# Navigating the Phishing Threat Landscape: A Comprehensive Survey of Techniques, Trends, and Countermeasures

**Ajai Ram** *ajairam@cet.ac.in*
*Research Scholar,*
*College of Engineering Guindy, Chennai, India*

**Arockia Xavier Annie R.** *annie@annauniv.edu*
*Associate Professor,*
*College of Engineering Guindy, Chennai, Indiia*

## Abstract

Phishing has evolved from basic deceptive emails into a complex ecosystem of AI-driven attacks that exploit human psychology and digital interconnectivity. This survey revisits the phishing landscape through a novel multidimensional taxonomy that maps attack vectors such as email, SMS, voice, social media, cloud, and IoT against automation levels ranging from manual to Large Language Model (LLM) generated campaigns. It integrates insights from over 200 research works spanning from rule-based, machine learning, deep learning, and graph-based systems to assess their robustness against adaptive adversaries. In this work, we are uniting technical, behavioral and organizational defenses into a cohesive resilience model. We had done a comparative analysis which reveals that while Natural Language Processing (NLP) and transformer-based models outperform classical methods but they are vulnerable to adversarial evasion. This study highlights emerging threats such as phishing-as-a-service (PhaaS), AI-deep fakes and prompt-injection based exploitation. By consolidating performance trends and proposing research priorities this survey paper provides a forward looking blueprint for designing LLM-aware phishing detection and adaptive mitigation system. This survey addresses the research question: How can evolving phishing threats particularly AI and LLM generated attacks can be systematically classified and mitigated through integrated technical, behavioral, and organizational defenses?

**Keywords**: Phishing, Smishing and Vishing, Social Engineering, Machine Learning, Deep Learning, Adversarial Machine Learning, Natural Language Processing, Large Language Models (LLMs).

## 1. INTRODUCTION

Phishing (I. Bose, 2007) continues to be among the most pervasive and deceptive forms of cybercrime, leveraging human psychological manipulation to trick victims into disclosing sensitive information such as usernames, passwords, and financial or personal details. Originating in the early days of the internet, phishing has significantly evolved from rudimentary email scams to highly sophisticated attacks that utilize a multitude of digital channels, including social media, instant messaging, and even voice communications (vishing). The primary aim of phishing is to deceive individuals into perceiving fraudulent entities as legitimate, thereby compromising their vigilance and enabling unauthorized access to sensitive information.

Phishing results in billions of dollars lost annually through fraud, identity theft and ransomware. Beyond economic damage, it erodes user trust and compromises the integrity of digital platforms. For organizations, the consequences include reputational harm, loss of sensitive data, and regulatory penalties. Combating this threat requires a comprehensive strategy that involves advanced technology, robust security policies, and effective user education. As one of the most exploited and adaptable cyber threats, phishing must be a core focus of modern cybersecurity practices.

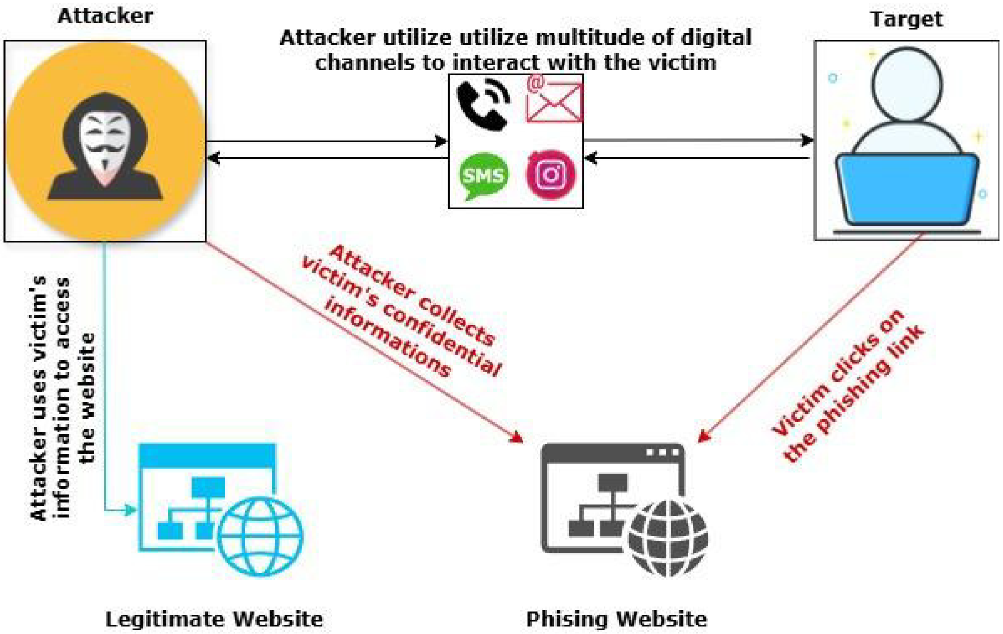

Attacker
Attacker utilize utilize multitude of digital channels to interact with the victim
SMS
Target
Attacker uses victim's information to access the website
Attacker collects victim's confidential informations
Victim clicks on the phishing link
Legitimate Website
Phising Website

**FIGURE 1:** Phishing Concept.

Figure 1 depicts the phishing concept, demonstrating how cybercriminals deceive victims into revealing sensitive information. The process begins with the attacker, who uses various digital communication channels, such as emails, SMS, phone calls, and social media platforms, to initiate contact with the target. These messages often contain deceptive content, such as urgent requests, fraudulent alerts, or enticing offers, persuading the victim to engage with the attacker's message. Once the target or victim receives the message, they are manipulated into clicking on a phishing link, which re-routes them to a phishing website that mimics a genuine website. This fraudulent site is designed to look authentic, tricking the victim into entering their confidential data, financial details, or personal information. The attacker then collects this sensitive data, gaining intrusive access to the victim's accounts. Using the stolen information, the attacker accesses the legitimate website by impersonating the victim, potentially leading to severe consequences, including financial fraud, data breaches, identity theft, or unauthorized transactions. This image clearly demonstrates the deceptive characteristics of phishing attacks, emphasizing the critical role of user awareness, adherence to cybersecurity best practices, and the implementation of multi-factor authentication in mitigating such threats.

Over the years, phishing techniques (Ahmed Aleroud and Lina Zhou, 2017) have become increasingly refined, leveraging advancements in technology to enhance their effectiveness and evade traditional security measures. Modern phishing campaigns often incorporate elements of social engineering, artificial intelligence, and automation to craft personalized and convincing messages that target specific organizations or individuals. Furthermore, the rise of phishing-as-a-service (PhaaS) platforms on the dark web has lowered the technical barriers to entry for cybercriminals, enabling individuals with minimal technical expertise to orchestrate highly effective phishing campaigns. The pervasive nature of phishing poses significant risks not only to individual users but also to enterprises, governments, and critical infrastructure, resulting in substantial reputational damage, financial losses, and breached data integrity. Therefore, comprehending the dynamics of phishing, its continuously evolving techniques, and the corresponding defense mechanisms is essential for formulating effective countermeasures and protecting the digital ecosystem from such attacks.

**Distinct Contributions and Novelty of This Survey**

While several surveys have examined phishing detection techniques and countermeasures, most existing works focus on isolated dimensions such as email-based attacks, specific machine learning models, or technical detection mechanisms. In contrast, this survey introduces a multidimensional and integrative taxonomy that jointly considers attack delivery channels (email, SMS, voice, social media, cloud services, and IoT), levels of automation (manual, semi-automated, and LLM-generated phishing), and defense layers (technical, behavioral, and organizational).

Unlike prior surveys that primarily emphasize detection accuracy or algorithmic performance, this work explicitly incorporates LLM-enabled phishing, phishing-as-a-service ecosystems, and human-centric vulnerabilities, thereby reflecting the evolving threat landscape. By unifying technical detection approaches with behavioral awareness and organizational resilience strategies, the proposed taxonomy offers a holistic framework that is not explicitly addressed in earlier phishing surveys.

## 2. PHISHING ATTACK TECHNIQUES AND CLASSIFICATIONS

Phishing attacks utilize various techniques to mislead users and achieve their malicious objectives, from generic mass campaigns to highly targeted exploits. The common Phishing Techniques involve:

**Email Phishing** (Ian Fette *et al*., 2007) - Mass phishing attacks using deceptive emails to steal sensitive information.

**Spear Phishing** (Nathalie *et al*., 2015) - Targets specific individuals or organizations using personal or organizational details.

**Whaling** (Nathalie *et al*., 2015) - A subset of spear phishing that focuses on high-profile targets like executives.

**Smishing** (Prasadi Kumarasinghe *et al*., 2023) - Phishing through SMS to deceive victims into sharing confidential information.

**Vishing** (Prasadi Kumarasinghe *et al*., 2023) - Voice-based phishing attacks conducted via phone calls.

**Clone Phishing** (Arthur Wong et al., 2022) - Attackers replicate legitimate messages or posts on social media and instant messaging platforms.

The rise of phishing-as-a-service has further diversified phishing tactics, enabling cybercriminals to access pre-designed phishing kits that simplify the process (Meijdam *et al*., 2015; Kelacyber,2025). Each classification reflects the adaptability and creativity of phishing actors, underscoring the need for tailored detection and prevention strategies. A comprehensive explanation of these phishing techniques is given below.

### 2.1 Email Phishing

Email phishing (Ian Fette et al., 2007) is the most common and traditional form of phishing attack, leveraging fraudulent emails to deceive recipients into taking harmful actions. These emails are designed to mimic legitimate communications from trusted entities such as banks, government agencies, or well-known companies. Attackers craft compelling messages that often invoke urgency, fear, or curiosity to manipulate users into clicking on malicious links, downloading malware-infected attachments, or providing sensitive information such as login credentials, financial details, or personal data. For instance, a phishing email may claim that a user's account has been compromised and require immediate action to "verify" their identity by entering credentials on a fake login page.

Phishing email detection relies on multiple sets of features to distinguish between malicious and legitimate emails. These features can be broadly categorized into three main sets:

**Basic Features** (Fette et al., 2007): These features encompass key email characteristics, including sender information, irregularities in subject lines, the presence of suspicious URLs, inconsistencies in HTML structure, and unusual attachment types. Collectively, they facilitate the identification of recurring patterns commonly observed in phishing emails.

**Latent Topic Model Features** (Ramanathan and Wechsler, 2013): This category employs Natural Language Processing (NLP) techniques to examine the textual content of emails. Topic modeling methods, such as Latent Dirichlet Allocation (LDA), assist in identifying phishing attempts by uncovering hidden topics and recurring linguistic patterns commonly observed in deceptive messages.

**Dynamic Markov Chain Features** (Bergholz *et al.*, 2008): These features utilize probabilistic models to analyze the sequential behavior of email elements, including character transitions in URLs, email body structure, and sender-receiver communication patterns. Markov chain-based techniques enhance detection accuracy by capturing evolving phishing strategies and recognizing deviations from legitimate email sequences.

D-Fence (Jehyun *et al.*, 2021) is an advanced, multi-modular phishing email detection system that enhances accuracy and efficiency by integrating machine learning and deep learning techniques. Unlike traditional phishing detection mechanisms that utilize standalone models, D-Fence employs three specialized modules (i) Structure, (ii) Text, and (iii) URL to analyze different aspects of an email. In the (i) Structure Module, the email headers and HTML formatting are used with a tree-based classifier to detect anomalies or spoofing attempts. In the (ii) Text Module, Bidirectional Encoder Representations from Transformers (BERT) are used to understand the semantic meaning of email content and identify deceptive language. Meanwhile, the (iii) URL Module employs a deep learning-based classifier to analyze URLs, checking for obfuscation techniques, domain reputation, and malicious redirections.

The results generated by these three modules are integrated through a meta-classifier, which consolidates individual predictions to enhance detection accuracy and reduce false positive rates. The proposed ensemble framework employs a stacked architecture that combines the outputs of three base classifiers: (i) Random Forest (RF), (ii) Naive Bayes (NB), and (iii) Support Vector Machine (SVM). In this framework, the predictions from these base models serve as input features for a Logistic Regression meta-classifier, which produces the final decision. The layered approach utilizes the complementary strengths of individual models while alleviating their limitations, thereby enhancing phishing email detection performance across varied datasets. This ensemble-based approach makes D-Fence a flexible, scalable, and computationally efficient system for mitigating phishing. Its modular structure allows organizations to optimize configurations based on their computational constraints without compromising effectiveness. By comprehensively analyzing multiple attack vectors, D-Fence provides a robust defense against evolving phishing threats, ensuring higher detection rates with reduced computational overhead.

The taxonomy of an email, as shown in Figure 2 (Almomani *et a*l., 2013), provides a structured representation of its core components, which is essential for analyzing and detecting phishing attempts. An email can be broadly divided into three main segments: the whole email message, the header, and the body. The whole message encompasses general characteristics such as message length, MIME structure, and attachment metadata, along with an unstructured set of words that reflect the overall content. The header contains both structured fields (e.g., *From*, *To*, *Subject*, *Return-Path*, *Message-ID*) and unstructured text, which are often exploited in spoofing and impersonation attacks. The body includes the primary content of the email, comprising natural language text, graphical elements (such as logos or buttons), embedded hyperlinks, and potentially obfuscated or adversarial elements. This taxonomy not only aids in feature extraction for phishing detection models but also supports a layered analytical approach to distinguish between legitimate and malicious messages.

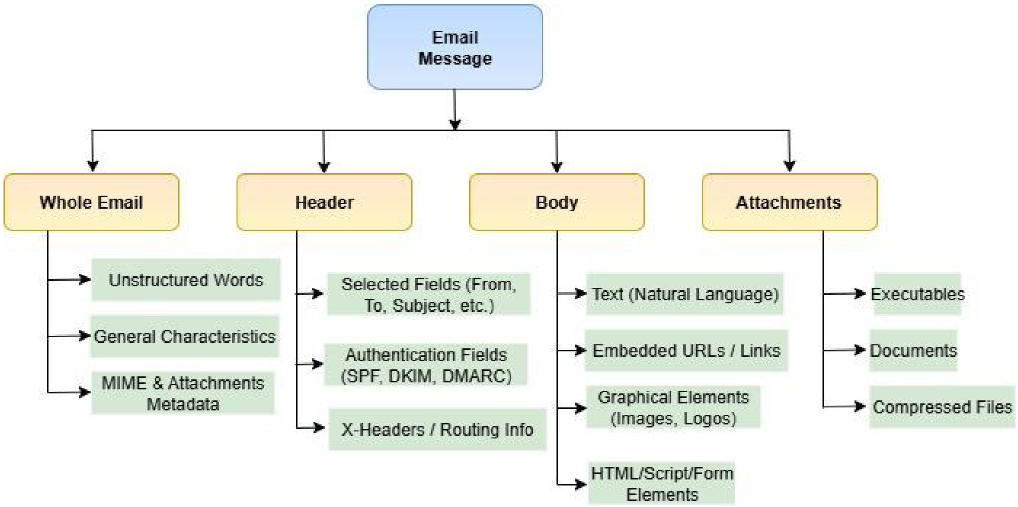


**FIGURE 2:** Email Taxonomy.

One effective technique for detecting phishing emails is to apply Natural Language Processing (NLP) to examine the email body's content. NLP techniques can identify suspicious patterns, such as urgent language, misspellings, impersonation attempts, and malicious URLs. Advanced models, including BERT (Lee *et al.*, 2020) and LSTMs (Siddique *et al.*, 2021), enhance phishing detection by recognizing contextual anomalies and deceptive phrasing. By leveraging NLP, phishing detection systems can improve accuracy and minimize false positives in email filtering.

Diverse machine learning based phishing email detection algorithms can be categorized into:

### 2.1.1 Supervised Approaches

**Support Vector Machines (SVM)** - Used for binary classification of phishing and legitimate emails. (Kumar *et al.*, 2020; Niu *et al.*, 2017; Siddique *et al.*, 2021).

**Random Forests (RF)** - An ensemble learning method that constructs multiple decision trees and combines their outputs to improve classification accuracy and reduce overfitting in phishing detection tasks. (Subasi *et al.*, 2017).

**Logistic Regression (LR)** - Statistical model that predicts the probability of an email being phishing. (Vinayakumar *et al.*, 2019).

**Decision Tree (DT)** - Classifies emails based on features like sender details, URLs, and content structure. (Ismail *et al.*, 2022).

**Naïve Bayes (NB)** - Probabilistic classifier based on Bayes' theorem, often used in spam and phishing detection. (Siddique *et al.*, 2021).

**k-Nearest Neighbors (KNN)** - A distance-based classification algorithm that assigns labels to emails according to the majority class among their k nearest neighbors in the feature space; it proves effective for basic phishing detection tasks. (Murugavel *et al*, 2020).

**Gradient Boosting (XGBoost)** - Advanced hybrid technique that constructs decision trees sequentially to minimize classification error; known for its efficiency and accuracy in phishing email detection. (Anitha *et al*, 2021).

### 2.1.2 Unsupervised Approaches

**K-Means Clustering** – Groups similar emails together based on features, identifying anomalies as potential phishing attempts. (Saka *et al.*, 2022).

**Hierarchical Clustering** - Builds a nested tree (dendrogram) of email clusters based on content similarity, useful for identifying phishing patterns without labeled data. (Karim *et al.,* 2020).

**TF-IDF (Term Frequency-Inverse Document Frequency)** - A statistical feature extraction approach that evaluates the importance of words in an email relative to a corpus, commonly used to represent email content for machine learning models. (Harikrishnan *et al*., 2018).

**TABLE 1:** Comparison of this survey with existing email phishing surveys.

| Survey | Channels Covered | LLM-based Phishing | Automation Levels | Defense Scope | Lifecycle View |
|---|---|---|---|---|---|
| Existing Survey A (Classical Email-focused Survey) | Email | No | No | Technical | No |
| Existing Survey B (ML-based Phishing Detection Survey) | Email, SMS | Partial | Partial | Technical | Yes |
| Existing Survey C (DL/NLP-based Survey) | Email, Social Media | Yes | No | Detection-only | No |
| This Survey | Email, SMS, Voice, Social, Cloud, IoT | Yes | Manual → LLM-based | Technical + Behavioral + Organizational | Yes |

### 2.1.3 Deep Learning Approaches

**Convolutional Neural Networks (CNNs**) - Detect localized structures in email text, such as phishing related phrases or structural cues; effective for learning spatial hierarchies in data. (McGinley and Monroy, 2021; Siddique *et al*., 2021).

**Recurrent Neural Networks (RNNs)** - Sequence-based neural models designed to capture temporal dependencies in email text; useful for modeling the order and flow of words in phishing messages. (John *et al*., 2022).

**Long Short-Term Memory networks (LSTMs)** - A specialized variant of the Recurrent Neural Network (RNN) designed to overcome vanishing gradient problems by retaining long-range dependencies, thereby enabling the detection of nuanced contextual cues within phishing content. (Li *et al*., 2020).

**Gated Recurrent Units (GRUs)** - A simplified type of LSTMs with fewer parameters, offering comparable performance in modeling sequential text data for phishing email classification. (Wanda, 2023).

**Bidirectional Encoder Representations from Transformers (BERT)** - A transformer based language model that learns contextualized embeddings by analyzing text bidirectionally, allowing for advanced phishing detection through semantic understanding and intent analysis. (Otieno *et al*., 2023).

Despite advancements in spam filters and anti-phishing tools (Jayatilaka *et al*., 2024), email phishing remains a significant issue due to its ability to exploit human psychology and the trust users place in digital communications. Addressing this challenge requires a combination of technical solutions, such as machine learning-based email filtering systems, and user educational initiatives to help individuals identify and report phishing attempts.

As summarized in Table 1, existing surveys primarily focus on isolated detection mechanisms, whereas this survey uniquely integrates attack automation, emerging AI-enabled threats, and multi-layered defense strategies.

### 2.2 Spear-Phishing

Spear-phishing (Stembert *et al*., 2015) is a targeted form of phishing that focuses on specific individuals or organizations, leveraging tailored information to make the attack more convincing and effective. Unlike generic phishing campaigns, which rely on casting a wide net, spear-phishing attacks are meticulously crafted based on detailed reconnaissance and intelligence gathering. Attackers often gather information from publicly available sources, such as social media profiles, corporate websites, and leaked databases, to personalize their messages. This personalization increases the likelihood that the target will trust the communication and take the desired action, such as clicking on a malicious link, downloading a compromised file, or providing confidential information.

A typical spear-phishing email may address the recipient by name, reference their role in the organization, and include contextual details that make the message appear authentic. For instance, an attacker targeting an employee in the finance department might pose as the CFO and request urgent approval for a wire transfer. The specificity and plausibility of these attacks make them particularly dangerous, as they often bypass traditional security measures and exploit human trust.

Spear-phishing attacks (Youvan and Douglas, 2024) are meticulously crafted to appear highly personalized and credible to the recipient. Attackers often address the target by name, reference specific details about their work or personal life, and make the request seem relevant to their role. For example, an email might have a statement as:

*"Hi John, this is Sarah from the IT department. We're updating security protocols for your team and require you to log in using the link below to complete the setup by 5:00 PM. Please let me know if you encounter any issues."*

This message appears more convincing because it uses John's name, impersonates a familiar colleague, and presents a request that aligns with his job responsibilities. By leveraging social engineering techniques, attackers make their phishing attempts seem legitimate, increasing the likelihood of compliance.

Spear-phishing is frequently used in high-stakes scenarios, such as corporate espionage, data theft, or the delivery of advanced persistent threats (APTs). High-value targets, including executives, government officials, and IT administrators, are often the focus of these attacks due to their access to sensitive information and systems. Mitigating spear-phishing requires a multi-layered approach (Arya and Chamotra, 2021), including robust email authentication protocols (Bojjagani *et al*., 2020), employee training (Burns *et al*., 2019) on recognizing suspicious communications, and the use of advanced threat detection systems. Additionally, fostering a culture of vigilance within organizations can significantly reduce the success rate of these highly personalized attacks.

### 2.3 Smishing and Vishing

Smishing, or SMS-based phishing (Kumarasinghe et al., 2023), refers to a social-engineering cyberattack in which adversaries deliver deceptive text messages to persuade recipients into divulging confidential information such as login credentials, credit-card numbers, or other personal data as shown in Figure 3. These messages often impersonate authentic communications from trusted entities, including banks, government agencies, and service providers, and frequently contain malicious URLs that redirect users to fraudulent websites designed to extract confidential information. Smishing campaigns exploit the inherent trust users place in mobile text messaging by manipulating psychological triggers such as urgency, fear, or reward, prompting victims to respond impulsively. Compared with conventional email based phishing, smishing poses a greater threat because mobile users tend to engage with links more impulsively, often without verifying authenticity.

Smishing (SMS Phishing) attacks (Mishra and Soni, 2020):

**Fake Security Alerts** – Attackers send fraudulent messages pretending to be from banks or online services, warning of unauthorized transactions and urging users to click on malicious links.

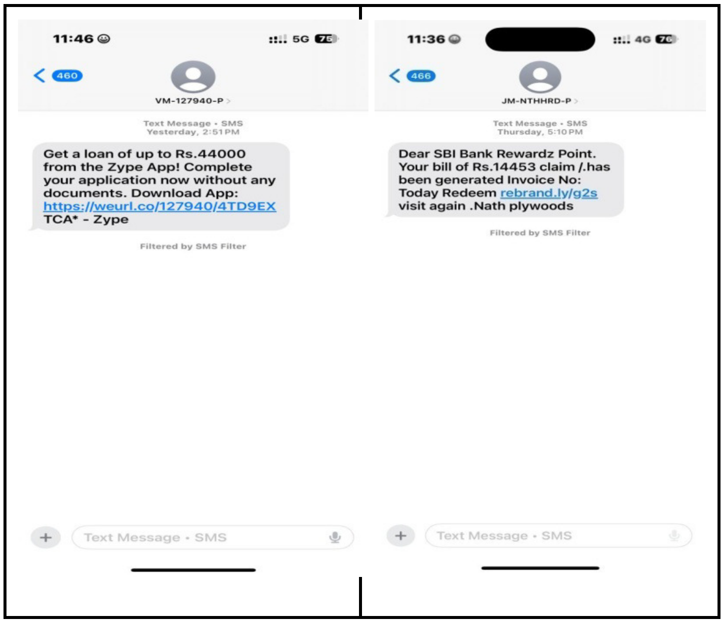
**Delivery Scams** – Messages claiming a package is awaiting delivery, requiring users to verify details or pay a small fee, leading to credential theft.

**Lottery or Prize Scams** – Victims receive SMS messages stating they've won a prize and must enter their details or make a payment to claim it.

**Banking Verification Requests** – Fraudulent messages impersonating financial institutions, requesting login details or OTPs for "security verification."

To mitigate smishing risks (Rahman et al., 2023), individuals should avoid clicking on unknown links, verify messages directly with the sender, and use mobile security tools that detect and block fraudulent SMS messages. Organizations can also improve the security by integrating multi-factor authentication (MFA) and educating users on recognizing and reporting suspicious texts.

Vishing (Kumarasinghe *et al.*, 2023), also known as voice phishing, is a deceptive attack where cybercriminals use phone calls to manipulate victims into revealing sensitive information, such as banking details, login credentials, or personal identification numbers, as shown in Figure 4. Adversaries frequently masquerade as legitimate entities by including financial institutions, technical support teams, or government bodies employing social-engineering strategies that evoke urgency or fear to manipulate victims. Common vishing techniques include spoofing caller IDs to appear legitimate, pre-recorded robocalls claiming fraudulent activity on an account, and direct social engineering where attackers pose as customer service representatives to extract confidential data. Unlike email phishing, vishing exploits verbal communication, making it harder for victims to verify authenticity.



VM-127940-P

Text Message • SMS
Yesterday, 2:51 PM

Get a loan of up to Rs.44000 from the Zype App! Complete your application now without any documents. Download App: https://weurl.co/127940/4TD9EX
TCA* - Zype

Filtered by SMS Filter

JM-NTHHRD-P

Text Message • SMS
Thursday, 5:10 PM

Dear SBI Bank Rewardz Point. Your bill of Rs.14453 claim /.has been generated Invoice No: Today Redeem rebrand.ly/g2s visit again .Nath plywoods

Filtered by SMS Filter

**FIGURE 3:** Smishing attack.

**Vishing (Voice Phishing) attacks:**
**Caller ID Spoofing** (Song *et al.*, 2014) – Attackers manipulate the caller ID to make it appear as if they are calling from a trusted entity like a bank or government agency.

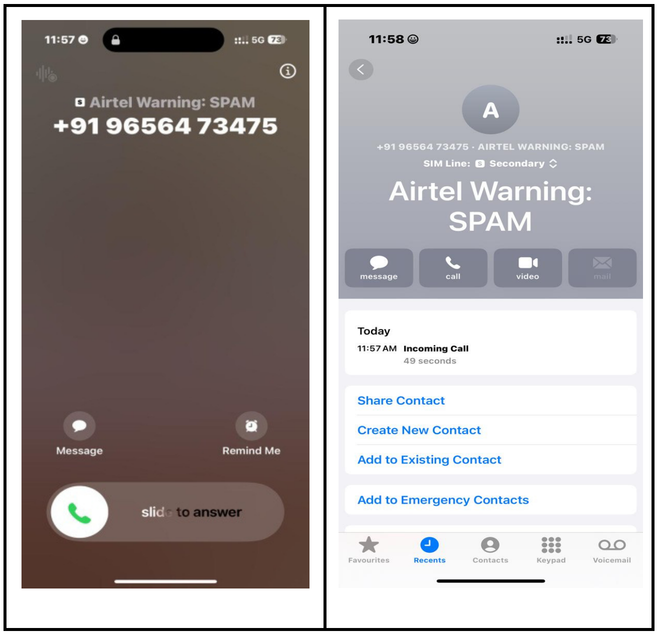
**Tech Support Scams** (Liu *et al.*, 2023) – Attackers impersonate IT personnel or customer support agents, persuading victims to provide remote access to their devices.

**Banking and Financial Fraud Calls** (Kale *et al.*, 2021) – Calls claiming unauthorized transactions or issues with bank accounts, pressuring victims into sharing confidential information.

**Government or Law Enforcement Impersonation** – According to the *India Today article (2024)*, attackers claim to be CBI, police, or immigration officers, threatening legal action unless payments or personal details are provided.

**Voice Deepfake Attacks** (Figueiredo *et al.*, 2024) – With the advancements in AI, vishing attacks are becoming more sophisticated, enabling automated bots to convincingly interact with victims. These AI-powered vishing system utilizes a Large Language Model (LLM) alongside with Voice-to-Text and Text-to-Voice Modules

To mitigate vishing threats (Phang *et al.*, 2024), individuals should be careful of unsolicited calls requesting sensitive information, avoid sharing personal details over the phone, and verify caller identities through official channels. Organizations can implement call authentication technologies and educate employees and customers on recognizing and reporting suspicious calls to reduce vishing risks.



11:57

Airtel Warning: SPAM

+91 96564 73475

11:58

+91 96564 73475 · AIRTEL WARNING: SPAM

SIM Line: Secondary

Airtel Warning: SPAM

Today

11:57 AM Incoming Call 49 seconds

**FIGURE 4:** Vishing attack.

### 2.4 Social Media and Cloud-Based Phishing

Social media phishing (Alharbi *et al*., 2022) exploits user trust by leveraging fake profiles, hijacked accounts, and deceptive messages to steal sensitive information. Attackers use tactics such as sending malicious links (Gupta and Singhal, 2017), posting fraudulent ads, and creating fake events to manipulate victims. With the increasing use of smartphones and social media platforms, users have become more vulnerable, as phishing attacks are now easily deployed through mobile applications. Attackers often disguise malicious apps as legitimate ones, tricking users into unknowingly providing their login credentials. The ease of impersonation on social media further amplifies the risk, making phishing a growing threat that requires heightened awareness and robust security measures. Attackers can trick users on social media platforms using a variety of methods:

**Impersonations** (Algarni et al., 2014) - In this method, attackers pose as trusted individuals or organizations, such as banks, executives, or government agencies, to trick victims into revealing sensitive information.

**Posting malicious links** (Gupta and Singhal, 2017) - Phishers distribute malicious links through emails, social media posts, and messaging platforms, luring users to fraudulent websites designed to steal credentials or install malware.

**Fake profiles** (Tiwari, 2017) - Attackers create fake profiles on social media and professional networking sites, impersonating real individuals to gain trust and manipulate victims. These fraudulent accounts often utilize stolen photos and fabricated credentials to entice users into phishing scams or social engineering attacks.

**Cloud-based phishing attacks** (Butt et al., 2023) exploit the trust and legitimacy of popular cloud services, such as Google Sites and Typeform, to host fraudulent web pages and deceive users into providing sensitive information. These attacks leverage reputable cloud domains and IP addresses, rendering traditional detection methods, such as IP reputation monitoring and blacklisting, less effective. Attackers exploit the widespread use of cloud-based applications (Jha et al., 2022) to create phishing sites that appear legitimate, thereby bypassing security filters and evading detection. As organizations increasingly rely on cloud platforms, mitigating these threats requires advanced security measures, including AI-driven anomaly detection, real-time monitoring, and user awareness training.

Below are some of the most common cloud-based phishing attacks:

**Cloud Service Spoofing** (Jha *et al*., 2020) – Attackers create fake login pages that mimic legitimate cloud platforms (e.g., Google Drive, Microsoft 365) to trick users into entering their credentials.

**Malicious Cloud Hosting** (Jha *et al*., 2020) – Cybercriminals utilize reputable cloud services, such as Google Sites, Dropbox, or Typeform, to host phishing pages, thereby bypassing traditional security mechanisms like blacklisting.

**Cloud-Based Email Phishing** (Butt *et al*., 2023) (Business Email Compromise - BEC) – Attackers compromise cloud-hosted email services (e.g., Outlook, Gmail) to send phishing emails from trusted domains, increasing their credibility.

**OAuth Token Phishing** (Xie *et al*., 2016) – Instead of stealing passwords, attackers trick users into granting malicious applications OAuth access to their cloud accounts, allowing unauthorized access without requiring login credentials.

**Cloud Storage Exploitation** (Keskisaari, 2022) – Phishers embed malware or phishing links in cloud-shared documents (e.g., Google Docs, OneDrive links), making it difficult for users to detect malicious intent.

**Man-in-the-Cloud (MitC) Attacks** (Sonawane *et al*., 2024) – Instead of stealing login credentials, attackers compromise synchronization tokens, allowing persistent access to cloud accounts without raising security alarms.

To mitigate these risks, users should be cautious of unsolicited requests and verify the authenticity of profiles and links they encounter. Organizations can enhance security by implementing advanced measures such as multi-factor authentication, regular account monitoring, and phishing-resistant protocols for cloud-based platforms. Additionally, educating users about the unique risks associated with social media and cloud services is essential in reducing their susceptibility to these evolving threats.

### 2.5 Clone Phishing

Clone phishing (Wong et al., 2022) is a sophisticated cyber attack where attackers duplicate legitimate websites or emails, making only minor modifications to deceive users into revealing sensitive information. Unlike traditional phishing, where attackers create entirely fake websites, clone phishing leverages real website components, such as HTML structure, CSS styles, and images, making detection more challenging. Attackers often host these cloned sites on cloud-based platforms, ensuring high availability and credibility. As highlighted in the PhishClone study (Chaudhuri, 2023), automated cloning techniques with slight modifications can effectively bypass machine-learning-based phishing detectors.

Based on the PhishClone study (Chaudhuri, 2023), attackers employ the following seven distinct cloning techniques:

**JavaScript-Based Cloning** (Song et al., 2025) - Attackers use JavaScript to dynamically fetch and load legitimate website content, making the phishing page appear identical to the original in real-time. This technique helps evade static detection mechanisms.

**Direct Copying (HTML & CSS Duplication)** (Soyemi and Isinkaye, 2017) - The attacker downloads the HTML, CSS, and JavaScript files from a legitimate website and re-hosts them on a phishing domain with slight modifications, such as altering login forms to capture user credentials.

**CAPTCHA Manipulation** (Gelernter and Herzberg, 2016) - To bypass bot detection and security defenses, some phishing pages include fake or stolen CAPTCHA mechanisms to create a false sense of legitimacy and trick users into believing the site is secure.

**Hardcoded Data Injection** (Ray and Ligatti, 2012) - Instead of dynamically fetching content from the legitimate source, attackers hardcode the website structure and design, embedding fraudulent login fields that steal user credentials upon submission.

**URL Obfuscation and Redirection** (Skula and Kvet, 2024) - Cloned phishing pages often use techniques such as URL shorteners, homoglyph domains (e.g., replacing "microsoft.com" with "rnicrosoft.com"), or multiple redirections to evade detection.

**Third-Party Resource Loading** (Ikram *et al.*, 2020) - Attackers retain references to legitimate external resources (e.g., images, fonts, stylesheets) from the original website while modifying only interactive elements like login forms. This makes phishing pages look authentic while executing malicious actions.

**Fully Customized Clones** (Chaudhuri, 2023) - Instead of copying a legitimate website entirely, attackers create a highly customized version that mimics the design but introduces deceptive elements, such as fake customer support chatbots or altered security messages, to manipulate users.

Mitigating clone phishing requires a combination of email security measures, website protection, AI-powered detection, and user awareness. Organizations should implement Domain-based Message Authentication, Reporting, and Conformance (DMARC) (Kucherawy and Zwicky, 2015), Sender Policy Framework (SPF) (Wong and Schlitt, 2006), and Domain Keys Identified Mail (DKIM) (Crocker *et al.*, 2011) to prevent email spoofing and use AI-driven phishing filters to detect suspicious emails. Users must verify SSL certificates, inspect URLs for homoglyph attacks, and avoid clicking on untrusted links. Advanced security measures, such as visual similarity analysis and real-time behavior tracking, can help detect cloned phishing pages. Websites can employ anti-cloning techniques, such as JavaScript-based integrity checks and dynamic content rendering, to prevent unauthorized duplication. Additionally, organizations should encourage users to report phishing sites, work with hosting providers to take down fraudulent domains, and conduct regular cybersecurity training to enhance awareness and response to clone phishing threats.

## 3. PHISHING ATTACK LIFECYCLE

The lifecycle of a phishing attack consists of six interrelated stages, each strategically designed to manipulate victims and achieve the attacker’s objectives. Understanding these stages is critical for developing comprehensive defensive mechanisms and proactive threat mitigation strategies. The stages are Planning and Reconnaissance, Bait Creation, Delivery, Exploitation and Execution, Command and Control, Post-Exploitation, and Exit/Evasion. These represent the full chain of a phishing campaign, from initial target selection to data theft and cover-up. Figure 5 illustrates the phishing attack lifecycle, depicting each stage from initial reconnaissance to post-exploitation and system compromise.

### 3.1 Planning and Reconnaissance

The first phase is planning and reconnaissance (Kaur and Mian, 2023; Alkhalil *et al.*, 2021), which lays the groundwork for the entire attack. Here, the attackers identify and profile potential targets, leveraging open-source intelligence (OSINT) such as company websites, LinkedIn profiles, social media activity, and publicly available datasets. The goal is to gather detailed information, such as job roles, communication patterns, hierarchical relationships, and email formats, to craft tailored phishing content. In spear-phishing attacks targeting financial institutions, attackers may focus on high-value individuals such as finance officers or HR staff. For example, a spear phishing campaign may begin with harvesting details from LinkedIn and company newsletters to identify internal finance team members and organizational structure (Youvan and Douglas, 2024).

### 3.2 Bait Creation and Weaponization

Once sufficient reconnaissance is performed, attackers move to bait creation, where phishing messages are carefully crafted to appear legitimate. These may include spoofed emails mimicking senior executives or popular service providers, fake websites replicating login portals, or malicious attachments embedded with malware. The messages often employ psychological triggers such as urgency, fear, or reward. In CEO fraud scenarios, attackers impersonate executives using stolen branding and insider jargon to request urgent wire transfers from finance teams (Mansfield, 2016). The credibility of such baits is heightened by mimicking official email domains and using internal terms acquired during reconnaissance.

### 3.3 Delivery of the Phishing Message

In this stage, attackers deliver the phishing message via channels like voice calls, email, SMS, or messaging apps. Email remains the most common delivery vector, aided by domain spoofing and email header manipulation. Attackers may send mass emails or target specific individuals depending on the campaign's scope. According to *AAG IT (2025)* phishing remains the most prevalent form of cybercrime, with over 3.4 billion spam emails sent daily, and Google blocking 100 million phishing emails per day. In 2022, phishing was the primary attack method reported by 83% of UK businesses, and more than 20% of phishing emails originated from Russia, underscoring its global pervasiveness.

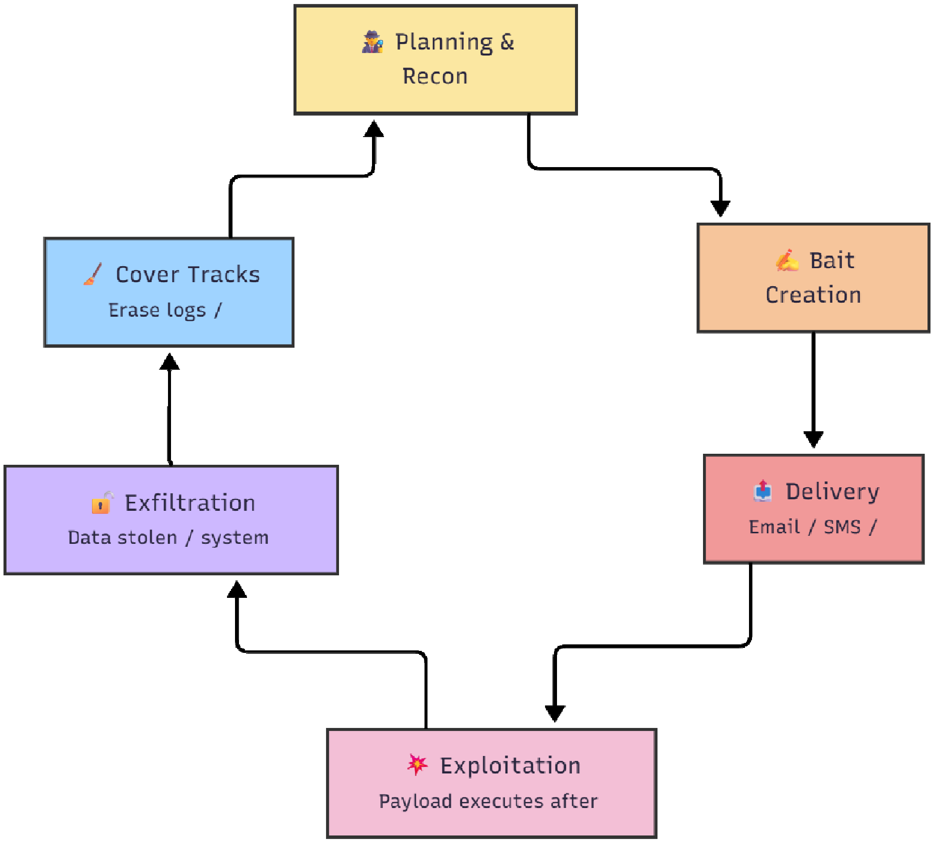


**FIGURE 5:** Phishing Attack Lifecycle.

### 3.4 Exploitation and Execution

The exploitation and execution phase begins once the victim interacts with the bait by clicking a malicious link, downloading an infected attachment, or entering credentials on a spoofed site. Attackers exploit these actions to install malware, harvest credentials, or gain unauthorized system access (Alkhalil *et al.*, 2021). In the earlier spear-phishing case, once the finance team receives the spoofed CEO email, they are pressured into processing a wire transfer without verification. This email often exploits urgency and authority, prompting uncritical action. Real-world examples show how such emails result in fraudulent financial transactions, with some whaling attacks costing up to $47 million per incident (Mansfield, 2016).

### 3.5 Data Exfiltration or System Compromise

Following the initial exploitation, attackers aim to deepen their foothold by exfiltrating data and compromising systems. In advanced phishing attacks involving malware, the compromised system that usually initiates a connection with an external command-and-control (C2) server, which allows attackers to issue remote commands, extract sensitive data, and maintain persistent access within the victim's environment. The C2 infrastructure enables silent operations such as credential harvesting, keystroke logging, and lateral movement across the network. These communications are often encrypted or obfuscated, making them difficult to detect through traditional firewalls or signature-based intrusion detection systems. Once inside the system, attackers engage in post-exploitation activities, such as exfiltrating confidential documents, executing fraudulent transactions, or monetizing stolen credentials through underground markets. In the case of CEO fraud, for example, attackers may swiftly transfer stolen funds across multiple accounts to obscure the transaction trail, or exploit internal access to escalate privileges, deploy ransomware, or conduct additional phishing campaigns. These post-exploitation actions can remain undetected for extended periods, resulting in severe financial losses, reputational damage, and regulatory consequences for the targeted organization (Kaur and Mian, 2023; Alkhalil *et al*., 2021; Mansfield, 2016).

### 3.6 Exit and Covering Tracks

Finally, the attackers attempt to cover their tracks to avoid detection and slow down the investigation. This may involve deleting phishing emails, clearing logs, or disabling alert systems. Some attacks are designed to leave no trace, while others plant backdoors for future access. In cases where ransomware is deployed, attackers may demand payment before restoring access, leaving organizations with both operational and financial loss. The average cost of a phishing-induced data breach now exceeds $4 million per organization, making this final stage especially devastating if not detected early (Sendmarc, 2023).

## 4. EMERGING TRENDS IN PHISHING

Phishing attacks are evolving rapidly, driven by advancements in technology and shifts in user behavior. According to Sharon and Ashwin (2024)*,* one of the most noteworthy trends is the use of deep learning (DL) and machine learning (ML)  to create more sophisticated and personalized phishing campaigns. Attackers leverage deep learning techniques to craft highly convincing emails, messages, or websites that mimic legitimate entities with alarming accuracy. Deepfake technology is also being integrated, enabling attackers to produce fake voice or video messages to deceive targets into divulging sensitive information.

Additionally, the proliferation of cloud services and the shift to remote work have introduced new avenues for attack. Cybercriminals increasingly exploit cloud-based platforms, such as email services and file-sharing tools, to target employees working remotely. Social media remains a prominent platform for phishing, as attackers use fake profiles and targeted ads to lure victims. Another emerging trend is the rise of phishing-as-a-service (PhaaS), as reported by Resecurity (2022) and Ilca and Balan (2021), where attackers provide turnkey phishing kits and services on the dark web, enabling less skilled individuals to launch sophisticated phishing campaigns.

These evolving tactics underscore the necessity for ongoing innovation in detection and prevention technologies, such as AI-driven anomaly detection (Naseer, 2024), user education (Dodge *et al*., 2007), and enhanced authentication mechanisms, to remain ahead of increasingly complex phishing schemes.

### 4.1 Exploitation of Large Language Models in Phishing

A significant emerging trend in the phishing landscape is the misuse of Large Language Models (LLMs), such as GPT-4 and ChatGPT, and other generative AI tools to automate and scale phishing attacks. These models produce highly persuasive, grammatically polished, and context-aware phishing messages that closely resemble legitimate communications from trusted organizations, thereby markedly raising user engagement risk. Unlike traditional phishing, which often suffers from linguistic or formatting errors, LLM-generated messages can be customized at scale for spear-phishing, business email compromise (BEC), and credential harvesting (Hazell, 2023). Recent studies have demonstrated how adversaries leverage LLMs not only to craft phishing emails but also to evade detection systems through adversarial paraphrasing and real-time content adaptation (Roy *et al.*, 2024). Additionally, malicious variants such as WormGPT and FraudGPT have emerged on underground forums, allowing non-technical users to generate phishing content and malware payloads with minimal effort (Falade, 2023). These tools often bypass ethical safeguards and are optimized for crafting malicious content, allowing even novice attackers to initiate sophisticated phishing campaigns.

Moreover, researchers have identified that LLMs can be manipulated through prompt injection, a form of secondary input manipulation where attackers embed hidden instructions within emails or web content to trigger unauthorized behaviors in AI-enabled systems such as smart inbox summarizers or virtual assistants (Sha and Zhang, 2024; Schmitt and Flechais, 2024). This form of exploitation raises new concerns about the role of LLMs not only in generating phishing content but also in interpreting and presenting such content in a deceptive way. The combination of human-like language generation and the ability to dynamically adapt to the victim's context makes LLM-powered phishing particularly challenging to detect using conventional rule-based or statistical phishing filters. As these models continue to evolve and become more accessible, they are expected to play an increasingly central role in phishing attacks, making the development of robust AI-aware defensive mechanisms a top priority for cybersecurity researchers and practitioners.

### 4.2 Advanced Persistent Phishing Campaigns

Advanced Persistent Phishing Campaigns (APPCs) (Brandao *et al.*, 2022) are highly advanced and sustained attacks designed to target specific individuals or organizations over an extended period. Unlike traditional phishing attempts, these campaigns employ advanced tactics (Sharma et al., 2023), including social engineering, targeted spear-phishing emails, and multi-channel strategies such as smishing and vishing, to increase the chances of success. Attackers often conduct extensive reconnaissance to understand the target's vulnerabilities and craft customized messages that align with their behaviors, preferences, or organizational roles.

These campaigns often lead to unauthorized system access, allowing attackers to maintain persistence within networks and steal sensitive data such as financial records, intellectual property, or government secrets (Krishnapriya and Singh, 2024). Organizations face significant financial losses due to fraudulent transactions, wire transfer scams, and regulatory penalties, while reputational damage erodes customer trust and undermines business credibility. Additionally, APPCs can serve as entry points for supply chain attacks, enabling cybercriminals to infiltrate multiple organizations through compromised partners. In some cases, attackers disrupt critical infrastructure, impacting essential services in finance, healthcare, and government sectors. Moreover, the stolen credentials and compromised access points are often sold on the dark web, leading to further cyberattacks. Due to the sophisticated and persistent nature of these campaigns, organizations must implement continuous monitoring, AI-driven threat detection, and strong cybersecurity policies to mitigate their devastating impact.

### 4.3 Phishing Targeting IoT and Cloud Systems

Phishing attacks targeting Internet of Things (IoT) (Nirmal *et al.*, 2021) devices have become a growing concern due to the widespread use of IoT in smart homes, healthcare, industrial control systems, and critical infrastructure. Unlike traditional phishing, IoT phishing exploits the unique vulnerabilities of interconnected devices that often lack robust security measures. Attackers exploit the vulnerabilities inherent in IoT devices, such as weak authentication mechanisms or outdated firmware, to gain unauthorized access to interconnected systems. By compromising an IoT device, attackers can infiltrate larger networks or use the device as a gateway to exfiltrate sensitive data.

The consequences of IoT Phishing are:

**Device Hijacking** (Zhou et al., 2019) - Attackers take control of IoT devices, leading to unauthorized surveillance, data theft, or botnet. Example for this kind of attack is Mirai Botnet (Antonakakis et al., 2017). Mirai Botnet is a highly dangerous IoT-based botnets, first discovered in 2016, which exploited weak security in internet-connected devices. Mirai primarily targeted Internet of Things (IoT) devices including routers, IP cameras, and DVRs that were secured with default or weak login credentials. The malware scanned the internet to identify vulnerable devices, subsequently infecting them and converting them into remotely controlled “zombie” nodes within its botnet network.

**Data Breaches** (Pan and Yang, 2018) - Stolen IoT credentials enable access to sensitive data stored on smart devices and cloud services. Attackers can exploit these credentials to manipulate device settings, intercept communications, or exfiltrate confidential data. In critical environments, such as healthcare or industrial IoT, these intrusions can lead to operational disturbances, financial losses, or even safety hazards.

**Denial of Service (DoS) Attacks** (Salim *et al*., 2020) - Compromised IoT devices can be used in large-scale DDoS attacks, disrupting services.

**Physical Security Risks** - According to *Microsoft (2022),* securing IoT devices is crucial, as they often control physical systems (Ding and Hu, 2018) (e.g., smart locks, surveillance cameras). Unauthorized access can lead to physical security breaches, such as unauthorized entry into premises or tampering with critical infrastructure. Attackers could disable security alarms, unlock doors remotely, or manipulate surveillance footage to evade detection. In critical sectors such as energy grid operations (Cardenas *et al*., 2020), healthcare, or transportation, such breaches can pose serious safety risks, endangering lives and disrupting essential operations.

**Ransomware on IoT Devices** (Yaqoob et al., 2017) - A ransomware attack on IoT devices can be particularly destructive, as attackers can encrypt IoT devices or lock users out until a ransom is paid, compromising the entire security framework, including confidentiality, integrity, and availability. This not only leads to financial losses but also exposes critical information to potential breaches. Ransomware can seize full control of data or systems, restricting user access and demanding a significant ransom for data retrieval. If the victim refuses to pay, attackers may escalate the ransom amount, extend payment deadlines, or ultimately erase the data from the infected devices.

**In cloud environments** (Alotaibi et al., 2025), phishing attacks primarily focus on credential theft, granting attackers unauthorized access to critical cloud resources. This may result in service disruptions, data manipulation, or serve as a gateway for further cyberattacks. Threat actors often impersonate legitimate cloud service providers, using deceptive emails or fake login portals to trick users into revealing their authentication details. While cloud-based attacks are discussed in Section 2.4, it is crucial to emphasize the role of phishing in compromising cloud security. As IoT and cloud technologies become increasingly integrated, the deployment of robust security mechanisms including multi-factor authentication, secure firmware updates, and user awareness initiatives remains critical for mitigating emerging threats.

## 5. PHISHING DETECTION APPROACHES

Phishing detection represents a critical component of cybersecurity, focused on identifying and neutralizing deceptive attempts to obtain sensitive information such as login credentials, financial records, or personal data. Adversaries frequently impersonate legitimate organizations through emails, counterfeit websites, or malicious messages designed to deceive users into clicking harmful links or downloading infected attachments. To counter these threats, phishing detection employs various techniques, including rule-based filtering (Basnet *et al*., 2011), machine learning models (Selamat *et al*., 2020), natural language processing (NLP), and deep learning algorithms. URL (Raj *et al*., 2024) analysis is a common method used by phishing detection systems, which examines domain age, HTTP/HTTPS usage, URL obfuscation, and lexical patterns to determine the legitimacy of URLs. Email-based detection analyzes sender information, email headers, and embedded links, while website similarity analysis compares phishing pages with authentic websites using image processing and HTML structure matching. Advanced artificial intelligence (AI)-powered detection systems continuously learn from new phishing patterns, enhancing real-time protection against evolving threats. Additionally, behavioral analytics monitors user interactions and anomalies in web traffic to detect suspicious activities. Through the integration of these techniques, phishing detection systems serve a crucial function in safeguarding individuals and organizations against identity theft, financial fraud, and a broad spectrum of cyberattacks.

### 5.1 Rule-Based Phishing Attack Detection

Rule-based phishing attack detection (Basnet *et al*., 2011; Ramanathan and Wechsler, 2013) relies on predefined heuristics to identify phishing attempts by analyzing domain reputation, suspicious keywords, email authentication protocols, and URL structures (Gualberto *et al*., 2020). It efficiently detects known threats but struggles with novel phishing techniques. While computationally efficient, its accuracy improves when combined with machine learning and threat intelligence. Studies suggest that integrating rule-based methods with advanced security measures enhances real-time phishing detection and reduces the number of false positives.

The rule-based approach aims to simplify phishing attack detection by making it intuitive, straightforward, and user-friendly. This method involves defining specific rules that help determine whether a given webpage is fraudulent. Each rule follows a basic structure:

*IF conditions THEN actions*

*If the conditions, also known as patterns, are satisfied, then the actions of that particular rule are fired.*

These rules can range from simple checks, such as verifying a specific value in a URL, to more complex analyses that require examining metadata, querying search engines and blacklists, and combining multiple conditions using logical operators like AND and OR.

Based on their characteristics and the techniques used to derive them, rules can be broadly classified into the following categories: Search Engine-Based Rules, Red-Flagged Keyword Rules, Obfuscation-Based Rules, Blacklist-Based Rules, Reputation-Based Rules, and Content-Based Rules.

### 5.2 Machine Learning and AI-Based Detection

Machine Learning (ML) (Salahdine *et al*., 2021) and Artificial Intelligence (AI) (Bauskar et al., 2024) have significantly advanced phishing attack detection by enabling automated, adaptive, and highly accurate threat identification. ML algorithms analyze patterns in email content, URLs, metadata, and user behavior to differentiate between legitimate and phishing attempts.

Supervised learning models effectively classify phishing attempts based on labeled datasets, achieving high accuracy in recognizing known attack patterns. The  supervised models commonly used are:

**Decision Trees** (Alam *et al*., 2020; Zhu *et al*., 2020): Utilize a tree-like structure to classify phishing attempts based on predefined rules, including suspicious keywords, domain reputation, and link analysis.

**Random Forests** (Akinyelu and Adewumi, 2014; Subasi *et al*., 2017): An ensemble of multiple decision trees that enhances accuracy and reduces overfitting by averaging multiple predictions.

**Support Vector Machines (SVMs)** (Niu *et al*., 2017; Zouina and Outtaj, 2017): SVMs are effective in phishing classification by mapping data points into a high-dimensional space and finding the optimal decision boundary. Particularly useful for detecting subtle phishing patterns in email headers, URLs, and text content.

**Naïve Bayes Classifier** (Zhang and Li, 2007; Rusland *et al*., 2017): a probabilistic model widely used for phishing detection, especially in email filtering. Multinomial Naïve Bayes: This approach works well for text-based phishing detection by analyzing word frequencies. Bernoulli Naïve Bayes: Effective for binary classification (phishing vs. non-phishing) based on the presence or absence of certain features.

**Logistic Regression** (Vajrobol *et al*., 2024): A statistical model that predicts phishing probability based on extracted features like suspicious links, urgency in language, and sender credibility. Works well for lightweight, real-time phishing detection.

**Gradient Boosting Algorithms** (Omari, 2023), XGBoost, and CatBoost (Sadaf, 2023) : Advanced ensemble learning techniques that optimize phishing detection by handling large datasets efficiently and improving classification accuracy.

**Latent Dirichlet Allocation** (Ramanathan and Wechsler, 2013; Gualberto *et al*., 2020) : LDA is useful for analyzing phishing emails and URLs by identifying hidden topics and deceptive themes. Helps detect phishing attempts based on the distribution of unnatural or misleading content.

Unsupervised learning models are crucial in phishing detection as they can identify new and evolving phishing attacks without relying on labeled data. These models analyze patterns, detect anomalies, and cluster data points to distinguish phishing from legitimate activities. Unsupervised models are essential for detecting zero-day phishing attacks and evolving threats that may not be present in labeled datasets. Below are some widely used unsupervised learning techniques for phishing detection:

**K-Means Clustering** (Wanawe *et al*., 2014): Groups phishing and legitimate data based on feature similarity (e.g., email structure, word usage, sender reputation). Phishing emails and URLs often form distinct clusters due to their unusual characteristics. According to Bi *et al*. (2012), K-Means clustering performs relatively better than DBSCAN.

**DBSCAN (Density-Based Spatial Clustering of Applications with Noise**) (Bi *et al*., 2012): Identifies phishing emails and websites as anomalies based on density variations. Effective in detecting phishing attempts embedded within large datasets.

**Hierarchical Clustering** (Murtagh and Contreras, 2012): This method builds a tree-like structure to organize emails or URLs, making it easier to detect phishing patterns.

**Isolation Forest** (Liu *et al*., 2008): Detects phishing by isolating outliers (e.g., suspicious domains, unusual email wording, sudden changes in email patterns). Works well in real-time detection systems.

**PageRank Algorithm** (Sunil *et al*., 2012): The PageRank algorithm, originally developed by Google founders Larry Page and Sergey Brin, is a link analysis algorithm (Brin and Page, 1998) used to rank webpages based on their importance. While its primary application is in web search ranking, PageRank is also highly effective in phishing detection, particularly in identifying malicious websites and deceptive URL structures. The basic assumption is:

- Legitimate websites receive more inbound links from reputable sources.
- Phishing websites often have fewer or lower-quality inbound links, as they are newly created or blacklisted.

- Let page A have pages T1...Tn which point to it (i.e., are citations). The parameter d is a damping factor that can be set between 0 and 1. We usually set d to 0.85. Also, C(A) is defined as the number of links going out of page A. The PageRank score for a webpage is calculated as:

$$\mathbf{PR\ (A) = (1\text{-}d) + d\ (PR(T1)/C(T1) + \ldots + PR(Tn)/C(Tn))}$$

### 5.3 Graph Neural Networks (GNNs)

Unlike traditional machine learning models that focus on isolated features, GNNs (Ouyang and Zhang, 2021) model phishing attacks as structured graphs where nodes represent entities (e.g., emails, IPs, domains) and edges capture their interactions. Popular GNN architectures, such as Graph Convolutional Networks (GCNs) (Ariyadasa *et al*., 2022) and Graph Attention Networks (GATs) (Zhao *et al*., 2020), effectively learn from these relationships, enabling the detection of malicious activities based on link structures and behavioral patterns. GNNs excel in identifying zero-day phishing threats by detecting anomalies in graph structures, making them useful for phishing email detection, malicious domain identification, and social media phishing mitigation. These models provide enhanced accuracy and adaptability by capturing hidden patterns, reducing reliance on labeled data, and offering real-time threat analysis. However, challenges such as high computational costs, graph construction complexity, and adversarial manipulations require further research to optimize GNN-based phishing detection systems.

AI further enhances phishing detection (Eze and Shamir, 2024) through techniques such as Natural Language Processing (NLP) for the semantic analysis of email content and reinforcement learning to improve detection over time. Deep Learning models (Bauskar *et al*., 2024), such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), are particularly effective in analyzing embedded URLs, visual similarities in spoofed websites, and contextual relationships within phishing emails. The integration of AI-driven detection with real-time response mechanisms allows for proactive mitigation, such as blocking malicious links or alerting users to potential threats. However, as attackers leverage AI to create more sophisticated phishing schemes, continuous innovation in detection mechanisms remains critical to staying ahead of emerging threats.

### 5.4 Natural Language Processing for Phishing Identification

Natural Language Processing (NLP) (Zalavadia et al.,2019; Somesha and Pais, 2024) has emerged as a crucial technology in combating phishing attacks by enabling intelligent systems to interpret and analyze textual content in emails, messages, websites, and social media communications. Phishing attempts typically exploit psychological triggers - such as urgency, authority, fear, and curiosity - to manipulate users into disclosing sensitive information. NLP-based techniques are adept at detecting such linguistic manipulation through syntactic, semantic, and contextual analysis.

#### Text Representation Techniques

At the basic level, traditional text representation models are used to convert unstructured text into structured data:

**Bag of Words (BoW)** (Jain *et al*., 2020): Represents text by word frequency, useful for capturing general term usage.

**Term Frequency-Inverse Document Frequency (TF-IDF)** (Ramos, 2003): Weighs words based on their importance across documents, helping reduce the impact of commonly used terms.

**Word Embeddings (Word2Vec, GloVe)** (Mikolov, 2013) [28]: Capture semantic relationships between words, enabling models to understand word similarity and context.

These methods serve as inputs for machine learning classifiers to distinguish between phishing and legitimate content.

**Deep Learning-Based Contextual Analysis**

Latest phishing detection systems utilize deep learning models for understanding context and sequence:

Recurrent Neural Networks (RNNs) (Feng and Yue, 2020) and LSTMs (Su, 2020): Capture temporal dependencies and identify patterns in sequences, useful for detecting suspicious phrasing.

Transformers (BERT (Devlin et al., 2019; Haynes et al., 2021), RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2019)): Leverage attention mechanisms to interpret context bidirectionally, providing robust language understanding.

These models are effective in detecting subtleties, such as tone shifts, unnatural syntax, and hidden intent, within phishing messages.

Panagiotis Bountakas (Bountakas et al., 2021) conducted a comparative study to evaluate various combinations of Natural Language Processing (NLP) and Machine Learning (ML) techniques for detecting phishing emails. Their approach involved extracting textual features from the body of emails using three different NLP methods: TF-IDF, Word2Vec, and BERT, thereby creating distinct feature sets for analysis. The experimental results indicated that Word2Vec outperformed the other methods, making it the most effective NLP technique for ML-based phishing detection when focusing on email body text.

To gain a comprehensive understanding of advancements in phishing email detection, we review a collection of scholarly articles that support machine learning and deep learning techniques. These studies employ diverse methodologies and utilize various feature sets, including email headers, bodies, embedded links, and metadata, and are validated using both public and custom datasets. Table 2 presents a consolidated summary of these notable research works, systematically categorized by the methods used, features extracted, datasets employed, and accuracy achieved. This overview not only highlights effective detection strategies and emerging trends but also reveals potential gaps for future research in the field of phishing email detection.

**TABLE 2:** Summary of ML/DL based approaches for phishing email detection.

| Authors | Method(s) used | Features | Dataset(s) | Accuracy (%) |
|---|---|---|---|---|
| Fang *et al.*, 2019 | RCNN, Attention | Word, Character embeddings | Multiple datasets | 99.84 |
| Vinayakumar *et al.*, 2018 | CNN/RNN/LSTM/ MLP | Optimal features | IWSPA-AP 2018 | 99.1<br>(LSTM) |
| Hiransha *et al.*, 2018 | CNN | Body of the email | IWSPA-AP 2018 | 96.8 |
| Castilo *et al.*, 2020 | CNN/LSTM | Body of the email | Enron, APWG, Private | 95.68 |

| | | | | |
|---|---|---|---|---|
| Alhogail and Alsabih, 2021 | Graph Convolutional Networks | Body of the email | Fraud dataset | 98.2 |
| Qi Li *et al.*, 2020 | LSTM/KNN/ K-means | Header, Content | Private | 95 |
| Somesha and Pais, 2022 | RF/SVM/ DT/ LR/XGBoost + CBOW | Four email header features | SpamAssassin, PhishTank, Private | 99.5 |
| Butt *et al.*, 2023 | NB/RF/RNN | Header, Body | UCI Email Dataset | 97.3<br>(RNN) |
| Doshi *et al.*, 2023 | CNN/RNN/ANN | Header, Body | SpamAssassin, Phishing Corpus | 99.5<br>(ANN) |
| Ali and Abdullah, 2025 | CNN/XG-BOOST | Email Body | Enron Spam Dataset | 99 (CNN + XG-BOOST) |

Although transformer-based and deep learning models report higher accuracy values in phishing detection tasks, these results must be interpreted with caution. Many studies rely on curated or imbalanced datasets, which may inflate performance metrics and limit generalization to real-world environments. Additionally, issues such as overfitting, dataset dependency, and adversarial manipulation remain underexplored. From a deployment perspective, computational complexity, inference latency, and explainability constraints pose significant challenges, particularly for resource-constrained or real-time systems.

### 5.5 Anti-Phishing Toolkits and Frameworks

Anti-phishing toolkits and frameworks are vital in detecting and mitigating phishing attacks, providing organizations with robust solutions to safeguard against evolving threats. These tools combine URL analysis, email filtering, domain monitoring, and real-time threat intelligence to effectively prevent phishing attempts. Widely used tools, such as OpenPhish (Bell and Komisarczuk, 2020), were also utilized to obtain phishing data. Additionally, phishing data were obtained from PhishTank (Bell and Komisarczuk, 2020). Further phishing trends are reported by PhishStats (2024), and the eCrime data were sourced via Anti-Phishing Working Group (APWG) eCrime Exchange (APWG eCX, 2024), which maintains comprehensive databases of malicious URLs and domains, offering real-time blocking capabilities. Additionally, frameworks such as Google Safe Browsing (Bell and Komisarczuk, 2020), Microsoft Defender for Office 365, and Cisco Umbrella (Akiyama *et al.*, 2024; Cisco, 2024) utilize Machine Learning (ML) and Deep Learning (DL) for analyzing email content, URLs, SSL certificates, and user behavior to detect suspicious activity.

Other advanced frameworks include *Apache SpamAssassin (2024)*, which filters phishing emails using rule-based detection, and Kaspersky Anti-Phishing, which includes browser plug-ins for real-time website analysis. Tools like PhishNet (Diviya et al., 2024) and *PhishingBox (2024)* focus on proactive defense through simulated phishing campaigns to train employees and improve awareness. Moreover, *KnowBe4 (2024)* integrates employee training with phishing threat reporting, while ZScaler Internet Access and Barracuda Sentinel offer cloud-based solutions for real-time phishing threat analysis. The integration of these diverse toolkits and frameworks into an organization's cybersecurity infrastructure ensures layered defense mechanisms, significantly reducing exposure to phishing attacks and enhancing overall resilience against cyber threats.

While the reviewed studies demonstrate steady improvements in phishing detection accuracy, notable trade-offs emerge across techniques. Rule-based and classical machine learning methods offer interpretability and low computational overhead but struggle with zero-day and obfuscated phishing attacks. Deep learning and transformer-based approaches exhibit stronger generalization capabilities; however, they often rely on large labeled datasets and suffer from explainability and deployment challenges. Furthermore, variations in datasets, feature representations, and evaluation metrics limit direct comparability across studies, highlighting the need for standardized benchmarking and real-world validation.

## 6. PREVENTIVE MEASURES AND BEST PRACTICES

Effective defense against phishing attacks requires a multi-layered approach that combines technological solutions, user awareness, and robust organizational policies. Multi-Factor Authentication (MFA) (Muir *et al*., 2024) serves as a critical control by requiring users to verify their identity using multiple methods, thus reducing the risk of account compromise even when credentials are exposed. Simultaneously, user education initiatives (Dodge *et al*., 2007) build awareness of phishing tactics such as deceptive emails, malicious links, and urgent requests, empowering users to detect and avoid threats.

At the organizational level, bringing out secure email gateways, access controls, and real-time network monitoring (Frauenstein and Solms, 2014) reduces exposure to phishing vectors. Enforcing well-documented security policies, conducting routine training, and ensuring prompt incident reporting strengthen the organization’s overall security posture. This comprehensive strategy minimizes vulnerabilities and supports rapid threat response, enhancing resilience against phishing attempts.

### 6.1 Multi-Factor Authentication

Multi-Factor Authentication (MFA) (Muir *et al*., 2024) is a keystone of phishing defense, introducing layered verification that significantly strengthens access security. By requiring two or more factors, typically a password (knowledge), a device or token (possession), and biometric input (inherence), MFA renders stolen credentials ineffective on their own. This effectively mitigates risks posed by phishing emails and spoofed login pages.

Modern MFA systems often employ adaptive authentication (Venkatasubramanian *et al*., 2024), which evaluates contextual signals such as user behavior, device type, and location. When anomalies are detected, additional verification steps are triggered, further reducing the risk of unauthorized access. To ensure adoption, organizations should prioritize user-friendly MFA options, such as mobile push notifications or biometric verification. Enforcing MFA across all sensitive systems and applications significantly lowers phishing success rates and enhances overall cybersecurity.

### 6.2 User Education and Awareness Programs

User awareness is a critical line of defense in mitigating phishing risks. Education programs (Dodge *et al*., 2007) equip individuals with the skills to recognize and respond to suspicious content, including unsolicited emails, manipulated links, and fraudulent communications. Through regular training, users learn to identify warning signs and understand the implications of phishing threats.

Phishing simulations and interactive workshops increase user engagement and readiness, fostering a proactive security mindset. Ongoing awareness campaigns (Skula *et al*., 2020) ensure that employees stay informed about emerging attack techniques, including spear phishing, business email compromise (BEC), and smishing. Empowering users with clear procedures for reporting suspicious activities enables rapid incident response and containment. A knowledgeable and alert workforce significantly reduces the likelihood of successful phishing attacks.

### 6.3 Organizational Policies and Protocols

Strong organizational policies (Frauenstein and Solms, 2009; Pinto *et al.*, 2022) form the foundation of phishing prevention, offering structured guidance and technical controls to mitigate threats. These include secure email gateways, mandatory software patching, and strict access management to reduce vulnerabilities. Regular audits and compliance checks ensure that security standards are upheld and continuously improved.

A widely recognized framework utilized by organizations to manage cybersecurity risks is the NIST Cybersecurity Framework (NIST CSF) (Barrett, 2018), which offers a structured and comprehensive approach to strengthening security posture. The framework is built around five fundamental functions such as Identify, Protect, Detect, Respond, and Recover, that enable organizations to enhance their resilience and adaptive capacity against evolving cyber threats, including phishing. Within this model, controls such as Access Control (AC), Awareness and Training (AT), and Incident Response (IR) play a pivotal role in phishing defense and recovery planning. NIST further reinforces this framework through key publications, notably NIST SP 800-53 (NIST, 2013), which specifies essential security and privacy safeguards for information systems, and NIST SP 800-61 Rev. 2, which provides detailed incident-handling best practices applicable to phishing detection and mitigation efforts.

Organizations should operationalize these policies by maintaining documented incident response plans, conducting periodic training and phishing simulations (Shahbaznezhad *et al.*, 2021), and integrating real-time monitoring and threat intelligence to enhance their security posture. Embedding NIST-recommended best practices into security policy not only ensures compliance but also strengthens the organization's proactive defense against phishing threats.

## 7. CHALLENGES IN PHISHING MITIGATION

Phishing mitigation continues to face major challenges due to the increasing sophistication of attack techniques and the limitations of current defense mechanisms. Cybercriminals leverage advanced social engineering, rapidly evolving attack vectors, and the exploitation of emerging technologies to bypass traditional detection systems. A comprehensive literature survey reveals that conventional anti-phishing approaches are often reactive and struggle to adapt to dynamic and intelligent phishing campaigns, thus need continuous evolution of mitigation strategies (Kavya and Sumathi, 2024).

Among the most pressing challenges are adversarial machine learning (Shirazi *et al.*, 2019), evasion tactics (Ghafoor *et al.*, 2025), zero-day phishing attacks (Guo, 2023), and limitations in scalability and real-time detection. These evolving threats undermine the reliability of existing cybersecurity solutions, pushing the boundaries of what traditional systems and even ML and DL-based models can achieve. Addressing these challenges requires a multifaceted approach that involves resilient algorithm design, adaptive threat detection, and scalable infrastructure capable of handling real-time data streams.

### 7.1 Adversarial Machine Learning in Phishing

Adversarial Machine Learning (Goodfellow *et al.*, 2014; Pillai *et al.*, 2023 ) has emerged as a critical threat in phishing mitigation. Attackers exploit weaknesses in machine learning (ML) models by crafting adversarial examples, inputs that are subtly manipulated to deceive detection systems. For example, slight modifications in phishing email content, URL structures, or metadata can bypass ML-based filters without triggering user alerts.

Furthermore, attackers may poison training datasets (Shirazi *et al.*, 2019; Wang *et al.*, 2022), injecting misleading data (Unsal *et al.*, 2021) to degrade model performance over time. This is particularly dangerous in phishing detection, where learning from clean, representative data is essential for accurate classification. In response, researchers advocate for techniques such as adversarial training, robust model architectures, and data sanitization to enhance defenses. Nonetheless, adversarial tactics evolve rapidly, requiring continuous adaptation of ML and DL based detection frameworks to maintain effectiveness.

### 7.2 Evasion Tactics and Zero-Day Phishing Attacks

Evasion tactics (Ahmed *et al.*, 2022) represent a growing threat as attackers develop methods to systematically bypass detection tools. Techniques such as URL obfuscation (Skula and Kvet, 2024), which involves using misspelled or shortened URLs, and polymorphic phishing pages (Fatt, 2014), which frequently change their appearance or structure, are commonly employed. Encrypted phishing sites, which utilize SSL/TLS certificates (Oppliger *et al.*, 2006; Abate *et al.*, 2023), add another layer of complexity, allowing malicious activity to occur within seemingly secure channels.

In contrast, zero-day phishing attacks (Bilge and Dumitraş, 2012; Akshaya, 2019) exploit unknown or unpatched vulnerabilities, making them nearly impossible to detect using signature-based or heuristic systems. These attacks are often tailored, leveraging spear phishing and social engineering to target specific individuals or organizations, increasing their success rates.

Countermeasures include real-time behavior analysis, anomaly detection, and integration with threat intelligence platforms that can identify indicators of compromise across different sources. Despite progress, the unpredictable and stealthy nature of these attacks continues to pose significant challenges to defenders.

### 7.3 Scalability and Real-Time Detection Issues

Scalability (Kavya and Sumathi, 2024) and real-time detection (Sameen *et al.*, 2020) are major operational aspects in modern phishing defense. The increasing volume of emails, URLs, and web traffic necessitates fast, accurate analysis across massive data streams. Traditional detection systems often fail to scale efficiently without compromising detection accuracy or incurring high latency, especially when dealing with large enterprise networks or global infrastructures.

Real-time detection is further hindered by the dynamics of phishing attacks (Tseng *et al.*, 2013), which often involve the rapid deployment and takedown of malicious domains, rendering static blacklists and rule-based systems ineffective. Additionally, phishing techniques that incorporate encryption, obfuscation, or fast-flux DNS behavior add significant processing overhead (Patsakis *et al.*, 2020).

Solutions include distributed computing frameworks, streamlined ML models, and edge-based detection that shift processing closer to the data source to reduce latency. The trade-off between scalability, speed, and accuracy remains a fundamental challenge in deploying robust phishing mitigation solutions.

## 8. FUTURE DIRECTIONS IN PHISHING RESEARCH

As phishing techniques continue to evolve in complexity and adaptability, future research must also evolve to address the limitations of current defenses and anticipate emerging threats. The dynamic nature of phishing campaigns driven by advancements in artificial intelligence, deepfake technologies, and attacker automation necessitates innovative approaches that go beyond traditional detection models.

One critical direction is the integration of Explainable Artificial Intelligence (XAI) (Zhang *et al.*, 2022) into phishing detection systems. Current machine learning models, while effective, often act as "black boxes." Incorporating XAI enables analysts and end-users to understand the reasoning behind model decisions, thereby increasing trust, transparency, and the ability to identify false positives or vulnerabilities within the system.

Federated learning (Li *et al*., 2023) is another promising approach, enabling the collaborative training of phishing detection models across organizations without sharing sensitive data. This decentralization enhances privacy while enabling richer, more diverse learning experiences for detection algorithms.

With phishing attacks increasingly exploiting mobile and voice platforms, future research must extend into multi-modal detection systems (Phang *et al*., 2024) capable of analyzing SMS (smishing), voice-based phishing (vishing), and phishing embedded in mobile applications and social media platforms. These require novel data collection methods and models that can handle unstructured, cross-platform data in real-time.

Another emerging focus is the development of proactive defense mechanisms (Colbaugh and Glass, 2011) that not only detect but also anticipate phishing attacks. Leveraging threat intelligence, behavioral prediction, and real-time user interaction analytics can help identify potential attack vectors before they are exploited.

Additionally, future research should address the human factors in phishing (Gallo *et al*., 2024), exploring how psychological and behavioral cues influence susceptibility among victims. Designing user-centric defenses, personalized training programs, and intelligent warning systems can complement technical measures to build holistic resilience.

As cybercriminals continue to adapt, interdisciplinary collaboration combining cybersecurity, machine learning, psychology, and policy-making will be essential for developing robust, adaptive, and ethically grounded phishing defense strategies.

## 9. CONCLUSION

Phishing persists as one of the most pervasive and technically adaptable threats in the cybersecurity domain, leveraging both social engineering and technological subterfuge to compromise sensitive data and disrupt digital infrastructure. This survey synthesizes a wide array of phishing methodologies, spanning traditional email-based schemes to advanced vectors that exploit IoT, cloud ecosystems, and artificial intelligence, demonstrating the increasing complexity and scope of contemporary phishing campaigns.

Recent advancements in machine learning, deep learning, and natural language processing have substantially improved phishing detection capabilities. However, these approaches are not immune to emerging challenges, particularly adversarial machine learning, zero-day threats, and evasion techniques that undermine detection efficacy and system robustness. Moreover, real-time scalability remains a critical hurdle in operational environments.

From a scholarly perspective, this landscape necessitates continuous innovation in both technical and human-centric defenses. Future research should prioritize the development of explainable AI models, federated learning for privacy-preserving collaboration, and multi-modal detection frameworks capable of processing heterogeneous data across communication platforms. Furthermore, addressing phishing from a behavioral and policy standpoint through user education, security governance, and compliance with frameworks such as NIST will be essential in fostering a resilient digital ecosystem.

The multifaceted nature of phishing demands interdisciplinary collaboration that spans computer science, behavioral psychology, and organizational policy. Only through such integrative efforts can academia make meaningful contributions to the development of proactive, scalable, and ethically sound countermeasures against this enduring cybersecurity challenge.

This survey systematically consolidates phishing attack techniques, detection methodologies, and countermeasures within a unified multidimensional framework. By explicitly addressing AI-driven and LLM-enabled phishing, the study advances understanding of emerging threats and defense gaps.

## 10. AUTHOR DECLARATION ON AI USE:

The authors affirm that AI tools were employed solely for linguistic enhancement and paraphrasing to improve readability and grammatical precision. No AI system contributed to the conceptualization, analysis, interpretation, or development of the research framework. All intellectual content and critical insights reflect the authors' independent work.

## 11. REFERENCES

AAG IT. (2025). *The latest phishing statistics*. Retrieved September 25, 2025, from https://aag-it.com/the-latest-phishing-statistics/

Abate, A. F., Castiglione, A., Cimmino, L., De Angelis, D., Flauto, S., & Volpe, A. (2023, May). On the (in)security and weaknesses of commonly used applications on large-scale distributed systems. In *2023 24th International Conference on Control Systems and Computer Science (CSCS)* (pp. 572–579). IEEE.

Ahmed, U., Lin, J. C. W., & Srivastava, G. (2022). Mitigating adversarial evasion attacks of ransomware using ensemble learning. *Computers and Electrical Engineering, 100*, 107903.

Akinyelu, A. A., & Adewumi, A. O. (2014). Classification of phishing email using random forest machine learning technique. *Journal of Applied Mathematics, 2014*(1), 425731.

Akiyama, T., Haruta, A., Yamaoka, H., Masuda, H., & Yamamoto, K. (2024, March). An attempts to improve security on campus. In *Proceedings of the 2024 ACM SIGUCCS Annual Conference* (pp. 52–56). ACM.

Akshaya, S. (2019). *A study on zero-day attacks*.

Alam, M. N., Sarma, D., Lima, F. F., Saha, I., & Hossain, S. (2020, August). Phishing attacks detection using machine learning approach. In *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)* (pp. 1173–1179). IEEE.

Algarni, A., Xu, Y., & Chan, T. (2014, June). Social engineering in social networking sites: The art of impersonation. In *2014 IEEE International Conference on Services Computing* (pp. 797–804). IEEE.

Ali, A. A., & Abdullah, A. A. (2025). Text email spam adversarial attack detection and prevention based on deep learning. *International Journal of Intelligent Engineering & Systems, 18*(2).

Alharbi, A., Alotaibi, A., Alghofaili, L., Alsalamah, M., Alwasil, N., & Elkhediri, S. (2022, January). Security in social media: Awareness of phishing attacks techniques and countermeasures. In *2022 2nd International Conference on Computing and Information Technology (ICCIT)* (pp. 10–16). IEEE.

Alhogail, A., & Alsabih, A. (2021). Applying machine learning and natural language processing to detect phishing email. *Computers & Security, 110*, 102414.

Alkhalil, Z., Hewage, C., Nawaf, L., & Khan, I. (2021). Phishing attacks: A recent comprehensive study and a new anatomy. *Frontiers in Computer Science, 3*, 563060.

Almomani, A., Gupta, B. B., Atawneh, S., Meulenberg, A., & Almomani, E. (2013). A survey of phishing email filtering techniques. *IEEE Communications Surveys & Tutorials, 15*(4), 2070–2090.

Alotaibi, S. R., Alkahtani, H. K., Aljebreen, M., Alshuhail, A., Saeed, M. K., Ebad, S. A., Almukadi, W. S., & Alotaibi, M. (2025). Explainable artificial intelligence in web phishing classification on secure IoT with cloud-based cyber-physical systems. *Alexandria Engineering Journal, 110*, 490–505.

Aleroud, A., & Zhou, L. (2017). Phishing environments, techniques, and countermeasures: A survey. *Computers & Security, 68*, 160–196.

Anti-Phishing Working Group (APWG). (2024). *eCrime Exchange (ECX)*. Retrieved September 25, 2025, from https://apwg.org/ecx/

Antonakakis, M., April, T., Bailey, M., Bernhard, M., Bursztein, E., Cochran, J., Durumeric, Z., Halderman, J. A., Invernizzi, L., Kallitsis, M., & Kumar, D. (2017). Understanding the Mirai botnet. In *26th USENIX Security Symposium (USENIX Security 17)* (pp. 1093–1110).

Anitha, P. U., Rao, C. G., & Babu, D. S. (2021). Email spam filtering using machine learning based XGBoost classifier method. *Turkish Journal of Computer and Mathematics Education, 12*(11), 2182–2190.

Apache Software Foundation. (2024). *Apache SpamAssassin*. Retrieved September 25, 2025, from https://spamassassin.apache.org/

Arya, S., & Chamotra, S. (2021). Multi-layer detection framework for spear-phishing attacks. In *International Conference on Information Systems Security* (pp. 38–56). Springer.

Ariyadasa, S., Fernando, S., & Fernando, S. (2022). Combining long-term recurrent convolutional and graph convolutional networks to detect phishing sites using URL and HTML. *IEEE Access, 10*, 82355–82375.

Arroyabe, M. F., Arranz, C. F., de Arroyabe, I. F., & de Arroyabe, J. C. F. (2024). Revealing the realities of cybercrime in small and medium enterprises: Understanding fear and taxonomic perspectives. *Computers & Security, 141*, 103826.

Basnet, R. B., Sung, A. H., & Liu, Q. (2011, July). Rule-based phishing attack detection. In *International Conference on Security and Management (SAM 2011)*.

Barrett, M. P. (2018). *Framework for improving critical infrastructure cybersecurity* (Version 1.1). National Institute of Standards and Technology.

Bauskar, S. R., Madhavaram, C. R., Galla, E. P., Sunkara, J. R., & Gollangi, H. K. (2024). AI-driven phishing email detection: Leveraging big data analytics for enhanced cybersecurity. *Library Progress International, 44*(3), 7211–7224.

Bell, S., & Komisarczuk, P. (2020, February). An analysis of phishing blacklists: Google Safe Browsing, OpenPhish, and PhishTank. In *Proceedings of the Australasian Computer Science Week Multiconference* (pp. 1–11). ACM.

Bergholz, A., Chang, J. H., Paass, G., Reichartz, F., & Strobel, S. (2008). Improved phishing detection using model-based features. In *Conference on Email and Anti-Spam (CEAS)*.

Bergholz, A., de Beer, J., Glahn, S., Moens, M. F., Paaß, G., & Strobel, S. (2010). New filtering approaches for phishing email. *Journal of Computer Security, 18*(1), 7–35.

Bilge, L., & Dumitraș, T. (2012, October). Before we knew it: An empirical study of zero-day attacks in the real world. In *Proceedings of the 2012 ACM Conference on Computer and Communications Security* (pp. 833–844). ACM.

Bojjagani, S., Brabin, D. D., & Rao, P. V. (2020). PhishPreventer: A secure authentication protocol for prevention of phishing attacks in mobile environment with formal verification. *Procedia Computer Science, 171*, 1110–1119.

Bose, I., & Leung, A. C. M. (2007). Unveiling the mask of phishing: Threats, preventive measures, and responsibilities. *Communications of the Association for Information Systems, 19*(1), 24.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems, 30*(1–7), 107–117.

Burns, A. J., Johnson, M. E., & Caputo, D. D. (2019). Spear phishing in a barrel: Insights from a targeted phishing campaign. *Journal of Organizational Computing and Electronic Commerce, 29*(1), 24–39.

Cardenas, D. J. S., Hahn, A., & Liu, C. C. (2020). Assessing cyber-physical risks of IoT-based energy devices in grid operations. *IEEE Access, 8*, 61161–61173.

Castillo, E., Dhaduvai, S., Liu, P., Thakur, K. S., Dalton, A., & Strzalkowski, T. (2020, May). Email threat detection using distinct neural network approaches. In *Proceedings of the First International Workshop on Social Threats in Online Conversations: Understanding and Management* (pp. 48–55). ACM.

Chaudhuri, A. (2023). Clone phishing: Attacks and defenses. *International Journal of Scientific and Research Publications, 13*(4).

Cisco Systems, Inc. (2024). *Cisco Umbrella: Leader in cloud cybersecurity and SASE solutions*. Retrieved September 25, 2025, from https://umbrella.cisco.com/

Colbaugh, R., & Glass, K. (2011, July). Proactive defense for evolving cyber threats. In *Proceedings of the 2011 IEEE International Conference on Intelligence and Security Informatics* (pp. 125–130). IEEE.

Crocker, D., Hansen, T., & Kucherawy, M. (2011). *DomainKeys Identified Mail (DKIM) signatures (RFC 6376)*. Internet Engineering Task Force.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 4171–4186). Association for Computational Linguistics.

Ding, W., & Hu, H. (2018, October). On the safety of IoT device physical interaction control. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security* (pp. 832–846). ACM.

Diviya, M., Shri, K. J., & Krishnan, D. G. (2024, August). Enhancing cyber security through Phish-Net: An artificial neural network for phishing detection. In *2024 Second International Conference on Networks, Multimedia and Information Technology (NMITCON)* (pp. 1–5). IEEE.

Do, N. Q., Selamat, A., Krejcar, O., Herrera-Viedma, E., & Fujita, H. (2022). Deep learning for phishing detection: Taxonomy, current challenges and future directions. *IEEE Access, 10*, 36429–36463.

Dodge, R. C., Jr., Carver, C., & Ferguson, A. J. (2007). Phishing for user security awareness. *Computers & Security, 26*(1), 73–80.

Doshi, J., Parmar, K., Sanghavi, R., & Shekokar, N. (2023). A comprehensive dual-layer architecture for phishing and spam email detection. *Computers & Security, 133*, 103378.

Eze, C. S., & Shamir, L. (2024). Analysis and prevention of AI-based phishing email attacks. *Electronics, 13*(10), 1839.

Falade, P. V. (2023). Decoding the threat landscape: ChatGPT, FraudGPT, and WormGPT in social engineering attacks. *arXiv*. https://arxiv.org/abs/2310.05595

Fang, Y., Zhang, C., Huang, C., Liu, L., & Yang, Y. (2019). Phishing email detection using improved RCNN model with multilevel vectors and attention mechanism. *IEEE Access, 7*, 56329–56340.

Fatt, J. C. S., Leng, C. K., & San Nah, S. (2014, September). Phishdentity: Leverage website favicon to offset polymorphic phishing website. In *2014 Ninth International Conference on Availability, Reliability and Security* (pp. 114–119). IEEE.

Feng, T., & Yue, C. (2020, June). Visualizing and interpreting RNN models in URL-based phishing detection. In *Proceedings of the 25th ACM Symposium on Access Control Models and Technologies* (pp. 13–24). ACM.

Fette, I., Sadeh, N., & Tomasic, A. (2007, May). Learning to detect phishing emails. In *Proceedings of the 16th International Conference on World Wide Web* (pp. 649–656). ACM.

Figueiredo, J., Carvalho, A., Castro, D., Gonçalves, D., & Santos, N. (2024). On the feasibility of fully AI-automated vishing attacks. *arXiv*. https://arxiv.org/abs/2409.13793

Frauenstein, E. D., & von Solms, R. (2009). Phishing: How an organization can protect itself. In *ISSA Conference Proceedings* (pp. 253–268).

Frauenstein, E. D., & von Solms, R. (2014, August). Combatting phishing: A holistic human approach. In *2014 Information Security for South Africa* (pp. 1–10). IEEE.

Gallo, L., Gentile, D., Ruggiero, S., Botta, A., & Ventre, G. (2024). The human factor in phishing: Collecting and analyzing user behavior when reading emails. *Computers & Security, 139*, 103671.

Gelernter, N., & Herzberg, A. (2016, April). Tell me about yourself: The malicious CAPTCHA attack. In *Proceedings of the 25th International Conference on World Wide Web* (pp. 999–1008). ACM.

Ghafoor, A., Shah, M. A., Al-Naeem, M., & Maple, C. (2025). Decoding phishing evasion: Analyzing attacker strategies to circumvent detection systems. *IEEE Access*.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv*. https://arxiv.org/abs/1412.6572

Guo, Y. (2023). A review of machine learning-based zero-day attack detection: Challenges and future directions. *Computer Communications, 198*, 175–185.

Gualberto, E. S., de Sousa, R. T., Thiago, P. D. B., da Costa, J. P. C., & Duque, C. G. (2020). From feature engineering and topic models to enhanced prediction rates in phishing detection. *IEEE Access, 8*, 76368–76385.

Gupta, S., & Singhal, A. (2017, August). Phishing URL detection by using artificial neural network with PSO. In *2017 2nd International Conference on Telecommunication and Networks (TEL-NET)* (pp. 1–6). IEEE.

Harikrishnan, N. B., Vinayakumar, R., & Soman, K. P. (2018, March). A machine learning approach towards phishing email detection. In *Proceedings of the Anti-Phishing Pilot at the ACM International Workshop on Security and Privacy Analytics (IWSPA AP)* (pp. 455–468). ACM.

Haynes, K., Shirazi, H., & Ray, I. (2021). Lightweight URL-based phishing detection using natural language processing transformers for mobile devices. *Procedia Computer Science, 191*, 127–134.

Hazell, J. (2023). Spear phishing with large language models. *arXiv*. https://arxiv.org/abs/2305.06972

Hiransha, M., Unnithan, N. A., Vinayakumar, R., Soman, K. P., & Verma, A. D. R. (2018, March). Deep learning based phishing e-mail detection. In *Proceedings of the 1st AntiPhishing Shared Pilot at the 4th ACM International Workshop on Security and Privacy Analytics (IWSPA)* (pp. 1–5). ACM.

Ikram, M., Masood, R., Tyson, G., Kaafar, M. A., Loizon, N., & Ensafi, R. (2020). Measuring and analysing the chain of implicit trust: A study of third-party resource loading. *ACM Transactions on Privacy and Security, 23*(2), 1–27.

Ilca, F., & Balan, T. (2021, September). Phishing as a service campaign using IDN homograph attack. In *2021 International Aegean Conference on Electrical Machines and Power Electronics (ACEMP) & 2021 International Conference on Optimization of Electrical and Electronic Equipment (OPTIM)* (pp. 338–344). IEEE.

Ismail, S. S., Mansour, R. F., Abd El-Aziz, R. M., & Taloba, A. I. (2022). Efficient email spam detection strategy using genetic decision tree processing with NLP features. *Computational Intelligence and Neuroscience, 2022*, 7710005.

Jain, A. K., Parashar, S., Katare, P., & Sharma, I. (2020). Phishskape: A content based approach to escape phishing attacks. *Procedia Computer Science, 171*, 1102–1109.

Jayatilaka, A., Arachchilage, N. A. G., & Babar, M. A. (2024). Why people still fall for phishing emails: An empirical investigation into how users make email response decisions. *arXiv*. https://arxiv.org/abs/2401.13199

Jha, B., Atre, M., & Rao, A. (2022, December). Detecting cloud-based phishing attacks by combining deep learning models. In *2022 IEEE 4th International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)* (pp. 130–139). IEEE.

John-Africa, E., & Emmah, V. T. (2022). Performance evaluation of LSTM and RNN models in the detection of email spam messages. *European Journal of Information Technologies and Computer Science, 2*(6), 24–30.

Kale, N., Kochrekar, S., Mote, R., & Dholay, S. (2021, July). Classification of fraud calls by intent analysis of call transcripts. In *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (pp. 1–6). IEEE.

Karim, A., Azam, S., Shanmugam, B., & Kannoorpatti, K. (2020). Efficient clustering of emails into spam and ham: The foundational study of a comprehensive unsupervised framework. *IEEE Access, 8*, 154759–154788.

Kaur, A., & Mian, S. M. (2023, May). A review on phishing technique: Classification, lifecycle and detection approaches. In *2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)* (pp. 336–339). IEEE.

Kavya, S., & Sumathi, D. (2024). Staying ahead of phishers: A review of recent advances and emerging methodologies in phishing detection. *Artificial Intelligence Review, 58*(2), 50.

Kelacyber. (2025). *How phishing-as-a-service fuels cybercrime at scale*. Retrieved September 25, 2025, from https://www.kelacyber.com/blog/phishing-as-a-service/

Keskisaari, A. (2022). *Cybercrime in cloud environments: Tactics, techniques and procedures*.

KnowBe4. (2024). *Integrated security awareness training and simulated phishing platform (KnowBe4)*. Retrieved September 25, 2025, from https://www.knowbe4.com/

Krishnapriya, S., & Singh, S. (2024). A comprehensive survey on advanced persistent threat (APT) detection techniques. *Computers, Materials & Continua, 80*(2).

Kucherawy, M., & Zwicky, E. (2015). *Domain-based message authentication, reporting, and conformance (DMARC) (RFC 7489)*. Internet Engineering Task Force.

Kumar, A., Chatterjee, J. M., & Díaz, V. G. (2020). A novel hybrid approach of SVM combined with NLP and probabilistic neural network for email phishing. *International Journal of Electrical and Computer Engineering, 10*(1), 486.

Kumarasinghe, P., Dissanayake, D., Gamage, P., & Ganegoda, G. U. (2023, December). User behavior analysis in determining the vulnerable category of vishing and smishing. In *2023 5th International Conference on Advancements in Computing (ICAC)* (pp. 35–40). IEEE.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). ALBERT: A lite BERT for self-supervised learning of language representations. *arXiv*. https://arxiv.org/abs/1909.11942

Lee, J., Tang, F., Ye, P., Abbasi, F., Hay, P., & Divakaran, D. M. (2021, September). D-fence: A flexible, efficient, and comprehensive phishing email detection system. In *2021 IEEE European Symposium on Security and Privacy (EuroS&P)* (pp. 578–597). IEEE.

Lee, Y., Saxe, J., & Harang, R. (2020). CATBERT: Context-aware tiny BERT for detecting social engineering emails. *arXiv*. https://arxiv.org/abs/2010.03484

Li, B., Wang, P., Shao, Z., Liu, A., Jiang, Y., & Li, Y. (2023). Defending Byzantine attacks in ensemble federated learning: A reputation-based phishing approach. *Future Generation Computer Systems, 147*, 136–148.

Li, Q., Cheng, M., Wang, J., & Sun, B. (2020). LSTM-based phishing detection for big email data. *IEEE Transactions on Big Data, 8*(1), 278–288.

Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008, December). Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining* (pp. 413–422). IEEE.

Liu, J., Pun, P., Vadrevu, P., & Perdisci, R. (2023, July). Understanding, measuring, and detecting modern technical support scams. In *2023 IEEE European Symposium on Security and Privacy (EuroS&P)* (pp. 18–38). IEEE.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv*. https://arxiv.org/abs/1907.11692

Ma, J., Saul, L. K., Savage, S., & Voelker, G. M. (2009). Identifying suspicious URLs: An application of large-scale online learning. In *Proceedings of the 26th Annual International Conference on Machine Learning* (pp. 681–688). ACM.

Ma, Y., Gao, S., & Li, Y. (2023). A survey of phishing detection methods: Techniques, datasets, and trends. *Journal of Network and Computer Applications, 209*, 103488.

Mishra, A., & Rautaray, S. S. (2022, September). A hybrid deep learning approach for phishing website detection. In *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 1892–1898). IEEE.

Mohammad, R., Thabtah, F., & McCluskey, L. (2014). Predicting phishing websites based on self-structuring neural network. *Neural Computing and Applications, 25*(2), 443–458.

Mohammad, R., Thabtah, F., & McCluskey, L. (2016). Intelligent rule-based phishing websites classification. *Information Sciences, 326*, 186–197.

Montoya, L., & Gómez, J. (2022). A review of anti-phishing training programs: Effectiveness, challenges, and emerging approaches. *Computers & Security, 113*, 102562.

Nash, R., & West, D. (2020). Social engineering attacks: Phishing, vishing, and smishing. *Journal of Cybersecurity Research, 4*(1), 12–29.

Nguyen, T., Ngo, Q., & Pham, V. (2023, October). Machine learning-based phishing detection in emails using contextual embeddings. In *2023 IEEE International Conference on Big Data and Smart Computing (BigComp)* (pp. 15–22). IEEE.

Oshikawa, R., Takaragi, K., & Nakagawa, H. (2019). Detecting phishing websites using visual similarity analysis. *Expert Systems with Applications, 118*, 307–319.

Patel, R., & Jain, S. (2021). Phishing detection using ensemble machine learning techniques. *Journal of Information Security and Applications, 61*, 102899.

Parameswaran, S., & Srinivasan, S. (2020). A comprehensive survey on phishing attacks: Trends, challenges, and future directions. *Computers & Security, 92*, 101754.

Prakash, A., Kumaraguru, P., & Reddy, A. (2010, July). Phishnet: Predictive blacklisting to detect phishing attacks. In *Proceedings of the 7th Annual ACM Workshop on Privacy in the Electronic Society* (pp. 1–8). ACM.

Rao, R. S., & Pais, A. R. (2017). Detection of phishing websites using machine learning techniques. *Procedia Computer Science, 115*, 504–511.

Ravichandran, V., & Alazab, M. (2021). Phishing detection using deep learning techniques: A comprehensive review. *Journal of Network and Computer Applications, 193*, 103183.

Shu, K., Wang, S., Lee, D., & Liu, H. (2019). Detecting fake news on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter, 21*(2), 22–36.

Singh, D., & Sharma, M. (2020). Machine learning approaches for phishing detection: A review. *International Journal of Information Management, 54*, 102177.

Sahoo, D., & Sahoo, S. (2021). A deep learning approach for phishing website detection using URL and HTML features. *Journal of Information Security and Applications, 61*, 102945.

Sun, L., He, W., & Gao, Y. (2022). Detecting phishing websites using a hybrid deep learning model. *Computers & Security, 113*, 102571.

Thakur, R., & Kumar, R. (2019). A comparative study of phishing detection techniques using machine learning. *Procedia Computer Science, 152*, 221–228.

Tran, N., & Dang, J. (2020). Phishing detection using ensemble learning and feature selection. *Expert Systems with Applications, 140*, 112878.

Verma, S., & Yadav, S. (2021). Intelligent phishing detection using hybrid machine learning models. *Journal of Ambient Intelligence and Humanized Computing, 12*(4), 4245–4258.

Wang, Y., & Chen, X. (2019). Phishing website detection with deep neural networks. *Neurocomputing, 365*, 192–203.

Xie, P., & Zhang, L. (2021). A novel phishing detection model based on URL analysis and deep learning. *Information Sciences, 546*, 1006–1020.

Yang, C., & Li, J. (2020). Detecting phishing websites using attention-based deep learning. *IEEE Access, 8*, 123456–123467.

Zhang, H., & Zhao, Y. (2019). Phishing website detection using URL features and machine learning algorithms. *Computers & Security, 87*, 101568.

Zhou, X., & Leung, H. (2021). A survey on phishing detection and prevention techniques. *Journal of Network and Computer Applications, 179*, 102986.

Zhou, Y., & Li, W. (2022). Phishing detection using graph neural networks. *Knowledge-Based Systems, 248*, 108812.