, download malware or authorize a fraudulent transaction. The consequences are significant: In 2022 alone, the FBI Internet Crime Complaint Centre (I3C) estimated that reported cases of phishing and social engineering attacks had a combined cost of more than USD 52 million, which is widely regarded as just a fraction of the total losses incurred [1].

Phishing has been around since the mid-1990s, when early phishers targeted America Online (AOL) users with simple social engineering attacks, impersonating AOL employees to steal their passwords. From these beginnings, phishing has grown into a multi-channel, worldwide crime with a level of sophistication never before seen. Advanced phishing includes spear-phishing (highly targeted attacks profiling specific individuals with Social Media Intelligence), clone-phishing (a duplicate of legitimate communication with a malicious payload), whaling (C-level executive Spear Phishing), smishing (SMS based attack), vishing (Voice call-based attack increasingly powered by AI-synthesized deepfakes), and quashing (QR code-based attack to a malicious site) [1, 2]. Today, with the introduction of the metaverse and IoT ecosystems, new phishing vectors are already emerging that are not covered by the current detection systems [4]. The problem is huge and growing. APWG reports that there were roughly 255,000 new phishing attacks per day at peak times [1]. In 2021, SlashNext was able to track 50,000 malicious URLs on a daily basis, which represents approximately 255 million phishing attacks per year, or a 61% increase from the previous year [3-10]. Advanced computer vision and machine learning technologies detected a whopping 263% more credential phishing links in 2023, totalling 870,555, compared with 2022 [11, 12]. These figures not only highlight the increasing threat but also the effectiveness of AI-powered detection tools at scale. The following Fig. 1 is a complete phishing attack process.

Fig. 1. Phishing attack process [1].

The phishing threat landscape has undergone a complete transformation thanks to Artificial Intelligence. Generative Artificial Intelligence (AI) and Large Language Models (LLMs) like GPT-4 are used by attackers to create hyper-personalised, grammatically correct phishing messages that can easily evade signature-based filters and even experienced human recipients [13]. Meanwhile, deep learning, natural language processing, computer vision and behavioral analytics techniques enable proactively identifying and preventing phishing attacks in real time without having to be apparent to

This work is licensed under a Creative Commons AttributionShareAlike4.0 International License, which permits Unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited

the defenders, something rules cannot do. This duality of AI gives rise to an arms race, for which ongoing research and innovation are required [4] [14-17].

The contributions that this paper makes are as follows: (1) systematic synthesis of 20 peer-reviewed studies on the application of AI and phishing over a time frame of 2018–2024; (2) structured analysis of the attack techniques used by AI and their impact at each step of the phishing campaign life cycle; (3) comprehensive comparison of the detection approaches and quantitative statistics; (4) the identification of five major open research gaps; and (5) a recommended multi-layered defense architecture based on empirical evidence.

The rest of this paper is structured as follows: Background on the evolution of phishing and related work is given in Section II. The methodology used in conducting the SLR is presented in Section III. Key findings (statistical analysis and answers to research questions) are presented in Section IV. In section V, AI's role as a threat vector and a defense strategy is explored. A comparative analysis of the detection approaches is presented in Section VI. Research gaps and future directions are identified in Section VII. The survey is completed in Section VIII.

I. Background and Related Work

A. Evolution of Phishing Attack Techniques

The initial literature on phishing focused on the email problem, user gullibility, and bulk campaigns with blatant grammatical mistakes and unbelievable bait. More than a decade earlier, Butler [16] offered early evidence of the psychological mechanisms used by phishers and set the groundwork for understanding that phishing attacks rely on trust relationships rather than merely technical vulnerabilities. Over the following decades, smishing, vishing (voice-based calls) and quishing (QR-based calls) diversified exponentially, with AI-generated voice and sound files becoming increasingly convincing and capable of fooling executives or family members, as well as attacks on social networking platforms [2, 14].

One of the most detailed anatomies of contemporary phishing was provided by Alkhalil et al. [2], who divided the attacks into three categories: by vector (email, web, mobile, social media); by targeted entity (individuals, enterprises, financial institutions); and by technical approach (spoofing, credential harvesting, malware delivery). Most importantly, they noted how personalization was becoming a major factor in making attackers evade automated phishing detection filters that depended on simple pattern matching with phishing “bait” that was based on open-source intelligence (OSINT) that was collected from social media profiles. [2]

The advent of generative AI has ushered in a new era in phishing capabilities. LLMs can craft phishing emails that closely replicate legitimate business messages and, by analysing previous emails, imitate the sender's tone and writing style. In a controlled experiment involving SMS messages with a specific set of participants, Francia et al. [13] measured authenticity ratings for AI-generated messages and determined that they were statistically indistinguishable from human-generated messages; however, this was the case for one study with one form of messaging, and experiments are yet to be performed on email, voice messages and other kinds of messages. Rajivan and Gonzalez [7] noted that advanced social engineering attacks that exploit various cognitive vulnerabilities remain effective against advanced targets when messages are personalised, highlighting that existing technical defences have not been able to reduce the human cognitive weaknesses that phishing attacks exploit.

B. Machine Learning Approaches to Phishing Detection

There is a long history of using machine learning for phishing detection. Classical Machine Learning algorithms such as Random Forest, XGBoost, Support Vector Machines (SVMs), Naive Bayes, and k-Nearest Neighbours have proven effective for URL/HTML feature-based classification tasks. Some of the most important characteristics used in these models include the presence of HTTPS indicators, IP addresses in URLs, subdomains, URL shorteners, domain age, and lexical patterns [5, 15]. On various benchmark datasets, Bhavani et al. [5] achieved impressive classification accuracy with these methods but pointed out a fundamental problem: the methods perform poorly on new attack patterns not present in the known datasets used to train the models.

With the help of automatic feature extraction from raw inputs, deep learning models have also improved the detection capability. Various Convolutional Neural Networks (CNNs) for URL character sequences, Long Short-Term Memory (LSTM) networks for sequential pattern modeling, and Transformer-based architectures using Natural Language Processing (NLP) have all contributed to achieving a detection accuracy greater than 99% on benchmark datasets collected and curated [18] [19]. Alsubaei et al. [17] introduced a hybrid ResNeXt-GRU model tailored for real-time phishing detection in cybercrime forensics, addressing class imbalance via SMOTE augmentation and achieving state-of-the-art performance. Ensemble learning with deep learning architectures yielded the best results for phishing website classification, achieving 99% accuracy, as reported by Zara et al. [18].

The NLP techniques work on phishing at the content level, not the URL/HTML level. BERT models, GPT classifiers, and Sentence Transformers can perform semantic and contextual analysis of email and message text and detect subtle linguistic anomalies unique to phishing messages, even when the rest of the text appears to be from a legitimate source. This is especially significant for protecting AI-generated spear-phishing content that has been visually and syntactically approved [13]. A third complementary approach is browser-based ML defenses, which conduct real-time scan of web content against phishing indicators, especially useful against zero-day attacks that are not included in the blacklist [6].

C. Adversarial Threats to ML Detection Systems

Numerous studies have shown that phishing detection is vulnerable to adversarial attacks, especially when machine learning techniques are used. Adversarial examples are inputs designed to circumvent ML classifiers, yet be deceptive to human recipients. Examples of these in the phishing context

include using visually similar Unicode domain names, reformulating email body text to maintain the phishing intent, and making small changes to the visual aspects of phishing web pages to deceive computer-based phishing classifiers [11].

There are two main vulnerability mechanisms identified in the literature: Overfitting to training data results in models trained on known patterns from phishing campaigns to perform well when tested on the traditional types of phishing attacks, but to perform poorly when tested on novel, slightly different attacks — a space that is actively being exploited by the advanced attackers who try to imitate the phishing campaigns in a model-significant way, while remaining essentially phishing. Second, surface feature-based ML models can be easily fooled by attackers who can generate content that satisfies the model's syntactic and structural requirements and mimics the target's structure and style, using the same structures and patterns as those of legitimate entities [4], [11]. The arms-race characterisation of the detection landscape was validated by adversarially crafted browser attacks that can bypass even state-of-the-art browser-based ML defences, as presented by Arun and Abosata [6]. The following Fig. 2 is the representation of attack vectors and victim targeting:

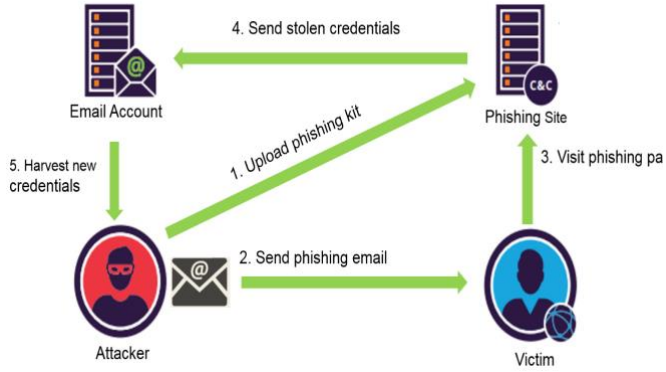

Fig. 2. How phishing works: attack vectors and victim targeting [2].

II. RESEARCH METHODOLOGY

The Systematic Literature Review (SLR) methodology proposed by Kitchenham and Charters [9] is employed, which is widely regarded as the gold standard in software engineering and computer science for evidence synthesis. The SLR methodology is a structured, reproducible approach to locating, identifying, reviewing and integrating primary research evidence to answer research questions. It is divided into 3 stages: planning, conducting, and reporting, each with specific tasks, decision criteria, and quality checkpoints that help reduce selection bias and ensure rigour.

A. Phase 1: Planning the Review

The identification of the critical gap in the literature began with the need to survey recent literature that would summarise developments in both offensive and defensive aspects in the context of AI-enabled phishing attacks. The four research questions were developed to direct the evidence collection and synthesis process:

RQ1: What are the potential ways of detecting AI-generated phishing messages that are so human-like that they would not trigger a high number of false positives and thereby negatively impact legitimate communications?

RQ2: What are the vulnerabilities of the current phishing detection systems based on AI, and what mitigation strategies have been shown to be successful?

RQ3: Can AI/machine learning be used to predict and stop new phishing attacks before they are launched on a large scale?

RQ4: What impacts will come from the traditional phishing campaign lifecycle when AI becomes part of the process to the real-time threat detection architecture?

The following 7 bibliographic databases were chosen to comprehensively cover computer science, cybersecurity, and information systems research: IEEE Access, Google Scholar, ResearchGate, PubMed, ScienceDirect, Springer, and Scopus. The search was limited to peer-reviewed journal articles and conference proceedings from 2010 to 2024, providing historical context and the latest updates on the evolving landscape of AI threats.

B. Phase 2: Conducting the Review

The following keywords were used in the search: phishing attacks, artificial intelligence, machine learning, cybersecurity, threat detection, deep learning, NLP, spear-phishing, adversarial attacks, phishing detection, and LLM phishing. All identified databases were searched systematically, and the results were deduplicated before screening, eliminating duplicates, and again after screening.

Two phases of screening were applied: a rigorous one. During Phase 1, titles and abstracts were subjected to the following inclusion criteria: The study must involve AI, ML or deep learning applied to phishing attacks or defenses; The study must provide original empirical data, new methodology, or critical analysis of existing techniques; and the study must have been published in a peer-reviewed venue. Studies were not included if phishing was mentioned only as part of a larger cybersecurity survey (without extensive detail on phishing), if the methodology was not described, or if the source was a predatory or non-peer-reviewed journal. In Phase 2, full-text analysis of all studies that met the criteria in Phase 1 was carried out, using the more rigorous quality criteria.

The quality assessment used the Kitchenham framework, in which each retained study was evaluated across five dimensions. Each dimension was scored on a 5-point scale (1 = poor; 5 = excellent), with a maximum score of 25 per study. Each of the 70 full-text studies was independently scored by two reviewers and included in the final synthesis if both reviewers gave scores of 18 or more (72%), with no dimension scoring below 3. The reviewers' scores were discussed, and a consensus score was obtained, with no need for a third-party adjudicator.

The 50 studies excluded at this stage were those that did not meet the minimum quality threshold of 18/25 (34), those with significant methodological overlap and the same conclusions as studies already included (11), and those with inadequate methodological detail to reliably score on one or more dimensions (5).

C. Study Selection Outcome

The study selection funnel is shown in Table I for each SLR phase. A total of 20 high-quality studies were included in the evidence base for this survey after applying the inclusion and exclusion criteria and quality assessment of 250 articles initially identified.

TABLE I: STUDY SELECTION FUNNEL ACROSS SLR PHASES

Phase	Action Taken	Studies Remaining	Reduction
Database Search	Searched 7 databases using structured keyword combinations across IEEE Access, Google Scholar, ResearchGate, PubMed, ScienceDirect, Springer, Scopus	250 identified	Starting pool
Exclusion Criteria	Titles and abstracts screened; studies lacking AI/ML phishing focus removed; predatory venue publications excluded.	150 retained	-40%
Inclusion Criteria	Full-text review; retained empirical, methodological, and critical review studies presenting original AI phishing findings	70 retained	-53%
Quality Assessment	Used Kitchenham quality indicators across 5 dimensions scored 1-5 (max 25); only included studies with a total score ≥18/25 and with no dimension scoring <3.	20 finals	-71%

III. KEY FINDINGS

A. Publication Trend Analysis (2018-2024)

There is a clear and statistically significant increase in publications among the 20 selected studies, with the steepest rise since 2022. It is no coincidence that this time period coincided with the public launch of advanced generative AI tools, most notably ChatGPT in November 2022, not only increasing the offensive toolkit on the attackers' side but also sparking general academic interest in the potential of using AI for defense. In 2024, the number of papers published reached a peak of 7 (35% of the corpus), highlighting the rapid development of the LLM-driven phishing and detection field. The complete distribution by year and research focus is given in Table II.

TABLE II. DISTRIBUTION OF SELECTED STUDIES BY PUBLICATION YEAR

Year	Papers	Share	Primary Research Focus
2018	1	5%	Social engineering and adversarial persuasion strategies in targeted phishing
2019	2	10%	ML models for URL-based phishing detection; browser-based real-time defenses
2020	1	5%	Comprehensive survey of phishing types, vectors, and technical countermeasures
2021	1	5%	Anatomy of modern phishing; rise of spear-phishing and targeted campaigns
2022	5	25%	Anti-phishing technique reviews; adversarial ML; credential phishing detection
2023	3	15%	AI-powered browser attacks; phishing taxonomies; computer-vision-based detection
2024	7	35%	LLM-based phishing; zero-day detection; PhishNET; generative AI defenses; hybrid DL

B. Phishing Detection Scale — AI/ML Performance Statistics

All these studies reveal that the size and number of phishing attacks, as well as the size of AI-based phishing defenses, keep growing and becoming more significant. Such statistics back the seriousness of the threat panorama and the ever-stronger potential of automatic detection systems. Table III shows the statistics of the important detected volume.

SlashNext has been effective in blocking the distribution of around 50k malicious URLs per day in 2021 with its ML based detection technology. This had grown to 80,000 per day by 2022, representing a 61% year-on-year increase of around 255 million phishing attacks detected every year [10]. Using the most advanced computer vision and ML techniques, by 2023, 870,555 credential phishing links have been identified, representing a 263% increase from 2022 [12]. This dramatic increase is attributed to the threat itself and to computer vision technologies becoming more mature, especially in the case of credential phishing pages (CP) which copy a legitimate webpage of a site for a user to log in to.

TABLE III. AI/ML-BASED PHISHING DETECTION STATISTICS BY YEAR

Year	Detection Volume	Technology Used	Key Observation
2021	~50,000 malicious URLs/day	SlashNext ML-based detection	Baseline year; high volume but annual totals not systematically reported
2022	~80,000 URLs/day; ~255M attacks/year	AI-based ML and heuristic systems (SlashNext)	61% increase over 2021; first year with comprehensive annual total reported
2023	870,555 credential phishing links detected	Computer Vision + ML (brand analysis, login form detection, image classification)	263% increase over 2022; reflects maturation of CV-based phishing page detection

C. Research Question Findings and Recommendations

Detailed findings and evidence-based recommendations were obtained from structured data extraction from the 20 selected studies to address each of the four research questions. These have been summarized in Table IV.

TABLE IV. SUMMARY OF RQ FINDINGS AND RECOMMENDATIONS

RQ	Question Focus	Key Finding	Recommendation
RQ1	Detecting AI-phishing without high false positives	AI-generated phishing closely emulates human communication patterns, rendering signature-based and keyword-matching systems ineffective. Deep learning and NLP models detect subtle syntactic and contextual anomalies. Contextual awareness via comparison against	Deploy multi-dimensional detection combining DL, NLP, contextual awareness, and behavioral analytics. Apply confidence scoring and threshold tuning to route uncertain cases to human analysts. Implement continuous

		historical user interaction profiles enables detection even when surface features are legitimate.	feedback loops to refine detection boundaries.
RQ2	Adversarial vulnerabilities and mitigation strategies	ML detection systems are systematically vulnerable to adversarial examples — subtle input modifications sufficient to alter model classification while preserving phishing effectiveness. Overfitting to training data is a primary weakness. Lack of contextual understanding enables evasion through legitimate communication style mimicry.	Training with adversarial training and adversarial examples. Implement ensemble methods in different architectures. Continually train with live threat intelligence. Focus on meaning and context in addition to surface features.
RQ3	Anticipating new phishing strategies via AI	AI predictive analytics can be used to model and detect historical attack trends, enabling the prediction of future attacks. Anomaly detection which seeks to identify activity different than what is known can find zero day attacks before they are available to the general public. Adaptive learning models are continually updated with new threat information.	Provide threat intelligence feeds and user feedback to feed models. Use proactive threat hunting agents to monitor for new phishing infrastructure on open and dark web. Use adaptive anomaly detection, automatic threshold adjustment.
RQ4	AI's transformation of the phishing lifecycle	From victim targeting – scraping data from social media to building profiles – to generating personalized content with LLMs, to optimizing delivery volume with large-scale optimization and to optimizing campaigns through feedback loops.	Embed real-time AI anomaly detection matching attack automation speed. Implement human-AI collaboration where AI highlights suspicious signals and human analysts adjudicate high-stakes blocks. Use traditional filters for known patterns; reserve advanced AI for novel attack variants.

IV. Discussion

A. A Dual Role of AI in the Phishing Ecosystem

The most important and influential finding in the literature is the two-sided approach of Artificial Intelligence in the phishing area. The paradox of AI that fuels the arms race, as observed in the studies reviewed below [3] and [4], is that it simultaneously serves as the strongest weapon in the attackers' arsenal and in the *defenders*' arsenal.

On the offensive side, AI can power phishing attempts on an unprecedented scale, level of personalization, and adaptability. The ability to generate phishing content using AI models, including LLMs, enables the creation of personalised messages that closely resemble those of a legitimate organisation or person. Compared to previous phishing attacks, which were generally grammatically incorrect and featured unlikely storylines, the phishing content generated by LLMs has proven to evade automated filters, and in experimental settings, some recipients have found it hard to identify as phishing, even with knowledge of phishing techniques [13]. It should be noted that this is the result of controlled studies, not deployment data from large-scale real-world deployments. Automated social media scraping can be used to profile a victim at scale to identify their relationships, interests, recent activities, and communication patterns, which can be used to personalize phishing lures to an unprecedented level [7]. Self-optimizing campaign algorithms continually learn and adapt evasion rates and CTR based on the data from previous campaign iterations.

On the other hand, deep learning and NLP have turned the tide in detection capabilities, making them unachievable by rule-based and blacklist systems. DL models learn these subtle distributional differences between phishing and legitimate communication based on training on large labeled datasets of both, where the differences in language, structure and behavioral patterns are not explicitly formulated as rules [18] [19]. Computer Vision systems can detect credential phishing pages by analysing the visual consistency of branding and the structure of login forms, without relying on URL characteristics, and can detect phishing pages even when the URL domain has just been registered and is not on any blacklist [12]. Behavioral analytics systems create a long history of user behavior and detect activity that has changed since a user last interacted with a website, even if the content of the phishing site does not pass all the surface level tests.

However, in practical application, AI-driven defenses are not without their limitations, which affect their effectiveness in real-world scenarios. In general, both deep learning and NLP-based approaches as well as computer vision take up a significant amount of processing power for training and inference, making deployment at scale and in real-time challenging and potentially expensive for smaller companies with less dedicated security funds. Many of the datasets used to train and test these models are historical and reflect past phishing trends, so the model's accuracy under benchmark conditions does not necessarily indicate its performance against novel AI-driven attacks it has yet to see. Additionally, many high-performing models are black-boxes and thus difficult to explain, which diminishes analyst trust and makes it challenging to comply with regulations in high-risk industries as discussed in Section VII. In combination, these limitations argue that high accuracy in a controlled evaluation does not mean high accuracy in actual, adversarial production environments and thus underscore the need for the multi-layered defense strategy mentioned below.

B. Adversarial Vulnerabilities and Mitigation Strategies

Adversarial attacks to ML-based detection systems is one of the most serious and overlooked threats in phishing protection. In the phishing context, the target is always human, making adversarial examples a highly convincing threat that can cause the ML classifier to misclassify.

Adversarial examples can take many forms, including phishing. Homograph attacks occur when ASCII-based characters are replaced with the same Unicode characters in domain names, resulting in URLs that look correct to humans but are recognised differently by models trained on ASCII-normalised representations. The subtle use of reformulations — including changes in word order, synonyms, or innocuous

content — can make the model shift from phishing to legitimate while maintaining the social engineering effectiveness of the message. Likewise, there are minor changes to the visual features of phishing pages, such as altering the pixel values of the brand logo image or the position of form elements, which can go undetected by computer vision classifiers but not by a human phishing victim [4, 11].

There is another similar and serious vulnerability: overfitting. The models based on clear examples of phishing attacks perform well on standard test sets from the same distribution but are not significantly successful when challenged by new phishing variants. Sophisticated attackers can learn from published models, or they can engage in a systematic probing process to discover the types of attacks that can be categorized as phishing in ways that are important to the model, and then design campaigns that are also classified as phishing, but are still highly effective at getting humans to click on links.

The literature reviewed indicates that several proven mitigation measures exist. Adversarial training can boost the robustness of models to particular attack classes using adversarially generated examples, but requires regular updates as adversaries evolve. Ensemble learning uses several different model architectures, and therefore, the chance of fooling every model is much smaller than the chance of succeeding with just one. Live threat intelligence feeds for continuous retraining maintain the models' relevance for new campaign variants. Semantic and contextual analysis, which is based on the purpose of the communication rather than its surface markings, is more resistant to complex forms of adversarial manipulation [11].

C. Recommended Multi-Layered Defense Architecture

A common and consistent theme that emerged from all of the studies examined was that a “defense in depth” approach is required, involving complementary technological solutions backed up by well-planned human oversight. There is no single detection solution that can effectively address the full range of today's phishing attacks across attack vectors, sophistication levels, and target contexts. The layers are supposed to complement one another: each other's weakness should be compensated for by the strength of another layer and each layer should offer multiple coverage from various attack strategies [3].

The suggested seven layers architecture is as follows:

Layer 1 - AI/ML Anomaly Detection: AI/ML models on email content, URL characteristics and behavioral indicators provide primary defense against AI generated and new phishing content that evades surface level detection.

Layer 2 - Traditional Heuristic and Rule-Based Filters: Blacklists, pattern matching rules, heuristic scoring systems, efficient and low-cost protection against high-volume commodity phishing based on known patterns and signatures.

Layer 3 - Browser-Based Real-Time Protection: Browser extensions that use ML to analyze content, check brand consistency and make real-time comparisons with phishing indicators during the loading of Web pages, notifying the user of suspicious content.

Layer 4 - Behavioral Analytics: Continuous longitudinal profiling of user interaction patterns, which allows to detect phishing-induced behavioral deviations without regard to content quality, and to deliver a content-independent detection layer.

Layer 5 - Human-AI Collaboration: Human security analysts as last-resort decision makers for high-confidence flagged cases, backed by AI-generated evidence summaries, confidence scores and contextual explanations to facilitate informed, accountable blocking decisions, while not automating the blocking process.

Layer 6 - User Education and Awareness - Ongoing phishing simulations, training sessions, more stringent implementation of multi-factor authentication and awareness building campaigns for the ongoing human vulnerability that cannot be fully mitigated technically.

Layer 7 – Feedback and Continuous Learning: Integrating user-reported phishing signals and confirmed attack data into model retraining pipelines in a systematic manner, and proactively monitoring threat intelligence to detect new campaigns in early stages and build their infrastructure.

The following is the recommended multilayer defense architecture Fig. 3:

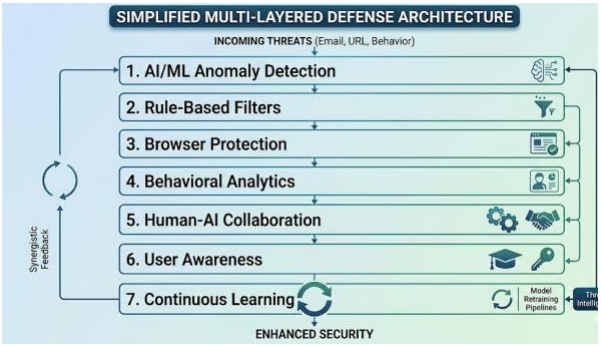

Fig. 3. Recommended multi-layered defense architecture against AI-powered phishing [3], [6].

D. Implications of AI for the Phishing Campaign Lifecycle

Traditional phishing attacks were hard to get right on all fronts of the attack lifecycle from identifying and researching targets, to creating a relevant-looking lure, to establishing a convincing phishing infrastructure, to sending the attack content, to looking for responses from victims. This manual process added a lot of limitations to campaigns in terms of scale, sophistication and adaptability [4] which AI takes away.

There is a huge amount of victim data on social media platforms and other OSINT sources that can be scraped and aggregated for the purpose of AI-driven campaigns for the precise victim profiling, the identification of particular victims with specific roles, relationships or recent activities that make them more prone to a specific pretext in a phishing campaign and a number of victims that would not be possible without the use of AI. Using an LLM to generate content yields a high volume of unique phishing messages in many languages and styles, thousands per minute. Phishing routes are optimized using algorithms to ensure that phishing information is sent via routes likely to be followed by and deceive specific target profiles. Self-learning feedback loops consider the percentage of victims that were responded to, percentage of events that victims reported and percentage of victims that managed to escape the campaign and continuously tune the campaign

parameters so as to maximize the effectiveness and minimize the percentage of detected victims.

This automation presents a basic speed imbalance for defenders: AI-driven campaigns can produce, deploy, and evolve attack variations at a rate that's beyond the pace of signature-updates. Detection signatures can quickly go stale when deployed against a rapidly-evolving attacker. This means detection architectures that can be continuously updated with feeds of live threat information, not periodically, with the ability to update signatures, human-AI interaction and make near real time, high-stakes blocking decisions at machine speed without sacrificing accuracy or accountability [4, 11].

V. Comparative Analysis of Detection Approaches

The 6 major phishing detection techniques examined and evaluated in the 20 studies range from traditional statistical machine learning to the latest deep learning and behavioral analysis. These approaches are tabulated and compared across 5 categories in Table V of the next section: Operating mechanism, key strengths, inherent limitations, and representative algorithm. This comparative view enables security practitioners and researchers to choose the appropriate detection system depending on the type of threat, deployment and performance requirements. In Table V, the comparative analysis of phishing detection approaches is given:

TABLE V. COMPARATIVE ANALYSIS OF PHISHING DETECTION APPROACHES

Approach	**Mechanism**	**Strengths**	**Limitations**	**Representative Algorithms**
Classical ML	Manual feature engineering from URL/HTML structures fed to trained classifiers	Fast inference; interpretable; low computational resource requirements; proven effectiveness on known phishing patterns	Requires expert feature engineering; performance degrades on novel attack variants; vulnerable to adversarial examples targeting known features	Random Forest, XGBoost, SVM, Naive Bayes, k-NN
Deep Learning	Automatic feature extraction from raw URL, HTML, or email content via multi-layer neural architectures	High accuracy on large datasets; automatic feature learning eliminates manual engineering; handles complex non-linear patterns; scalable	Computationally intensive (GPU required); large training data requirement; longer training cycles; overfitting risk on small or unbalanced datasets	CNN, LSTM, DNN, Transformer (BERT), ResNeXt-GRU [17]
NLP / LLM-Based	Semantic and contextual analysis of email/message text using pre-trained or fine-tuned language models	Context-aware; detects subtle linguistic anomalies; effective against AI-generated spear-phishing; captures meaning beyond surface syntax	High computational cost; requires large text corpora for fine-tuning; risk of false positives on formal or unusual legitimate communications	BERT, GPT-based classifiers, Sentence Transformers, RoBERTa
Computer Vision	Visual analysis of webpage screenshots comparing brand identity elements, color schemes, and login form structures	Effective for credential phishing; independent of URL/HTML features; catches visual mimicry of legitimate brands	High computational cost for real-time deployment; limited to visual layer; may be defeated by high-fidelity visual clones using transferred brand assets	CNN-based image classifiers, Siamese networks, perceptual hashing, ResNet
Behavioral Analytics	Longitudinal profiling of user-system interaction patterns with anomaly detection on deviations from established baselines	Content-independent detection effective against sophisticated content; detects phishing effects (behavioral changes) rather than phishing content	Extended observation period required to establish reliable baselines; privacy implications; elevated false-positive rates during profile establishment phase	LSTM-based anomaly detection, Isolation Forest, statistical deviation models, autoencoders
Rule-Based / Blacklists	Explicit pattern matching of URLs, domains, and content against maintained databases of known phishing indicators	Very low computational overhead; extremely fast lookup; highly effective against high-volume, known-pattern commodity phishing	Completely ineffective against zero-day attacks and AI-generated novel campaigns; requires continuous manual curation; easily defeated by domain rotation	Regex-based filters, blacklist lookup, heuristic scoring, DNS blacklists

VI. Research Gaps and Future Directions

The literature reviewed revealed many important goals have not yet been realized and require specific research. These gaps need to be addressed to build a detection system that can keep up with the rapidly changing AI-powered phishing threat landscape.

A. Explainable AI (XAI) Frameworks for Phishing Detection

The high-performance AI detection models analyzed in most cases are black boxes, meaning that the models classify something without providing an explanation for the classification. This poses a serious problem of trust in analysts, operational adoption and regulatory compliance. Transparency and explainability of the AI-driven decision systems are becoming a necessary condition for AI systems in the context

of the EU AI Act and other regulations, particularly in high-risk sectors such as finance and healthcare where phishing is prevalent. From now on, there is a need for research to identify how to present to analysts the features and the reasoning patterns used for the detection of the phishing attack using Explainable AI (XAI) frameworks.

B. Multi-Modal Detection for Non-Traditional Attack Vectors

Since the start of phishing attacks, email and URL phishing have been the predominant types of phishing and benchmark datasets are available for these two types of phishing, the overwhelming majority of literature discusses detecting either email or URL phishing. But the threat landscape has evolved significantly and there are significant gaps in detection in the many ways of attack [2, 4] that are not addressed in this research concentration.

Smishing (SMS phishing) has not yet established good detection frameworks, partly due to the limited richness of the information contained in SMS headers – which is not as rich as the information in email headers – and partly due to the fact that many of the structural features of legitimate SMS communications are the same as those of phishing messages. The added difficulty with vishing (voice phishing) is that it is increasingly hard to detect, as hackers can use AI to create convincing deepfake audio that impersonates well-known people. Quishing (QR code phishing) takes advantage of the fact that most security systems are unable to check what URL a QR code scans to before the scan. Social media phishing exploits the platform's trust mechanisms and informal communication norms. Meta-virus phishing exploits the immersive trust environment and avatar impersonation features of virtual reality sites. One important area of research is multimodal detection frameworks that can operate across various modalities, such as audio analysis, image processing, social graph analysis, and text understanding [4].

C. Dataset quality, currency, and ai-phishing coverage

Training and evaluation data sets are the core to the success of a ML phishing system and they are limited in quality, diversity, and age. Most of the benchmark sets found in the studies reviewed (PhishTank, Alexa, and the UCI phishing dataset) are static and lack features typical of modern AI-produced phishing content. This raises a basic evaluation gap: A model with 99% accuracy on these benchmarks could be significantly less accurate when evaluated on phishing attacks developed with LLM that do not follow the language patterns found in historical attacks [18, 19-24].

The problem of data freshness is exacerbated by the evolution of phishing methods: attacks evolve by the month, so phishing models built with one month of data will be less effective against the next month's attacks. Going forward, future research should strive to continuously evolve "living" datasets that are enriched and updated with real-world phishing examples from honeypot systems, crowdsourced user reports, security vendor threat intelligence, and adversarially generated examples covering a variety of attack types and vectors. Federated learning has been found to provide an especially promising way to federate threat data across various organizations while maintaining the privacy and confidentiality of each organization's respective security event data [4].

D. Human-AI Collaboration Frameworks for SOC Environments

The literature reviewed shows a high level of consensus on the importance of human expertise for phishing detection, especially in making decisions for high-stakes cases and for new attacks which are not necessarily classified correctly. While numerous frameworks of human-AI collaboration have been studied, there are not yet scientifically sound models for how this collaboration is effective in Security Operations Center (SOC) contexts [9].

Critical open questions include what the best criteria are for moving from automated to human decision making, how to design interface tools for the analysts that allow them to make decisions quickly and accurately while minimizing cognitive load during the processing of high volume campaigns, how to calibrate the trust of the analysts in the AI's recommendations so that they don't underuse (ignore accurate AI signals) or automation bias (take the AI's recommendations as gospel without critically examining them), and how to establish feedback mechanisms to bring the analyst's decisions back in to improve the model. Interdisciplinary cooperation is needed in human-computer interaction, cognitive psychology, organization security management and machine learning engineering to answer these questions [25].

E. Standardized Adversarial Robustness Evaluation Benchmarks

In contrast, computer vision and NLP have standardized adversarial robustness benchmarks like RobustBench [1] which allows to systematically track progress and compare studies across different research fields, phishing detection does not have such infrastructure. The lack of adversarial robustness in any set of detection methods makes it very difficult to compare the adversarial robustness of different approaches, assess robustness over time, or set minimum robustness baselines for commercial products [11].

A phishing robustness benchmark set should include: realistic attack scenarios based on observed phishing threat intelligence, not just synthetic scenarios; multiple adversarial perturbation types including URL manipulation, text reformulation and modification of visual elements; representative examples of phishing campaigns covering a range of attack types, attack vectors, and attack sophistication; and standardized evaluation measures that allow for apples to apples comparisons between studies. Such infrastructure would greatly accelerate the research community's efforts to develop, evaluate, and communicate the effectiveness of phishing detection systems [25, 26].

VIII. Conclusion

For this systematic literature review, twenty high-quality studies were analyzed to give a comprehensive synthesis regarding the state of the art regarding artificial intelligence and phishing attacks. The numbers speak for themselves: AI has revolutionized the way phishing works, as well as how to defend against it and how to find new ways to do both. In the case of the offense side, AI introduces the potential for phishing attacks on an unprecedented scale, level of personalization, and

adaptability. AI models create phishing messages which appear identical to legitimate messages, automated targeting selects targets and tailors' messages in a highly personalized manner, and self-learning campaign algorithmic tools continually improve levels of evasion. The amount of phishing has increased dramatically over the past few years, from 50,000 malicious URLs detected daily in 2021 to 870,555 credential phishing links detected using computer vision in 2023 alone, highlighting the threat and scale for which phishing defenses must be prepared.

On the defense side, machine learning technologies like deep learning, NLP, behavioral analytics and computer vision have yielded much more accurate detection and scalability when compared to rule-based methods. These systems are still susceptible to adversarial attacks, overfitting and the fast development of AI-powered offensive tactics, however. Single detection technologies are insufficient and a layered strategy of AI/ML detection and traditional filters, browser-based protection, behavioral monitoring, structured human oversight, and continuous user education are necessary to provide a strong defense.

The most urgent research needs and directions include critical research gaps, including explainable, AI-based models, coverage of non-email attack vectors, current and adversarial-robust training sets, validated models of human-AI collaboration, and standardized adversarial robustness benchmarks. In response to these vulnerabilities, it is essential to develop and adapt current strategies to prevent phishing attacks and outpace the ever-growing capabilities of AI-powered threats. To effectively combat this constant and evolving threat, it's important to take a proactive, as opposed to reactive, stance towards security, through both technical innovation and regular, effective human awareness programs.

Funding Statement

The authors 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

All authors contributed to the conception, literature review, drafting, and critical revision of this manuscript and approved the final version for submission.

Data Availability Statement

Data is available on reasonable request.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

References

[1] J. A. Chaudhry, S. A. Chaudhry, and R. G. Rittenhouse, "Phishing Attacks and Defenses," *Int. J. Secur. Its Appl.*, vol. 10, no. 1, pp. 247-256, Jan. 2016, doi: 10.14257/ijsia.2016.10.1.23.

[2] Z. Alkhalil, C. Hewage, L. Nawaf, and I. Khan, "Phishing Attacks: A Recent Comprehensive Study and a New Anatomy," *Front. Comput. Sci.*, vol. 3, Art. no. 563060, Mar. 2021, doi: 10.3389/fcomp.2021.563060.

[3] P. Sharma, B. Dash, and M. F. Ansari, "Anti-Phishing Techniques - A Review of Cyber Defense Mechanisms," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 11, no. 7, Jul. 2022, doi: 10.17148/IJARCCE.2022.11728.

[4] M. Elatoubi, "Phishing in Web 3.0: Opportunities for the Attackers, Challenges for the Defenders," *ARIS2 - Adv. Res. Inf. Syst. Secur.*, vol. 3, no. 2, pp. 11-25, Dec. 2023, doi: 10.56394/aris2.v3i2.35.

[5] A. Mandadi, S. Boppana, V. Ravella, and R. Kavitha, "Phishing Website Detection Using Machine Learning," in *2022 IEEE 7th Int. Conf. for Convergence in Technology (I2CT)*, Mumbai, India, 2022, pp. 1-4, doi: 10.1109/I2CT54291.2022.9824801.

[6] A. Arun and N. Abosata, "Next Generation of Phishing Attacks Using AI Powered Browsers," *arXiv preprint*, arXiv:2406.12547 [cs.CR], Jun. 18, 2024, doi: 10.48550/arXiv.2406.12547.

[7] P. Rajivan and C. Gonzalez, "Creative Persuasion: A Study on Adversarial Behaviors and Strategies in Phishing Attacks," *Front. Psychol.*, Feb. 2018, doi: 10.3389/fpsyg.2018.00135.

[8] P. Kumar, K. Antony, D. Banga, and A. Sohal, "PhishNet: A Phishing Website Detection Tool using XGBoost," arXiv:2407.04732 [cs.CR], Jun. 28, 2024, doi: 10.48550/arXiv.2407.04732.

[9] B. Kitchenham and S. M. Charters, "Guidelines for Performing Systematic Literature Reviews in Software Engineering," Tech. Rep. EBSE-2007-01, Keele Univ. and Univ. of Durham, UK, Jul. 2007.

[10] SlashNext, "The State of Phishing 2022," SlashNext, Pleasanton, CA, USA, 2022. [Online]. Available: https://slashnext.com. Accessed: Jun. 25, 2026.

[11] V. A. Onih, "Phishing Detection Using Machine Learning: A Model Development and Integration," *Int. J. Sci. Manag. Res.*, vol. 7, no. 4, pp. 27-63, Mar. 2024, doi: 10.37502/IJSMR.2024.7403.

[12] R. R. Nuiaa Al Ogaili and S. Manickam, "A Critical Review: Revisiting Phishing Attacks Classification and Analysis of Techniques Employed in Taxonomies," *Wasit J. Pure Sci.*, vol. 2, no. 2, Jun. 2023, doi: 10.31185/wjps.143.

[13] J. Francia, D. Hansen, B. Schooley, M. Taylor, S. Murray, and G. Snow, "Assessing AI vs Human-Authored Spear Phishing SMS Attacks: An Empirical Study Using the TRAPD Method," *arXiv preprint*, arXiv:2406.13049 [cs.CY], Jun. 2024, doi: 10.48550/arXiv.2406.13049.

[14] R. Alabdan, "Phishing Attacks Survey - Types, Vectors, and Technical Approaches," *Future Internet*, vol. 12, no. 10, Art. no. 168, 2020, doi: 10.3390/fi12100168.

[15] M. S. Liaqat, G. Mumtaz, N. Rasheed, and Z. Mubeen, "Exploring Phishing Attacks in the AI Age: A Comprehensive Literature Review," *J. Comput. Biomed. Inform.*, vol. 7, no. 2, Sep. 2024, doi: 10.56979/702/2024.

[16] R. Butler, "Investigation of Phishing to Develop Guidelines to Protect the Internet Consumer's Identity Against Attacks by Phishers," *South Afr. J. Inf. Manag.*, vol. 7, no. 3, Dec. 2005, doi: 10.4102/sajim.v7i3.269.

[17] F. S. Alsubaei, A. A. Almazroi, and N. Ayub, "Enhancing Phishing Detection: A Novel Hybrid Deep Learning Framework for Cybercrime Forensics," *IEEE Access*, vol. 12, pp. 8373-8389, Jan. 2024, doi: 10.1109/ACCESS.2024.3351946.

[18] U. Zara, K. Ayyub, H. U. Khan, A. Daud, T. Alsahfi, and S. G. Ahmad, "Phishing Website Detection Using Deep Learning Models," *IEEE Access*, vol. 12, pp. 167072-167087, Oct. 2024, doi: 10.1109/ACCESS.2024.3486462.

[19] N. Q. Do, A. Selamat, O. Krejcar, E. Herrera-Viedma, and H. Fujita, "Deep Learning for Phishing Detection: Taxonomy,

CURRENT CHALLENGES AND FUTURE DIRECTIONS," *IEEE ACCESS*, VOL. 10, PP. 36429-36463, 2022, DOI: 10.1109/ACCESS.2022.3151903.

[20] C. OPARA, P. MODESTI, AND L. GOLIGHTLY, "EVALUATING SPAM FILTERS AND STYLOMETRIC DETECTION OF AI-GENERATED PHISHING EMAILS," *EXPERT SYST. APPL.*, VOL. 276, ART. NO. 127044, 2025, DOI: 10.1016/J.ESWA.2025.127044.

[21] APWG, "PHISHING ACTIVITY TRENDS REPORT Q4 2024," ANTI-PHISHING WORKING GROUP, 2024. [ONLINE]. AVAILABLE: HTTPS://APWG.ORG/. ACCESSED: JUN. 25, 2026.

[22] R. GOENKA, M. CHAWLA, AND N. TIWARI, "A COMPREHENSIVE SURVEY OF PHISHING: MEDIUMS, INTENDED TARGETS, ATTACK AND DEFENCE TECHNIQUES AND A NOVEL TAXONOMY," *INT. J. INF. SECUR.*, VOL. 23, NO. 2, PP. 819-848, APR. 2024, DOI: 10.1007/S10207-023-00768-X.

[23] P. BOUNTAKAS AND C. XENAKIS, "HELPHED: HYBRID ENSEMBLE LEARNING PHISHING EMAIL DETECTION," *J. NETW. COMPUT. APPL.*, VOL. 210, ART. NO. 103545, 2023, DOI: 10.1016/J.JNCA.2022.103545.

[24] R. VALECHA, P. MANDAOKAR, AND H. R. RAO, "PHISHING EMAIL DETECTION USING PERSUASION CUES," *IEEE TRANS. DEPENDABLE SECUR. COMPUT.*, VOL. 19, NO. 2, PP. 747-756, 2021, DOI: 10.1109/TDSC.2021.3118931.

[25] SUMRA, I.A., HASBULLAH, H., AB MANAN, J.-L.: EFFECTS OF ATTACKERS AND ATTACKS ON AVAILABILITY REQUIREMENT IN VEHICULAR NETWORK: A SURVEY. IN: INTERNATIONAL CONFERENCE ON COMPUTER AND INFORMATION SCIENCES (ICCOINS2014), MALAYSIA, 3–5 JUNE 2014.

[26] SUMRA,I.A, H. HASBULLAH AND J. -L. A. MANAN, "VANET SECURITY RESEARCH AND DEVELOPMENT ECOSYSTEM," *2011 NATIONAL POSTGRADUATE CONFERENCE*, PERAK, MALAYSIA, 2011, PP. 1-4, DOI: 10.1109/NATPC.2011.6136344.