



Phishing in the age of distributed intelligence: taxonomies, detection strategies, and the emerging role of federated learning

Esraa Daoud¹ · Javier Garcia-Blas² · Sadi Alawadi³ · Jesus Carretero²

Received: 17 December 2025 / Accepted: 19 March 2026
© The Author(s) 2026

Abstract

Phishing has evolved into one of the most adaptive and damaging cybersecurity threats, continually reshaping itself to exploit human behaviour, technical vulnerabilities, and, more recently, advances in artificial intelligence. As attack vectors diversify from traditional email scams to sophisticated, multi-stage, AI-generated, and hybrid phishing campaigns, defending against them has become significantly challenging. This survey provides a comprehensive and contemporary examination of the phishing landscape, tracing its evolution, analysing real-world incidents, and contextualising its growing impact through global statistics. We introduce a unified, multidimensional taxonomy that categorizes phishing attacks into distinct categories, providing a clearer understanding of how new attack techniques operate and escalate. In parallel, we review a broader range of phishing detection strategies, from list-based, heuristic, and similarity-driven techniques to modern machine learning and deep learning approaches. While these methods have advanced detection capabilities, they continue to face significant constraints related to data privacy, scalability, and the rapid emergence of novel attack patterns. Motivated by these limitations, the survey highlights the growing relevance of Federated Learning (FL) as a privacy-preserving and collaborative paradigm for phishing detection. To the best of our knowledge, this is the first comprehensive survey to examine phishing defence through the lens of FL. In particular, we examine the role of FL in enabling decentralized, privacy-aware detection without exchanging raw data, compared to centralized training in terms of performance, privacy guarantees, resilience, and scalability. Drawing from this analysis, we offer valuable insights into critical research gaps and future directions for developing robust, scalable, and privacy-aware phishing detection solutions.

Keywords Phishing · Cybersecurity · Privacy-Preserving · Machine Learning · Federated Learning · Anti-Phishing Strategies · Phishing Targets · Phishing Media · Phishing Phases · Phishing Techniques · Malware · Social engineering

1 Introduction

The rapid advancement of technology and internet usage over the last decade has significantly expanded network scale and associated applications, with the number of users exceeding two billion [1, 2]. According to [3], the number of internet users reached approximately 4.9 billion in 2020, highlighting the significant growth of digital connectivity in modern life. Notably, the COVID-19 pandemic has significantly increased individuals' reliance on online services, resulting in a 20% increase in internet usage, from 50% to 70%, during this period [4]. This wave of digitalization not only transformed societal behaviour but also led to the exponential growth of data generation. Based on the Dataprot study, the global deployment of Internet of Things (IoT) devices is expected to exceed 29 billion by 2030, further contributing to this data explosion [5, 6].

✉ Sadi Alawadi
sadi.alawadi@bth.se

Esraa Daoud
esraa_jamil@aabu.edu.jo

Javier Garcia-Blas
fjblas@inf.uc3m.es

Jesus Carretero
jcarrete@inf.uc3m.es

¹ Department of Programming, Al al-Bayt University, Mafraq, Jordan

² Computer Science and Engineering Department, Universidad Carlos III de Madrid, Madrid, Spain

³ Computer Science Department, Blekinge Tekniska Högskola (BTH), Karlskrona, Sweden

However, the lack of a regulatory framework governing online behavior, coupled with the absence of centralized internet surveillance or authority, exposes users' information and valuable assets to security threats [2, 3]. This has led to a surge in sophisticated cyberattacks, resulting in a substantial increase in data breaches and posing significant challenges for network security in accurately detecting intrusions [1, 2, 7, 8]. For example, a study by the International Association of IT Asset Managers (IAITAM), which involved over 3000 employees across 12 countries, revealed that 94% of them experienced data breaches due to cyber-attacks during the pandemic, with an average of over 2 breaches per employee [9]. Studies for 2025 show a dramatic increase in cyberattacks, with Check Point Research reporting a 21% rise in global attacks in Q2 2025 and a nearly 50% surge in Q1 2025 [10], most of them attempts of intrusions, being an intrusion something defined as any attempt to compromise the confidentiality, integrity, or availability (CIA) of data, computer, network, or more to bypass the implemented security mechanisms [1, 7]. These threats are particularly concerning in contexts like remote healthcare, exemplified by the Internet of Medical Things (IoMT). The IoMT generates and analyzes vast amounts of big data in a distributed manner, and compromising security in these scenarios could result in severe harm to patients' health [11, 12].

The growing threats highlight the urgent need for advanced, adaptive intrusion detection systems capable of promptly identifying and mitigating such attacks. Intrusion Detection Systems (IDSs) have been developed as security mechanisms specifically designed to detect abnormal activities in cyberspace at the earliest possible stage. By constantly monitoring network traffic, IDSs can identify suspicious behaviour such as phishing, Denial of Service (DoS), and malware attacks, thereby safeguarding the CIA of data [1, 5, 11, 13]. At the same time, in the realm of privacy, the United States and the European Union have introduced regulations governing the collection, storage, and use of personal data. The U.S. has enacted the Health Insurance Portability and Accountability Act (HIPAA), while the EU has implemented the General Data Protection Regulation (GDPR) [14].

Cyberattacks may target vulnerable computer networks or exploit human vulnerabilities, as seen in social engineering (SE) attacks. SE attacks represent a common type of cyberattack that involves manipulating victims into compromising security measures. This manipulation can result in gaining unauthorized access to the target's information or injecting malware into their system, rather than directly attacking the system. Therefore, SE attacks exploit human emotions such as fear, trust, or curiosity to compromise the CIA of data. For instance, a cyber attacker might employ tactics such as sending a spoofed email containing a malicious link to an organization or individual. Once the recipient clicks the link, the attacker begins collecting sensitive data or deploying

malware on the system [15–19]. These increasingly common techniques highlight the vulnerabilities associated with phishing and SE attacks, posing significant security risks to both individuals and organizations.

In recent years, SE attacks have emerged as one of the most prevalent attack methods used by adversaries [20], accounting for approximately 84% of cyberattacks and achieving a notably high success rate. According to the U.S. Justice Department, SE attacks represent a significant global threat in the United States, experiencing the highest number of attacks in 2016, resulting in a staggering cost of \$121.22 billion, followed by Germany and Japan. The U.S. Federal Bureau of Investigation (FBI) has reported instances where attackers impersonate executives and request fund transfers via email, resulting in companies losing over \$2.3 billion. These statistics emphasize the severity of SE attacks, which can exceed even the financial damages caused by natural disasters. Consequently, detecting and mitigating such cyber threats are essential [21].

SE attacks have various forms, including phishing, pretexting, baiting, vishing, piggybacking, tailgating, smishing, or spear-phishing [15, 18–20]. Among them, *phishing attacks* have consistently ranked among the most dominant and dangerous cyber threats over the past decade, prompting more studies on this topic over the last several years [22, 23]. These attacks often exploit human inattention or lack of awareness to steal sensitive information or gain unauthorized access to computerized systems. Notably, 95% of security breaches are attributed to human error, as emphasised by [24]. Phishing tactics range from fraudulent emails to sophisticated schemes on social media platforms like Twitter. The consequences extend beyond financial losses, posing serious risks to individuals' safety and lives [14, 25, 26].

Traditionally, phishing detection was, and still is in many cases, based on powerful centralized tools. However, the dynamic nature of these attacks, coupled with the rapid advancements in the digital landscape and the privacy regulations, has created a constant need for innovative and robust defence mechanisms to counter phishing threats effectively [27]. Thus, while centralized approaches remain the gold standard for raw detection accuracy, federated learning is rapidly becoming the preferred framework for privacy-sensitive environments, like banking and healthcare, for phishing detection and defence techniques [28].

FL was introduced by Google researchers in 2016 to collaboratively train a machine learning model directly on distributed edge devices without sharing raw data [29]. In the fight against phishing, the shift from Centralized Learning (CL) to Federated Learning (FL) represents a move from "bringing data to the model" to "bringing the model to the data". Instead of transferring data to a central server, FL follows a decentralized paradigm in which the model is sent to the data sources, trained locally, and then aggregated, thereby

preserving data privacy and reducing communication of sensitive information.

However, the application of FL to phishing detection remains in its early stages compared to centralized approaches. This is mainly due to challenges such as heterogeneous and non-IID data distributions across clients, coordination and deployment complexity, communication overhead, and security risks inherent in collaborative training environments. As a result, centralized methods continue to dominate the phishing detection literature, while FL-based solutions constitute an emerging research direction with considerable potential.

The main contributions of this paper are summarised as follows:

- We introduce a comprehensive, up-to-date, multidimensional taxonomy of phishing attacks that captures the evolution of modern phishing campaigns, including AI-driven and hybrid techniques, and provides a unified view of attack vectors, delivery channels, and levels of complexity.
- We present a structured, comparative taxonomy of phishing detection and defence approaches, spanning traditional methods and machine learning and deep learning techniques, and analyse them in terms of detection capability, scalability, and practical deployment constraints.
- We provide, to the best of our knowledge, the first comprehensive semi-systematic survey that examines phishing detection through the lens of federated learning, positioning FL as a privacy-preserving and decentralised alternative to conventional centralised ML/DL architectures.
- We conduct a systematic comparison of centralised and federated phishing detection paradigms and identify key open challenges specific to FL-based solutions, including data heterogeneity, communication overhead, and adversarial robustness, thereby outlining promising directions for future research toward scalable, privacy-aware anti-phishing systems.

The remainder of the paper is structured as follows. Section 2 details the methodology and selection criteria. Section 3 backgrounds the concept of phishing, including its definition, historical impact, and relevant statistics. To clarify the focus of the phishing attacks covered in this paper, in Sect. 4 we propose a taxonomy to classify current phishing attack methods. Followed by Sect. 5, which presents a structured classification of phishing detection techniques into two main families, centralised and distributed approaches. The section summarizes their underlying methodologies, explains their operational characteristics, and highlights their respective advantages and limitations. Section 6 presents a landscape of federated learning works for phishing detection and coun-

Table 1 Goal of this Study

<i>Purpose</i>	Identify, classify, and Analyse
<i>Issue</i>	Dimensions of
<i>Object</i>	phishing attacks and defence techniques
<i>Viewpoint</i>	Researchers and practitioners

termeasures, so as comparatives with centralized approaches. Section 7 discusses and presents the current challenges of FL in the field of phishing attacks. Finally, Sect. 8 concludes the paper and enumerates the future work.

2 Methodology

This section outlines the methodology used in this survey, which follows a semi-systematic approach to identify phishing attacks, detection techniques, and the role of FL in phishing detection. In alignment with established procedures for a semi-systematic survey [30], the study was conducted in four structured steps: defining the study goal, formulating research questions, identifying the search strategy, and applying explicit inclusion and exclusion criteria.

2.1 Study goal

Table 1 summarises the goal of this study based on the Goal–Question–Metric (GQM) approach [31]. The target audience for this study comprises researchers and practitioners who are interested in (i) contributing to the research about phishing attacks and detection and defence approaches, (ii) a comprehensive understanding of phishing threats, challenges, and contemporary detection strategies, including FL.

2.2 Research questions

To achieve the stated goal for this survey, we investigated the following Research Questions (RQs):

- RQ1: What are existing taxonomies of phishing attacks and defense techniques?
Objective: We aim to identify existing taxonomies of phishing attacks and defense techniques.
- RQ2: What are the main taxonomic dimensions used for Phishing attacks?
Objective: We aim to synthesize the taxonomies identified in the answer to RQ1 into a comprehensive taxonomy of phishing attacks.
- RQ3: What are the main taxonomic dimensions used for defense techniques against phishing attacks?
Objective: We aim to synthesize the taxonomies iden-

Table 2 Queries performed and number of journal publications obtained (2015–2025).

ID	Query	Number of publications			
		Scopus	IEEE Xplore	ACM DL	ScienceDirect
Q1	(TITLE-ABS-KEY("phishing") OR TITLE-ABS-KEY("phishing detection")) AND PUBYEAR > 2014 AND PUBYEAR < 2026 AND LIMIT-TO(SRCTYPE, "j") AND LIMIT-TO(LANGUAGE, "English")	2,809	283	99	541
Q2	(TITLE-ABS-KEY("phishing") AND TITLE-ABS-KEY("federated learning")) AND PUBYEAR > 2016 AND PUBYEAR < 2026 AND LIMIT-TO(SRCTYPE, "j") AND LIMIT-TO(LANGUAGE, "English")	34	3	1	1

- **I1.** Studies that present a taxonomy, classification, or systematic analysis of phishing attacks or detection techniques.
- **I2.** Studies that propose or evaluate phishing detection techniques using ML, DL, or FL.
- **I3.** Peer-reviewed publication (journals, conferences).
- **I4.** Studies written in English.
- **I5.** Studies providing technical descriptions, evaluations, or empirical results.

2. **Exclusion Criteria (E):** All studies are excluded if any of the following apply:

- **E1.** Studies that mention phishing only as an example without proposing a taxonomy, algorithm, or analytical framework.
- **E2.** Studies focusing solely on unrelated cybersecurity domains (e.g., malware, botnets) without connection to phishing.
- **E3.** Non-peer-reviewed dissertations, tutorials, posters, editorials, and presentations.
- **E4.** Studies lacking sufficient methodological detail or evaluation.
- **E5.** Duplicate publications (in this case, the most complete version was retained).

how search queries translate into the final inclusion set (e.g., screening steps and how different publication venues were handled)

Eligible studies are those that satisfy all inclusion criteria and none of the exclusion criteria. These studies form

the foundation for answering RQ1 and RQ3. Based on those criteria, we selected the papers for this survey from the set extracted using the queries. Table 3 lists the top publication venues represented in this survey, demonstrating that the majority originate from high-impact journals and leading international conferences.

3 Phishing attacks

Phishing attacks have evolved significantly, becoming increasingly sophisticated and widespread threats that affect a wide range of internet users, governmental institutions, and service providers [32–35]. The proliferation of publicly accessible phishing attack toolkits on the internet has enabled malicious actors to deploy advanced online fraud movements with unusual complexity [34]. In the hierarchy of organizational risks, phishing attacks are classified among the top 10 threats and rank as the second most common attack vector in data breaches. In 2020, phishing emerged as the most frequently reported cybercrime to the Internet Crime Complaint Center (IC3), with a rise in the complaints by over 100% compared to the previous year [36]. Despite the existence of automated detection systems, human awareness plays a significant role in contributing to phishing attacks, which are responsible for 82% of data breaches. Even technical experts are vulnerable, as demonstrated by spear phishing attacks targeting security firms such as RSA and HBGary [22, 37, 38]. Table 4 provides real-world examples of phishing incidents along with detailed descriptions.

Phishing attacks are relatively easy to carry out, often exploiting human errors, such as misspellings (e.g., www.twitter.com vs www.twtitter.com) or typosquatting, where attackers replace characters like “w” with “v” to create fake URLs [61, 62]. Attackers also employ a wide range of tac-

Table 3 Top 20 venues used in this survey to extract the papers.

Source title	# of occurrences
IEEE Access	16
ACM Computing Surveys (CSUR)	8
Computers & Security	8
IEEE Communications Surveys & Tutorials	4
Future Internet	3
International Journal of Information Security	3
Expert Systems with Applications	2
Cyber Security and Applications	2
Sensors	2
Informatics	1
Transactions on Emerging Telecommunications Technologies	2
Telecommunication Systems	2
IEEE European Symposium on Security and Privacy (EuroS&P)	2
Decision Support Systems	2
Procedia Computer Science	2
Journal of King Saud University-Computer and Information Sciences	2
IEEE 24th International Symposium on Communications and Information Technologies	1
22nd Annual International Conference on Privacy, Security, and Trust	1
IEEE 50th Conference on Local Computer Networks	1
Proceedings IEEE INFOCOM	1
IEEE International Smart Cities Conference (ISC2)	1
International Conference on Smart Structures and Systems (ICSSS)	1
IEEE Symposium on Computer Applications & Industrial Electronics	1
ACM Transactions on Privacy and Security	1
American Journal of AI Cyber Computing Management	1
Springer journals and proceedings	5
Annual Review of Biomedical Engineering	1
Global Journal of Applied Sciences and Technology	1

tics, including setting up fake websites, sending fraudulent emails, using QR codes, SMS phishing, and performing DNS spoofing [21, 34, 63, 64]. Scam emails remain the most common form of phishing, imitating legitimate sources to prompt victims to click on a link or open an attachment [65–67], as a result, 94% of malware infiltrates computers through email [68]. Given that the COVID-19 pandemic exacerbated this issue, leading to a staggering 600% increase in phishing emails [66]. Phishers frequently impersonate reputable organizations, such as banks or online service providers, to create a sense of urgency and prompt immediate action [38, 69]. A typical phishing message might be, *"Dear valued customer, we have detected unusual activity in your bank account. To ensure uninterrupted service, verify your details by clicking on the link below"* [4], therefore, victims often fall prey to phishing attacks due to insufficient assessment of the sender's identity and a lack of proper education to recognize such threats [20, 61, 62].

3.1 Phishing definition

Due to the dynamic nature of phishing, there is no universally agreed-upon definition; instead, the definitions vary depending on context and application [22, 25, 33, 70, 71]. Generally, phishing can be defined as the process of deceiving targets into performing actions that benefit the attacker [25]. However, some definitions are limited, focusing narrowly on the theft of sensitive information via email. In reality, phishing can occur through various methods, such as malware distribution. For instance, a Man-in-the-Browser (MITB) attack can manipulate banking transactions without stealing the victim's data [22]. Table 5 summarizes various definitions of phishing, categorized by phishing objectives and techniques. Many definitions, such as those by Merwe et al. [72], Kirda and Kruegel [73], Xiang et al. [74], and others, focus primarily on phishing scenarios aimed at stealing victims' information. On the other hand, broader definitions by

Table 4 Summary of significant real-world phishing incidents (2009–2025). The table reports major incidents across diverse sectors for several big companies, including the employed phishing techniques, and the resulting damages and consequences.

Period	Target	Technique	Consequences
2009	Bank of America and Wells Fargo Customers	Spoofed emails	Loss of \$1.5 million from approximately 500 victims [39].
2011	RSA employees	Spoofed emails	Loss of \$66 million [39].
2013-2015	Google and Facebook	BEC	Loss of over \$100 million [4, 39].
2015	Ukrainian Power Grid	Spear Phishing (water-hole), Malware (Black-Energy3)	Shutdown of 30 substations, leaving 230,000 people without power for up to 6 hours; recovery took months [40, 41].
2016	Austrian Aerospace Parts Manufacturer	BEC	Loss of around \$61 million [4].
2017	Bangladesh Bank	Watering Hole, LinkedIn	Loss of \$81 million [41, 42].
	Google and Facebook	BEC	Loss of \$100 million [43].
2018	Middle Eastern Company	Spear Phishing	Targeted a board member, exploiting CVE-2017-0199 vulnerability [44].
	Twitter	Tweets, Malware	Russian operatives targeted U.S. Department of Defense [45].
2018	Equifax Company	Spear Phishing	Exposed sensitive information of 145.5 million Americans over several months [21].
2019	Oregon Department of Human Services	Spear Phishing	Personal health information of 350,000 clients was leaked [39].
2020	Twitter	Phone spear phishing	Loss of over \$100,000 worth of Bitcoins [4, 39].
	COVID Vaccine Supply Chain	Phishing	Accessed sensitive information related to the COVID vaccine cold chain across multiple countries [4, 39].
2021	Oil Pipeline Company, USA	Ransomware, Spear Phishing	Caused a six-day shutdown of Colonial Pipeline; ransom paid in Bitcoin [44].
2022	OpenSea NFT Platform	Phishing	Resulted in a financial loss of \$1.7 million from 17 affected individuals [46, 47].
2023	MOVEit	SQL commands	Breach over 1150 organizations, more than 56 million users compromised [48, 49].
	Enzo Biochem	Ransomware, Phishing	2.5 million patient records exposed, 600,000 with social security numbers (SSNs) [50, 51].
2024	Change Healthcare	Ransomware, Phishing	Loss of leaked data of ~100 million people, \$22 million ransom, major operational losses [52–54].
	British engineering firm Arup (Hong Kong office)	Gen-AI Deepfake Spear Phishing	\$25 million loss from deepfake CFO scam; employee tricked into fraudulent transaction [55, 56].
2025	Pepco Group	BEC AI-crafted email	€15.5 million lost [57].
	npm maintainer Josh Junon's account	Phishing, Malware	Malware was injected into popular npm packages with ~2.6 billion weekly downloads, disrupting about 10% of cloud environments [58].
	Harvard University	Phishing	Harvard's Alumni Affairs and Development systems were compromised, exposing personal data of students, alumni, donors, staff, and faculty [59, 60].

Khonji et al. [22] and Whittaker et al. [75] include phishing tactics that go beyond credential theft. Among these, Alkhalil et al. [25] offers the most comprehensive definition, as it does not limit phishing to specific targets or techniques. Furthermore, [76] offers a detailed overview of the most common phishing attacks.

3.2 Phishing lifecycle

As illustrated in Fig. 3, the life cycle of a typical phishing attack comprises several stages, including the following:

- **Reconnaissance:** Also known as the planning phase, this is when phishers select their attack method, determine objectives, and gather information on potential victims [4, 17, 25, 40, 80].
- **Weaponization:** The process of setting up an attack by identifying vulnerabilities that can be used to deliver phishing materials to victims is referred to as the weaponization or preparation phase [25, 80]. This phase can be conducted manually or with the help of automated tools [23].

Table 5 Comprehensive overview of phishing attack definitions from academic and industry sources. The table summarises how different organizations and researchers conceptualize phishing in terms of its core definition, objectives, and the primary techniques or channels employed to attack.

Source	Definition	Objective	Technique
PhishTank [77]	Phishing is a fraudulent attempt, usually made through email, to steal your personal information.	Personal information	Email
APWG [76]	Phishing is a criminal mechanism employing both social engineering and technical subterfuge to steal consumers' personal identity data financial account credentials.	Personal identity data and financial account credentials.	SE and technical subterfuge
Merwe et al. [72]	A fraudulent activity that involves the creation of a replica of an existing web page to fool a user into submitting personal, financial, or password data	Personal, financial, or password	Web page
Kirda and Kruegel [73]	A form of online identity theft that aims to steal sensitive information such as online banking passwords and credit card information from users.	Sensitive information	Not specified
Xiang et al. [74]	Phishing is a form of identity theft, in which criminals build replicas of target Web sites and lure unsuspecting victims to disclose their sensitive information like passwords, personal identification numbers (PINs), etc.	Sensitive information	Web sites
Chanti and Chithralekha [63]	A fraudulent activity in which the attacker tries to gain illegal financial gain either by: stealing and spoofing user identity/credentials or usurping control of access to user information.	Financial gain	Not specified
Lastdrager et al. [70]	Phishing is a scalable act of deception whereby impersonation is used to obtain information from a target.	Information	Not specified
Mohammad et al. [78]	Phishing website is the practice of creating a copy of a legitimate website and use social skills to fool a victim into submitting his personal information	Information	Web sites
Ramesh et al. [79]	Phishing is a fraudulent act to acquire sensitive information from unsuspecting users by masking as a trustworthy entity in an electronic commerce	Sensitive information	Not specified
Whittaker et al. [75]	A phishing page is any web page that, without permission, alleges to act on behalf of a third party with the intention of confusing viewers into performing an action with which the viewer would only trust a true agent of the third party	Not specified	Web page
Khonji et al. [22]	Phishing is a type of computer attack that communicates socially engineered messages to humans via electronic communication channels in order to persuade them to perform certain actions for the attacker's benefit	Not specified	SE
Alkhalil et al. [25]	Phishing as a socio-technical attack, in which the attacker targets specific valuables by exploiting an existing vulnerability to pass a specific threat via a selected medium into the victim's system, utilizing social engineering tricks or some other techniques to convince the victim into taking a specific action that causes various types of damages	Not specified	SE or other

- **Distribution:** During this stage, phishers deploy bait by delivering phishing materials, such as botnets or malware, to targeted victims and await their responses. [4, 17, 25, 40, 80].
- **Exploitation:** The exploitation, or penetration, phase primarily relies on manipulating human psychology [4, 33, 40, 80]. At this stage, phishers execute their malicious activities, such as collecting data or gaining unauthorized access to the target's device.
- **Exfiltration:** The final stage involves the phisher attempting to remove evidence of their activities, such as deleting fake websites and redirecting victims back to the authentic site [4, 40].

3.3 Evolution of phishing attacks

The evolution of phishing attacks traces back to the mid-1990s [44], as depicted in Fig. 4. One of the earliest recorded incidents targeted America Online (AOL), where attackers leveraged SE tactics to trick users into changing their passwords for safety purposes [17, 39, 81]. According to the APWG stolen AOL accounts were used as a form of currency among attackers to trade for hacking software [22]. The term "phishing" first printed in media on March 16, 1997 [40, 82]. Etymologically, "phishing" stems from "fishing", where a fisher (phisher) employs "bait" (e.g., a SE message) to "fish" for sensitive information [22, 38, 40, 44, 83]. The

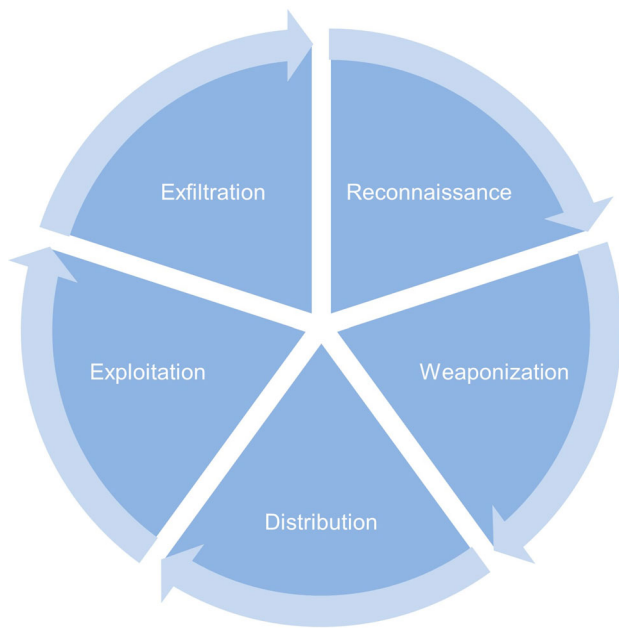


Fig. 3 Phishing Lifecycle: From Reconnaissance to Exfiltration.

substitution of 'f' with 'ph' is associated with “phone phreaking,” an attack that targeted telephone networks [22, 38, 82]. Phishing attacks have increased notably since 2016, continuing on an upward trend [80]. By the early 2000s, attackers utilized websites as their primary choice, especially mimicking online banking [44]. More recently, these attacks have expanded to focus on service providers and Software as a Service (SaaS) platforms [22, 40, 84].

Phishing grew rapidly during the period from 2014 to 2016, with an annual average increase of 97.36% [33]. Malicious URLs rose by 20% from 2017 to 2018 [40], and phishing attempts surged by 900% between 2018 and 2020 [35]. In 2020, phishing complaints saw a more than 100% increase compared to 2019 [36], with a 250% rise in unique phishing sites and a 600% spike in the first quarter alone [4, 61, 85]. This growth continued between 2020 and 2021, where unique phishing sites saw an additional 345% increase [18]. Between 2022 and 2023, phishing attacks saw a dramatic increase, with over 300,000 incidents reported in July 2023 alone [86]. According to APWG reports, 2023 was recorded as the worst year for phishing activity, with incidents reaching their highest value in Q1 2023^{1,2}. Although the number of attacks stabilised between June 2023 and March 2024², the situation escalated again thereafter. Q1 2025 registered the highest quarterly total since the 1.07 million attacks reported in Q4 2023³. Overall, phishing activity

has shown a consistent upward trend from Q2 2024 through Q2 2025⁴. Table 6 summarizes the quarterly phishing attacks detected between 2019 and 2025.

During this period, notable shifts in targeted sectors also emerged. Social media platforms became the primary target for phishing attacks in Q4 2023, a trend that continued throughout 2024. As illustrated in Fig. 5, attacks targeting social media increased sharply, in contrast to the decline in phishing attempts against financial institutions observed between Q3 and Q4 2023¹. Phone-based fraud also escalated significantly, rising by nearly 260% compared to Q4 2022, with hybrid vishing, which was rarely reported prior to 2023, becoming increasingly prevalent^{2,5}. This trend persisted throughout 2024⁴ and into 2025^{6,7}. Furthermore, more than 1.7 million unique malicious QR codes were detected across Q4 2024 and Q1 2025, followed by an additional 635,672 detections in Q2 2025^{3,4}. Following a 50% increase in BEC attack amounts in Q1 2024, fluctuations persisted throughout the year, with reported losses nearly doubling in Q4. In 2025, BEC attack values declined in Q1² but surged by 97% in Q2^{4,7}.

The financial impact of phishing has escalated over the years. In 2016, phishing caused \$9 billion in global losses [32]. In 2017, Google and Facebook were deceived into paying \$100 million [43]. Between 2016 and 2019, email-based phishing resulted in over \$26 billion in global losses [87]. By 2022, the global cost of phishing exceeded \$10.3 billion [18]. In 2023, BEC alone caused \$2.9 billion in the U.S. losses², followed by \$2.8 billion in 2024⁴. From October 2013 to December 2022, BEC attacks led to \$51 billion in global losses, as reported by the FBI's IC3¹. Table 7 summarizes quarterly averages of amounts requested in BEC wire transfer scams from 2021 to 2025.

Phishing attacks remain the most common cyberattack, and they not only cause financial losses but also lead to reputational damage and a loss of trust [44, 88]. Financial institutions remain the primary targets, with 83% of businesses affected annually [18, 35]. A notable incident was the phishing attack on OpenSea NFT marketplace users, resulting in \$1.7 million in losses [46, 47]. The complexity of phishing increases as it often serves as a gateway for launching other types of attacks [37, 65].

Despite ongoing research aimed at developing effective countermeasures, these scams remain a challenging problem without a definitive solution [34, 35, 39]. Traditional defence techniques like firewalls, signature-based, and heuristic-based have proven inadequate in protecting devices from

¹ https://docs.apwg.org/reports/apwg_trends_report_q4_2023.pdf.

² https://docs.apwg.org/reports/apwg_trends_report_q1_2024.pdf.

³ https://docs.apwg.org/reports/apwg_trends_report_q1_2025.pdf

⁴ https://docs.apwg.org/reports/apwg_trends_report_q2_2025.pdf.

⁵ https://docs.apwg.org/reports/apwg_trends_report_q2_2024.pdf.

⁶ https://docs.apwg.org/reports/apwg_trends_report_q3_2024.pdf.

⁷ https://docs.apwg.org/reports/apwg_trends_report_q4_2024.pdf.

Fig. 4 Evolution of phishing attacks from 1996 to 2025, showing an exponential growth in the number of attacks, victim count, and financial impact. Over time, phishing techniques have evolved from simple fraudulent emails to complex, socially engineered, and AI-enhanced campaigns, resulting in increasingly severe and widespread economic and organisational consequences.



Table 6 Phishing Attacks Detected by Quarter (2019-2025) .

Year	Q1	Q2	Q3	Q4	Total
2019	180,768	182,465	266,387	162,155	791,775
2020	165,722	146,994	571,764	637,302	1,521,782
2021	616,939	616,939	730,372	888,585	2,847,773
2022	1,025,968	1,097,811	1,270,883	1,350,037	4,744,699
2023	1,624,144	1,286,208	999,956	1,077,501	4,987,809
2024	963,994	877,536	932,923	989,123	3,763,576
2025	1,003,924	1,130,393	-	-	-

evolving cyber threats [5, 7, 89]. Machine learning (ML), including shallow models, especially deep learning (DL), can detect various types of attacks with minimal human intervention [7]. Recently, DL techniques have gained significant attention in the context of Intrusion Detection Systems (IDSs) [90]. Although these models have been successfully used for IDS, they typically rely on a centralized entity to process the data collected from all network users. This approach is not only costly but also introduces risks, including central

machine failure, limited computational power, and significant security and privacy concerns regarding the collected data. To address these limitations, Federated Learning (FL) emerges as a decentralized and privacy-preserving learning technique. Unlike traditional centralized ML, FL transfers the model to the data source for training without the need to share data [7, 91–94]. Nevertheless, the use of FL in cybersecurity is still in its early stages, with many practical aspects remaining unresolved [95, 96].

Fig. 5 Most Targeted Industries in Q3-Q4 2023 (APWG).

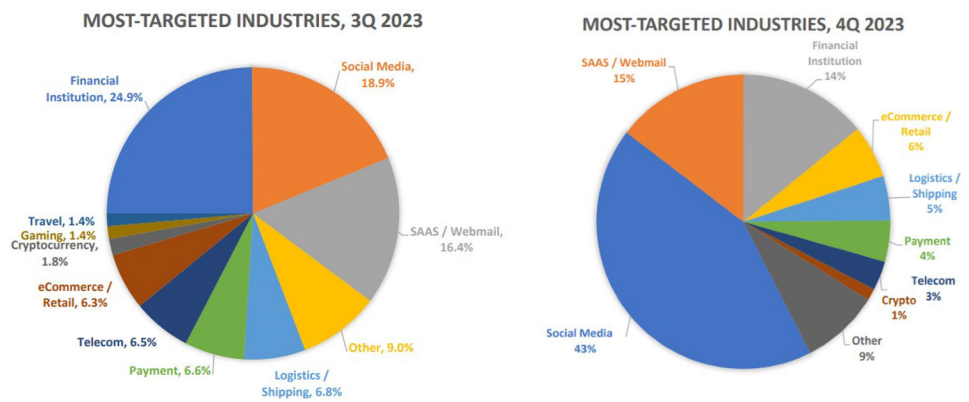


Table 7 BEC Attacks: Average Wire Transfer Request by Quarter (2021-2025).

Year	Q1	Q2	Q3	Q4	Total
2021	\$85,000	\$106,000	\$64,353	\$50,027	\$305,380
2022	\$91,436	\$109,467	\$93,881	\$132,559	\$427,343
2023	\$187,053	\$293,359	\$164,645	\$56,195	\$701,252
2024	\$84,059	\$89,520	\$67,145	\$128,980	\$369,704
2025	\$42,236	\$83,099	-	-	-

3.4 Existing phishing datasets

Several datasets are available for the experimentation and evaluation of phishing detection algorithms. Table 8 highlights some of the most commonly used phishing and ham datasets, categorized by origin, content, and usage in the literature. These datasets vary in size, feature representation, source, attack type, and targeted medium (e.g., email or websites). Accessibility and licensing conditions also differ across datasets.

While many of them are freely available for academic research, others (like the Millersmiles dataset) may require commercial licenses. Some, such as those available on Mendeley and UCI, contain engineered features derived from elements like URLs, HTML content, JavaScript, and DNS information. In contrast, datasets such as PhishTank and OpenPhish provide real-time phishing data feeds via APIs, enabling continuous updates and integration. Overall, these datasets have played a critical role in benchmarking phishing detection methodologies and continue to evolve in response to emerging cyber threats and advancements in detection techniques.

4 Taxonomy of phishing attacks techniques

Figure 6 shows the identified phishing attack types, which answers to RQ1. Based on our proposed taxonomy, we have

classified these attacks into distinct categories based on their characteristics and attack methods.

4.1 Email-based attacks

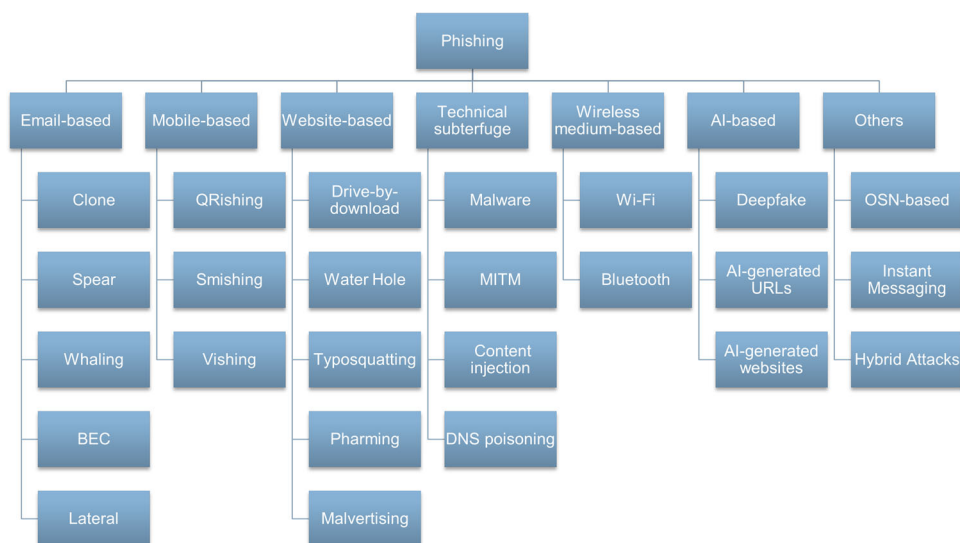
Email remains one of the most exploited channels for phishing attacks [38, 66]. A typical phishing technique involves creating a sense of urgency or scarcity, combined with well-crafted visuals and persuasive content [19, 40]. Beyond stealing sensitive information, phishing emails often serve as a good channel for malware distribution, increasing the risk of unauthorized access [86, 106]. These attacks range from generic spam to more targeted approaches, such as spear phishing, and evolve based on the perceived value of the target [38, 150].

Different types of email-based phishing include:

- Clone phishing: In this type, the attackers clone a previously sent or received legitimate email, replacing its links or attachments with malicious versions. By spoofing the sender’s email address and incorporating details relevant to the recipient, the email appears highly authentic, making it harder to detect [25, 151].
- Spear phishing: First introduced in 2005, spear phishing quickly gained attention as a highly effective attack method. Studies later revealed that it is 4.5 times more successful at compromising victims compared to general phishing attacks [97]. Unlike generic phishing, spear phishing is highly targeted and meticulously planned, focusing on specific individuals or organizations, such as government officials, business partners, or high-net-worth individuals [39, 40, 80, 152]. Spear phishing has become one of the most prevalent cybersecurity threats, accounting for over 90% of phishing-related incidents. It serves as a primary attack vector for phishing and remains a leading method for malware distribution [15, 39, 40, 97]. Phishers (Attackers) carefully craft personalized emails that seem to come from trusted sources, such as friends,

Table 8 Overview of datasets commonly used for website and email phishing research, summarising each dataset's source, a brief description, and representative studies that utilised it.

Data source	Description	Used by
PhishTank [77]	Phishing website data reported by users is accessible via an API [17, 71, 97, 98].	[79, 99–112]
APWG [113]	A record of phishing attacks reported to or detected by APWG [64, 71, 97].	[106, 109, 114, 115]
Mendeley [116–118]	[116] includes 247,950 URLs (128,541 phishing, 119,409 legitimate) with 42 features [119]. [117] contains 10,000 websites (5,000 phishing, 5,000 legitimate) with 48 features [62, 105]. [118] comprises 88,647 websites (58,000 legitimate, 30,647 phishing) with 111 features [4, 62, 120]	[121–125]
UCI [126]	Published by [127], it contains 11,055 websites (6,157 phishing, 4,898 legitimate) with 30 features from URLs, HTML, JavaScript, and DNS services [62]	[128–131]
Enron [132]	Generated by 150+ Enron employees and refined by [133], 500,000 emails are included [17, 71, 97]	[66, 134–136]
OpenPhish [137]	Utilizes an autonomous algorithm to detect zero-day phishing websites, maintaining 18 million phishing URLs with related metadata [17, 138]	[103, 105, 106, 109]
Phishload [139]	A set of phishing sites with HTML code, URLs, and other phishing-related information, along with over 1000 legitimate websites [17, 71, 97]	[102, 140, 141]
Alexa [142]	Top-ranked domains for legitimate websites, ordered by probability [62, 98, 138]	[33, 79, 100, 103, 105, 106, 109]
DMOZ [143]	A people-reviewed web directory of over 2 million benign URLs, discontinued in 2017 and replaced by Curlie [33, 62, 99]	[99, 107, 110, 112]
Kaggle [144, 145]	A publicly available website phishing datasets [80]	[102, 104, 128]
Millersmiles [146]	An archive of over 2 million spoof emails and phishing scams in the HoneyTrap database, available for commercial licensing [33]	[79, 99, 101]
SpamAssassin [147]	A collection of 4,150 ham and 1,897 spam emails from the SpamAssassin project [80, 97, 106, 148]	[106, 140, 148]
Nazario [149]	Created by Jose Nazario, the dataset contains only phishing emails [71, 80]	[66, 106]

Fig. 6 The proposed taxonomy of phishing attack techniques, categorized based on attack type.

supervisors, or reputable organizations like banks and social media platforms. These deceptive emails often incorporate personal details, including the target's name, address, or other identifiers, to avoid suspicion and enhance credibility [4, 25, 40, 80, 153]. While spear phishing requires more time and effort than traditional phishing, its success rate is significantly higher due to the tailored and convincing nature of the attacks [21, 40, 80, 82]. These attacks are becoming easier to initiate, as attackers can leverage widely available toolkits that automate and facilitate the process [154].

- Whaling phishing: Is an advanced form of spear phishing that explicitly targets high-ranking executives and decision-makers within organizations, often referred to as "big fishes" due to their privileged access to sensitive data and critical resources [21, 40, 152]. The main objective of whaling attacks is often to orchestrate large financial transactions, such as wire transfers [3]. However, these attacks can also serve as an entry point for more advanced cyber threats, including cyber extortion, by leveraging the high-level access and authority obtained through the initial breach [4, 39, 40].
- Business email compromise phishing (BEC): This type is a specialized form of spear phishing that targets government, non-profit, and commercial organizations with the intent to cause financial harm [38, 40]. Also known as a "launchpad attack", it can trigger a knock-on effect, where the compromise of one account leads to further breaches within the organization [40]. BEC shares similarities with whaling, as both focus on high-profile individuals to gain access to sensitive business data and manipulate internal processes. Attackers may alter previous emails, change meeting schedules, or contact clients and service providers [4, 21]. Unlike traditional phishing, attackers in BEC schemes do not directly steal money. Instead, they spoof executive email accounts and deceive lower-level employees into initiating fraudulent fund transfers [4]. Chief Financial Officers (CFOs) are the most frequently targeted, appearing in 41% of BEC cases [40]. A notable example is the "Fake President" attack on FACC, where attackers impersonated the CEO and tricked the finance department into transferring approximately \$61 million to the attackers' bank account [4].
- Lateral phishing: Exploits compromised internal accounts within an organization, resulting in billions of dollars in losses [153, 155]. Once attackers gain access, they establish a foothold in the organization's network, moving laterally to target sensitive data, such as executive credentials, financial records, and intellectual property, by monitoring the victim's email communications [153, 154, 156]. These attackers can remain within the organization undetected for extended periods, leveraging

compromised accounts to blend in with legitimate communications. For instance, university accounts have been misused to access restricted scholarly articles and bypass institutional restrictions. Additionally, stolen credentials are often sold on the dark web, with prices ranging from 8 to 190 per account [154]. Lateral spear-phishing is on the rise due to its high success rate. Since these attacks originate from internal email accounts, they are extremely difficult to detect, often bypassing automated security solutions, making them a significant cybersecurity challenge [153, 154].

4.2 Mobile-based attacks

- Smishing: This type operates by sending SMS or instant messages that contain malicious links or phone numbers to a victim's smartphone, aiming to deceive them into sharing sensitive information or downloading malware on their device [26, 61, 80, 151]. Once the malware (e.i. keylogger) is installed, it will grant attackers access to contacts, messages, credentials, search history, and other personal data [40, 41, 61].
- Vishing (voice phishing): Exploits voice communication channels, such as phone calls, to steal a victim's personal information [3, 39]. Attackers often conduct preliminary research on their targets, gathering details in advance to establish trust and avoid suspicion [39, 80]. Therefore, vishing is usually more successful than other phishing strategies due to the higher perceived trustworthiness of phone calls compared to other techniques [4, 21, 80, 157]. Phishers may use voice-over-internet protocol (VoIP) to make their calls; taking advantage of the low cost of this solution, mask their physical location effectively and make calls seem legitimate [21, 40, 82, 157]. Additionally, vishing can bypass two-factor authentication (2FA) by tricking victims into revealing one-time passwords (OTPs). For example, an Indian woman lost nearly 700,000 Rupees after sharing her OTP with a fraudster posing as a bank official [39].
- QR Phishing: Initially reported by Kaspersky in 2011, QRishing has grown alongside the increased use of smartphones for activities such as tracking, mobile payments, and accessing websites and apps [39, 40, 84]. Mobile users are particularly vulnerable due to small screen sizes and the everyday use of URL-shortening techniques, which obscure the entire URL and make it difficult to identify malicious links [39, 40, 158]. A key vulnerability is that QR codes are not human-readable until scanned and decoded [39, 40, 84, 159]. Compounding this risk, many QR code readers automatically execute embedded actions without requiring user approval [40]. As a result, this opens the door for malicious QR codes that impersonate legitimate ones,

redirecting users to harmful websites designed to install malware or steal sensitive information [39, 40, 159, 160]. Furthermore, QRishing becomes even more dangerous when combined with other phishing techniques, increasing its overall impact [40].

4.3 Website-based attacks

In this phishing attacks category, attackers replicate legitimate websites (i.e., Google, eBay, or PayPal) to steal users' personal and financial information [17, 26, 80, 82]. These websites can be either legitimate sites compromised and injected with malicious content or entirely fake domains owned by attackers [17]. The availability of online phishing kits and free hosting servers simplify the creation of such attacks with minimal effort [39]. Consequently, multiple sub-categories of website-based phishing attacks fall under this umbrella, including:

- **Drive-by-download:** This attack exploits browser vulnerabilities, leveraging scripting techniques such as JavaScript, to secretly install malware on a victim's device when they access a compromised or phishing website [83, 151, 152, 161], taking into consideration that these attacks do not require any direct user interaction.
- **Watering hole:** This attack is a highly targeted variant of the drive-by download attack. Instead of directly targeting victims, attackers compromise a legitimate website frequently visited by a specific individual or community. By injecting malicious code into the site, attackers exploit browser vulnerabilities to install malware on the visitors' devices, enabling unauthorized access to sensitive data [41, 152, 162].
- **Typosquatting (URL Spoofing):** The phisher exploits typographical errors users make when entering a URL, redirecting them to a malicious website. For example, mistyping "Facebok.com" instead of "Facebook.com" can lead victims to a fraudulent site. Attackers perform this attack by registering domain names that are almost identical to those of popular websites [39, 40, 161].
- **Pharming:** The attacker creates malicious websites to steal financial and sensitive information from victims [151, 161]. Additionally, they often deploy malware to manipulate local Domain Name System (DNS) settings, facilitating further attacks such as DNS poisoning and man-in-the-middle (MITM) attacks [41, 63, 162].
- **Malvertising:** Attackers utilize deceptive online advertisements to fall victims and encourage them to click on them, redirecting victims to phishing websites designed for stealing sensitive information purposes [39, 161].

4.4 Technical subterfuge

- **Malware:** Involves the execution of malicious software via links or attachments to perform unauthorized activities such as data theft, encryption, or device damage. Common malware types used in these attacks include keyloggers, trojans, spyware, and ransomware [17, 25, 63, 163].
- **Man-in-the-Middle (MITM):** The phisher positions themselves in the communication channel between the victim and the legitimate recipient, enabling them to eavesdrop, manipulate and collect personal information while maintaining a normal communication flow [4, 25, 40, 157]. Although SSL/TLS encryption typically secures online traffic, malware can modify system settings to bypass these protections and facilitate such attacks [23].
- **Content injection:** Involves inserting malicious content into legitimate websites to steal personal details or install malware. Popular methods include Cross-Site Scripting (XSS), which injects client-side scripts into web pages, and Structured Query Language (SQL) injection, which exploits database vulnerabilities through unauthorized commands [4, 25, 63, 164, 165].
- **Domain Name Server (DNS) poisoning:** Attackers establish a fake DNS server or manipulate the existing DNS table to redirect traffic to fraudulent sites or deliver malware. Further, this technique enables phishing, credential theft, and malware distribution by intercepting legitimate traffic and rerouting it to attacker-controlled sites [4, 17, 157, 164].

4.5 Wireless medium-based

- **Wi-Fi (WiPhishing):** Also known as an evil twin attack, attackers create fraudulent Wi-Fi access points that mimic legitimate public networks, often posing as free internet services. Due to the lack of encryption in many public hotspots, attackers can intercept and monitor network traffic, allowing them to steal sensitive information, capture login credentials, or inject malware into the victim's device [4, 17, 40, 157].
- **Bluetooth:** Enable attackers to establish unauthorized connections with victims' devices, particularly in crowded areas, using default or weak Bluetooth passwords or security settings to send phishing messages such as advertisements [17, 166].

4.6 AI-based attacks

With the advancement in AI technology, the malicious exploit AI to expand the scale, accelerate the attack process, and

refine the sophistication of phishing attacks, significantly increasing the chances of bypassing AI-based detection systems [4, 167, 168].

- **Deepfake:** Combining “deep learning” and “fake” represents a significant instance of offensive AI. While sometimes used for harmless purposes, they are often exploited for malicious activities, such as phishing. These technologies can mimic a victim’s voice or facial features to deceive others [61, 169–171]. In 2018, BuzzFeed released a deepfake video of former President Obama, which heightened concerns about identity theft and misinformation [171]. Deepfakes lead to substantial economic and social costs. Notably, the most significant loss reported from a single cybercrime incident involving a deepfake was approximately \$35 million [172]. In 2019, cybercriminals used deepfake voice phishing (vishing) to deceive the CEO of a UK energy company into transferring \$243,000 to fraudulent accounts [172, 173].
- **AI-generated phishing URLs:** This type leverages AI tools to generate synthetic phishing URLs, resulting in improved success rates of attacks while evading phishing detection classifiers effectively [4, 168]. Bahnsen et al. studied and demonstrated this ability in their work on DeepPhish [99, 174].
- **AI-generated phishing websites:** This type shares similarities with the earlier one by leveraging AI tools to build and create compelling phishing websites that automatically mimic legitimate ones and adapt to victim interactions, complicating detection [141].

4.7 Other attack techniques

- **Phishers exploit social media platforms,** such as Facebook, Twitter, and LinkedIn, that facilitate interaction and information sharing between users. By impersonating trusted individuals, they deceive target victims into revealing sensitive information [17, 40, 63, 80].
- **Instant Messaging (IM):** Attackers exploit chatting applications such as WhatsApp, Telegram, and Facebook Messenger, which have audio, video, hyperlinks, and file-sharing capabilities. This functionality exposes IM platforms to phishing attacks, as attackers can craft realistic messages and interactions that deceive users into disclosing sensitive information or credentials [4, 40, 157].
- **Hybrid-based:** In this approach, attackers combine different phishing tactics to increase the chances of a successful attack. For instance, in hybrid vishing, attackers send a fake purchase notification to the target, instructing them

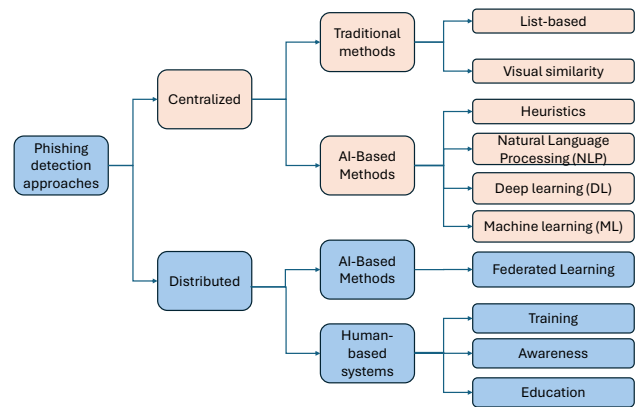


Fig. 7 Taxonomy of phishing detection approaches, classified into two main families: (i) centralized methods, which rely on consolidated data processing and model training, and (ii) distributed methods encompassing both human- and AI-based strategies, that leverage federated or decentralized architectures to detect phishing without centralizing sensitive information, and to reach broader human populations for training, awareness, and reducing vulnerability to phishing attacks.

to call a specified number to obtain a refund. After the call, the attacker uses voice communication to continue the scam and obtain sensitive information^{1, 2}.

5 Phishing detection techniques

Figure 7 shows a proposed taxonomy for phishing detection approaches. In this taxonomy, methods are classified into two main categories: 1) centralized methods, which rely on consolidated data processing and model training. 2) distributed methods, which include both human- and AI-based strategies, that leverage federated or decentralised architectures to detect phishing without centralising sensitive information, and to reach broader human populations for training, awareness, and reducing vulnerability to phishing attacks. The following sections review each group of techniques in detail, as summarised in the taxonomy.

5.1 Centralized approaches

Centralized methods for phishing detection have been widely studied. For instance, Yu et al. [175] introduced an IDS based on a DL technique called Few-shot Learning (FSL), with the NSL-KDD and UNSW-NB15 datasets. By integrating DNN and CNN models with a random sampling technique, their model outperformed traditional ML and other DL models. Similarly, Thaseen et al. [176] proposed an IDS that combines ANN with correlation-based feature selection (CFS), achieving accuracies of 98.45% on NSL-KDD and 96.44% on UNSW-NB15, outperforming other classifiers, such as DT, SVM, and RF. For phishing URL detection, Butnaru et al. [104] trained five ML models (e.g.,

RF, DT, and SVM) on Kaggle and PhishTank datasets. RF model showed the best results, even outperforming Google Safe Browsing (GSB). Yang et al. [112] presented a Multidimensional Feature Phishing Detection (MFPD) method, using CNN-LSTM to analyze URL sequences and XGBoost for classification. Their dataset was created by crawling PhishTank and DMOZ. Maini et al. [177] applied a voting ensemble method combining eight models, including RF, AdaBoost, and XGBoost, achieving 93.6% accuracy. Karim et al. [178] improved phishing detection by using a canopy feature selection method with an ensemble of LR, SVM, and DT classifiers, while Adane et al. [179] used univariate feature selection (UFS) to evaluate ensemble classifiers like RF, gradient boosting (GB), and CAT Boost (CATB), with RF achieving the fastest detection times. In phishing email detection, Chanis et al. [180] enhanced phishing email detection by integrating traditional text analysis with stylometric features to analyse email writing styles. Their stacking-based approach yielded reliable results with ML classifiers such as RF and Gradient Boosting. He et al. [181] proposed a double-layer detection method using LSTM, Bi-LSTM, and XGBoost to identify both phishing emails and insider threats. The proposed method achieved accuracy of 98.38% using the Enron dataset and phishing email dataset from *monkey.org*.

5.1.1 Similarity-based methods

Visual similarity detection typically involves comparing the content of suspicious web pages with legitimate ones. This process uses both textual (e.g., HTML and CSS) and visual (e.g., page snapshots and logos) elements [62, 80, 97]. It effectively identifies specific phishing techniques, such as Tabnabbing [39]. Similarity is measured using a variety of algorithms, such as content hashing and picture comparison. A page is considered mimicked if its similarity score exceeds a predefined threshold [34, 40, 62]. Although this approach is successful in detecting phishing embedded objects missed by heuristic techniques and eliminates the need to extract features from individual page by leveraging features that apply to the entire websites [4], it faces significant challenges such as high computational costs, large storage requirements, the need for advanced image processing tools, and significant runtime complexity [80, 87, 182]. It also struggles with zero-hour phishing attacks, has a higher false positive rate (FPR) than list-based methods, and can be bypassed through minor visual changes to phishing sites [4, 40, 64].

5.1.2 Heuristic methods

The heuristic approach uses features extracted from phishing sites, such as URLs, text content, and DNS data, to differentiate them from legitimate ones [33, 40, 64, 80]. These features are used to train classifiers for building detection models [80].

While heuristic methods are effective in detecting zero-day attacks [4, 22, 64] and have strong generalization capabilities for new phishing attempts [33, 80, 89], they are limited to a subset of common threats and cannot be applied to all newly evolving attacks [80, 89]. Another limitation is that not all phishing sites possess similar features [4, 22, 40]. Moreover, attackers who understand the detection scheme can easily evade it [4, 40], and these methods often fail to accurately identify phishing sites hosted on compromised domains [4], and often resulting in high FPR [32, 33, 80].

5.1.3 List-based methods

The list-based phishing detection approach differentiates between phishing and legitimate webpages by using collected lists of trusted and suspicious resources, such as URLs, domain names, images, and the Document Object Model (DOM) [4, 40]. These lists are typically generated through user reports or third-party detection systems and are employed by popular browsers such as Google Chrome [17, 39, 64]. This approach consists of two categories: whitelists and blacklists. Whitelists contain legitimate URLs, and any URL not included is marked as suspicious. On the other hand, blacklists contain malicious URLs; when users attempt to access these URLs, they are warned and often prevented from proceeding to the webpage [4, 64, 80]. The main advantages of this approach include its simplicity, high accuracy, and low FPR [4, 17]. However, it has notable limitations. It cannot detect zero-day phishing attacks [4], is easily bypassed with minor URL modifications, and requires frequent updates to address new phishing threats [4, 33, 44, 64].

5.1.4 Machine learning and deep learning methods

ML, a subset of AI, can automatically learn from and adapt to new data, providing experience-based solutions [27, 87, 96]. Large volumes of data are used to train ML models, which analyze features like URL structure, webpage content, and JavaScript [34, 64, 183]. ML algorithms come in a variety of types, such as reinforcement learning, supervised, semi-supervised, and unsupervised [27]. Commonly employed ML classifiers include Support Vector Machines (SVM), Neural Networks (NN), Decision Trees (DT), and Random Forests (RF) [32, 40, 80]. These algorithms can detect new threats and evolving attack patterns, as well as changes in phishing sites that might evade detection by traditional methods [40, 80, 87, 89]. Furthermore, ML models can reduce FPR and, prevent zero-day phishing attacks, making them more effective than conventional techniques like blacklist- and signature-based approaches [40, 80, 89, 108]. However, this methodology is highly dependent on the quality and size of the training dataset, the used algorithm, and the fine-tuning of hyperparameters [40, 64, 80]. Although high

accuracy of traditional ML, they are costly, rely on third-party services, and require high computation power due to manual feature engineering. Additionally, they demand extra resources and lack scalability, making them time-consuming even when applied to small datasets [17, 32, 39, 80].

DL is a subfield of ML that is derived from NN models. It can learn hierarchical representations from low-level to high-level features [27, 32, 80, 89]. Convolutional Neural Networks (CNNs), Deep Neural Networks (DNNs), and Multi-Layer Perceptrons (MLPs) are examples of popular DL architectures [64, 80, 89]. Although DL requires larger datasets and longer training times than traditional ML methods, it offers greater learning and pattern recognition capabilities. Additionally, it eliminates the need for manual feature engineering and dependency on third-party services. These advantages make DL more effective for identifying sophisticated phishing attacks, including zero-day attacks [32, 80, 87, 184]. ML and DL models with IDSs achieve high classification accuracy, particularly when used with large datasets [7, 34, 64, 93]. However, these models are centralized training techniques, which involve aggregating all datasets (like phishing and legitimate emails) into a central repository for model training [66, 93]. Centralizing such a vast amount of data is a costly process and comes with significant risks and challenges. In particular, concerns related to data protection, security, and privacy make it hard to share data and build models that can detect threats. Therefore, distributed learning methods such as FL are becoming increasingly popular. FL overcomes these constraints and allows models to be trained without centralized data aggregation [66, 93, 94, 108, 183, 184].

5.2 Decentralized approaches

In contrast to centralized learning detection techniques, in this approach, data remains local to each entity, and only model updates or insights are shared between participants, avoiding the need to centralize sensitive or private data. The goal is to collaboratively train a phishing detection model while ensuring data privacy and security.

5.2.1 Human education, awareness, and training

Human factors are a leading cause of security incidents, contributing to 95% of cases and 82% of data breaches [37, 61, 186]. Technical solutions alone are insufficient, particularly against sophisticated attacks like lateral phishing, which exploit legitimate but compromised enterprise accounts [187, 188]. This underscores the importance of a comprehensive approach that combines technological measures with user training and awareness programs [24, 39, 187]. Effective programs often include sessions, mock phishing exercises, and educational tools like games. For instance, the game

“Anti-Phishing Phil” improved participants’ ability to identify phishing by 61% [25, 82].

However, training programs face several challenges: people often forget the material they learned over time, daily training can be costly, and many programs focus narrowly on specific attack types while ignoring others. Moreover, effective training assumes that users have basic cybersecurity knowledge, which is not always true. [25, 43, 82]. Moreover, with recent deepfakes generated by AI, tools such as vishing and spear-phishing are becoming increasingly difficult to detect. In their work [189], the authors studied human ability to detect AI-generated content through a survey that included examples of various media types. The study concluded that improving users’ ability to detect AI-generated content would reduce the risk of successful phishing attacks.

5.2.2 Federated learning

Federated Learning (FL) is a distributed learning paradigm that is used to train a shared ML model across distributed edge devices (e.g., smartphones, PCs) over their data, while maintaining the data locally [66, 89, 182–184, 190]. In this sense, federated learning (FL) is adopted as the main example of decentralized learning, given its distributed nature.

Federated learning (FL) is a promising technique that has recently been applied to phishing detection and defence [28]. Federated learning has several advantages, such as privacy preserving, avoiding distributing original data for detection, allowing phishing detection at large-scale deployments, etc. The next section of the paper is fully devoted to FL phishing detection and defence techniques.

6 Federated learning for phishing detection

This section presents a landscape of federated learning works for phishing detection and countermeasures, so as comparative with centralized approaches. FL was introduced by Google researchers [29] in 2016 to mitigate key limitations of centralized training such as, data transfer cost, privacy risk, and scalability, by keeping data local and communicating only model parameters or updates with a coordinating server [7, 28, 93, 95, 191].

A standard FL round proceeds as follows:

1. The server initializes the ML model with random weights and selects a subset of clients for the training round.
2. The server sends the current model parameters (gradients, weights) to the selected clients.
3. Each client performs on-device optimization over its private data, typically using stochastic gradient descent (SGD), for a small number of local epochs.

Table 9 Summary of existing surveys on phishing detection and defense techniques (2013–2024). For each study, we report its year and main focus; coverage of countermeasure families including human factors, FL, ML, similarity-based, heuristic, list-based, and others; taxonomy elements (types, defence) and formal definitions; and whether it includes broader context (Historical Evolution, Statistics, Real-World Incidents), a Taxonomy (Types, Defence), Dataset, and Challenges. . Cells marked ✓ indicate coverage, ⊖ indicates partial/limited treatment, and blanks indicate not addressed.

Reference	Year	Main focus	Phishing countermeasures			Types	Definition	Historical Evolution	Statistics	Real-World Incidents	Taxonomy	Dataset	Challenges			
			Human											Others		
			FL	ML	Similarity									Heuristics	List	Others
Khonji et al. [22]	2013	Detection	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓				
Almomani et al. [185]	2013	Detection (mail)	✓	✓	✓	✓	⊖	⊖	✓	✓	✓	✓				
Mohammad et al. [78]	2015	Detection (website)	✓	✓	✓	✓	✓	⊖	✓	✓	✓	✓				
Varshney et al. [98]	2016	Detection (website)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓				
Aleroud et al. [23]	2017	Techniques and countermeasures	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓				
Dou et al. [33]	2017	Detection (website)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓				
Gupta et al. [97]	2017	Detection	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓				
Chiew et al. [83]	2018	Techniques	✓	✓	✓	✓	⊖	⊖	⊖	✓	✓	✓				
Goel et al. [17]	2018	Mobile attacks	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓				
Das et al. [106]	2019	Detection	✓	✓	✓	✓	✓	✓	⊖	✓	✓	✓				
Alabdian et al. [40]	2020	Techniques and countermeasures	✓	✓	✓	✓	⊖	✓	✓	✓	✓	✓				
Alkhalil et al. [25]	2021	Techniques and countermeasures	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓				

Table 9 continued

Reference	Year	Main focus	Phishing countermeasures				Types	Definition	Historical Evolution	Statistics	Real-World Incidents	Taxonomy		Dataset	Challenges
			Human		Others							Types	Defence		
			FL	ML	Similarity	Heuristics									
Basit et al. [32]	2021	Detection	✓				✓	⊖	✓		✓	✓		✓	
Salloum et al. [71]	2022	Detection (email)	✓					⊖	⊖	⊖	✓	✓	✓	✓	
Do et al. [80]	2022	Detection	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Al-Qahitani et al. [9]	2022	Attacks (during COVID-19)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Jain et al. [157]	2022	Attacks and detection	✓	✓	✓	✓	⊖	⊖	✓	⊖	✓	✓	✓	✓	
Chanti et al. [63]	2022	Attacks (classification)				✓			✓	⊖	✓	✓	✓	✓	
Naqvi et al. [37]	2023	Techniques (mitigation)	✓	✓	✓	✓	✓		✓		✓	✓	✓	✓	
Zieni et al. [62]	2023	Detection (website)	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	
Goenka et al. [4]	2023	Techniques and countermeasures	✓	✓	✓	✓	⊖	⊖	✓	✓	✓	✓	✓	✓	
Safi et al. [64]	2023	Detection (website)	✓	✓	✓	✓	✓		✓		✓	✓	✓	✓	
Varshney et al. [39]	2024	Comprehensive anti-phishing	✓	✓	✓	✓	⊖	⊖	✓	✓	✓	✓	✓	✓	
Thomopoulos et al. [82]	2024	EEG-based and eye-tracking	✓	✓	✓	✓	⊖	⊖	⊖		✓	✓	✓	✓	
<i>This survey</i>	-	-	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	

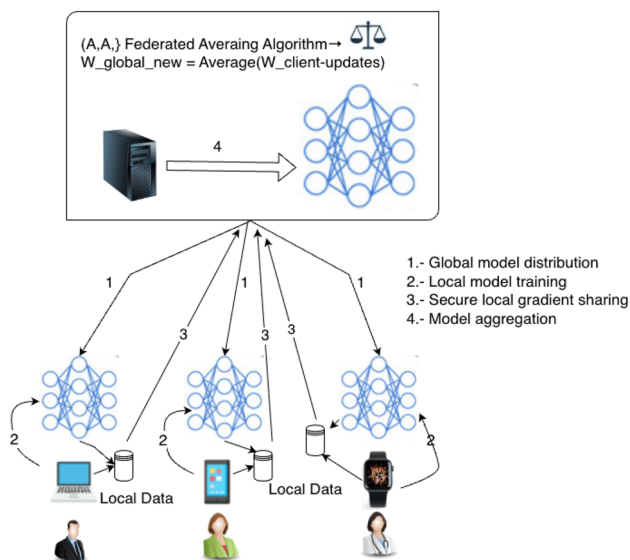


Fig. 8 Federated learning framework for phishing detection.

4. Each client reports the trained model updates (weights or gradients) to the server to construct the global ML model.
5. The server aggregates client updates, commonly via Federated Averaging (FedAvg), a data-size-weighted mean, to produce a new global model.
6. The updated global model is broadcast to clients, completing one communication round/training round; rounds repeat until convergence or an early-stopping criterion is met.

Therefore, as shown in Fig. 8, the global model in FL is updated iteratively by aggregating local models received from all nodes [182] in a privacy-preserving, secure way. This process enables resource-constrained clients to benefit from knowledge learned from other clients' data [7, 192]. Training on heterogeneous, multi-source data expands the problem space and facilitates the development of comprehensive models that generalize well [190, 192]. In the context of IDS, FL enables device-specific defence mechanisms while preserving privacy. IDS models can be trained and continuously updated with the latest attack profiles by sharing information from multiple parties [95, 193].

6.1 Federated learning for phishing detection

Federated learning provides a scalable and privacy-aware approach that improves detection performance while addressing the limitations of traditional solutions [194]. Federated learning supports phishing detection by training models directly on user devices, such as email clients or web browsers, and combining these local models to improve a shared global detection system. FL has been widely explored for phishing detection across multiple modalities,

with early work focusing on malicious URLs and more recent studies extending to email and voice. In Security Operation Centres (SOCs), Khramtsova et al. [108] trained on large, imbalanced URL corpora (e.g., URLhaus, PhishTank) and reported that FL increased the detection rate for small participants by about 30%. Building on the idea of cross-organization collaboration, Ongun et al. [195] proposed CELEST, an FL framework that couples Word2Vec and Fast-Text to detect malicious URLs at scale; CELEST achieved strong performance in production-like deployments across two university networks. Complementing these supervised approaches, Sakazi et al. [182] introduced a semi-supervised transfer learning framework (STFL) that utilizes a Bi-LSTM to adapt to heterogeneous, non-iid data and capture client-specific patterns. On three benchmarks, STFL outperformed strong centralized and FL baselines, demonstrating that leveraging unlabeled data can further improve robustness in realistic settings. To enhance detection, Adhithya et al. [196] introduced a custom attention-based classifier with residual connections, tailored for web phishing, that leverages attention mechanisms to capture intricate phishing patterns.

Beyond URL filtering, FL has also shown promise for email phishing detection, where client data are naturally siloed by the organization. Thapa et al. [66] compared RNN, BERT, and THEMIS on CSIRO and Phishbowl under varying client counts and data partitions. Under balanced data and a small number of clients, FL matched centralized learning; however, performance degraded with an increasing number of clients or higher skew. For instance, THEMIS achieved 97.9% accuracy at epoch 45, while BERT reached 96.1% at epoch 15 with five clients. Sun et al. [184] proposed FedPB, which integrates a global word-embedding layer with an LSTM, achieving 83% accuracy across different client counts and heterogeneity levels; they highlighted asynchronous FL to mitigate staleness. To address privacy leakage from gradients, Löbner et al. [135] combined FL with local differential privacy (LDP) on the 2020 Enron corpus (33,722 emails; 50.9% spam), showing that client-side F1-based thresholds can reduce typical DP accuracy losses while retaining competitive performance; per-client threshold tuning remains an open optimization. Kaushal et al. [197] presented a novel federated learning-based fair clustering technique for spam email detection, showing more effectiveness and performance than centralized methods.

FL has also been applied to voice phishing, where data imbalance and heterogeneity are pronounced. Using the KorCCVi dataset, Yoon et al. [198] grouped clients by shared characteristics to reduce skew. This balancing improved accuracy to 50% versus 40% under unbalanced assignments, underscoring the sensitivity of global models to client heterogeneity and the utility of simple stratification strategies.

Traditional spam detection methodologies often neglect user privacy preservation, potentially incurring data leak-

age risks. Across these domains, privacy preservation in FL remains a core driver. Elkhawas et al. [199] surveyed and proposed FL techniques that reduce privacy risks in phishing analytics. However, they showed that FL alone is not enough to comply to privacy regulations like GDPR, the EU AI act and privacy-preserving technology must be used in conjunction to ensure federated learning's compliance to privacy regulations. In contrast, Yoon et al. [198] demonstrated that collaborative training can protect local data yet improve models. Looking ahead, hybrid architectures can further bridge the gap between accuracy and privacy. Rose et al. [200] employed a ConvLSTM to enhance phishing detection in text-based communications under decentralized constraints, allowing for cross-organisation training without centralizing raw data. Recent advancements in large language models (LLMs) have introduced new opportunities for phishing detection by leveraging multimodal data. Li et al. LLM [201] proposed FedPhishLLM, a privacy-preserving and explainable phishing detection mechanism using federated learning and LLMs, and Hossain et al. [202] introduced a method to preserve privacy with real-time scam detection and conversational scambaiting by leveraging LLMs and Federated Learning. Very recently, federated learning and LLM have been applied to detect phishing in multilingual email. Staples et al. presented a work specifically targeting English, French, and Russian emails [203]. One remaining problem is the data distribution imbalance introduced by many FL frameworks. FPW-BC framework [204] integrates federated learning integrated with multi-feature fusion techniques, enhancing FedProx aggregation algorithm stability during server-side parameter aggregation via a horse-racing selection strategy.

Recently, the issue of securing Federated Learning (FL) systems against Byzantine adversaries capable of sabotaging model performance through malicious client behaviour has emerged as a major challenge. Singh et al. have proposed [205] an ensemble federated learning (EFL) framework to enhance federated learning security with reputation-based phishing defence. It incorporates a phishing mechanism and a Bayesian-based reputation system to effectively identify and mitigate such threats.

Federated Learning also allows to integrate multiple local models from user devices into a global model, which enhances detection accuracy compared to standalone models and ensures raw data remains secure. In this regard, Jha et al. [206] published NetHackAI, an advanced multi-model phishing-detection system using federated learning. The experimental evaluation across diverse datasets demonstrates that the proposed system achieves higher scalability, stronger adaptability to evolving threats, and better protection of sensitive user information.

Federated Learning has been applied to phishing detection in many sectors. In their work on banking systems,

Kumar et al. [207] have shown that FL enables collaborative, privacy-preserving model training across distributed banking nodes, thereby enhancing the detection of malware, phishing, and transactional anomalies. However, it has to be complemented with adaptive strategies and secure aggregation, and phased deployment to maximize effectiveness. FL is also very promising for combating vishing in banking, which is becoming increasingly worrisome due to deepfakes [208]. Healthcare is also heavily using FL. Examples include protecting privacy in data management through data silos [201], avoiding signal phishing [209], or mitigating spear-phishing [210]. However, as shown in [211], the transition from research to real-world deployment still raises several challenges.

Overall, the literature converges on a common theme: FL enables threat detection to scale across diverse parties and modalities (URLs, emails, voice) while managing non-IID data and privacy constraints, provided that algorithms and system designs explicitly address heterogeneity, communication staleness, and privacy-utility trade-offs. Taken together, these works outline a coherent path: FL enables organizations to share *signals*, not raw data; to harden models against poisoning via secure aggregation; to pair detection with *active* disruption; and to sustain privacy through DP/HE. This shifts phishing defence from isolated detectors to a coordinated, privacy-aware, and continuously improving security fabric.

6.2 Federated learning for phishing countermeasures

After phishing detection, Federated Learning has increasingly been applied to phishing prevention, extending the aim beyond traditional detection toward risk reduction, real-time threat disruption, and enabling privacy-preserving collaborative protection.

Gwassini et al. [212] introduced a comprehensive cyber-defense framework that employs FL for global cyber risk assessment across IoT-enabled smart organizations. The Cyber-XAI-Block system, combines XDRL, CapsNet, hoping IDS/IPS, and DQN-A3C to detect and mitigate threats, including phishing, achieving strong results with 97% detection and 98% prevention accuracy. FL enhances overall mitigation by aggregating risk signals into a shared global model stored in the Cyber Threat Repository (CTR). Blockchain-secured model updates protect the aggregation against tampering and poisoning attacks. The resulting global model is then redistributed to all participants, enabling adaptive security policy updates.

Hossain et al. [202] presented a privacy-preserving AI-in-the-loop framework that detects and disrupts real-time phishing and other scams using LLM-based risk scoring and automated scam-baiting. FL with differential privacy (DP) enables continuous model improvement while pre-

venting gradient-leakage attacks, achieving high engagement (around 0.80), strong relevance, and low PII leakage (≤ 0.0085). This approach shows how FL can actively counter such attacks while preserving strong privacy and safety.

Stryczek et al. [213] proposed a federation-based email-threat mitigation system (CyberDART) through the exchange of anonymized threat intelligence. Using the PATCH anonymization algorithm, decentralized nodes generate and share privacy-preserving signatures without exposing raw email content. Experiments on the TREC spam Corpus of around 90,000 emails demonstrate that combining local filters with federated intelligence improves detection performance by approximately 58%, while keeping FPR low. In other collaborative-defense work, Al-Khalisy and Al-Kateb [214] introduced MetaGuard, a proactive cyber-threat-hunting framework that leverages Federated Learning (FL) with a hybrid XGBoost–meta-learning model to detect and counter evolving attacks like spear-phishing. MetaGuard operates across distributed organizations, continuously identifying, adapting to, and responding to new threats while preserving data privacy via Differential Privacy (DP) and Homomorphic Encryption (HE).

Recent literature highlights the vulnerability of Federated Learning (FL) to poisoning attacks, where malicious actors submit fake gradients to degrade the model. Defense strategies over encrypted gradients are being investigated to protect critical networks. An example in wearable healthcare is presented by Ramahan et al. in their work "Federated Learning for Phishing Detection and Protection in Wearable Health Networks" [215].

6.3 Comparison between centralized and federated learning for phishing detection

To clearly differentiate centralized and federated phishing detection paradigms, Table 10 provides a high-level comparison across key operational, security, and evaluation dimensions. It summarizes differences related to privacy exposure and regulatory compliance, communication overhead, scalability, threat models and attack surfaces, and robustness mechanisms. As shown in the table, federated learning has many advantages over centralized methods, as it enhances data privacy by reducing data exposure and avoiding data pooling, provides better scalability by using distributed clients and avoiding data transfer costs, is more robust by using secure aggregation and differential privacy, and has better alignment with real-world settings.

Table 11 summarizes representative studies from both paradigms, including a systematic comparison between centralized and federated learning approaches for phishing detection. The table outlines the datasets used, the underlying models and techniques, the learning paradigms adopted, and the evaluation metrics reported in the literature and

addressed in this survey. As shown, recent studies mostly use the federated learning paradigm. Major evaluation metrics are accuracy, precision, and recall. Communication overhead/efficiency is covered by less works.

7 Discussion and challenges

This section presents the survey's key findings and outlines how each Research Question is addressed. The discussion integrates insights from the proposed phishing attack taxonomy, the taxonomy of detection and defence techniques, and the emerging role of FL as a privacy-preserving paradigm for phishing detection, along with the main challenges that restrict the practical deployment of FL-based solutions in real-world environments. Overall, the findings confirm that phishing remains a highly adaptive, multi-dimensional, sophisticated threat that increasingly exploits human limitations, technical vulnerabilities, and emerging AI capabilities. The examined real-world incidents, summarized in Table 4, along with the evolution timeline in Figure 4, demonstrate that phishing campaigns have expanded in delivery channels and complexity. This ongoing progression enables attackers to maintain high success rates by continually adapting their methods to increase effectiveness and evade detection. Early phishing attacks mainly used basic email impersonation, but modern attacks now shift toward highly convincing, personalized, and scalable phishing attempts spreading across a wide range of vectors.

With respect to **RQ1**, the survey shows that prior studies have introduced a wide range of taxonomies focused on phishing attacks and mitigation strategies, but these taxonomies are often focused on particular attack vectors, such as email, websites, or mobile platforms, or concentrate on specific defence approaches, such as blacklists, heuristics, or ML models. As a result, they provide only partial views of phishing rather than a comprehensive view of the phishing landscape. The comparative analysis of existing surveys in Table 9 highlights that only a limited number of works offer an integrated view that links the evolution of phishing attacks, real-world cases, phishing attack types, detection methods, datasets, and deployment challenges. In contrast, this survey brings these aspects together, providing a more holistic understanding of the phishing ecosystem.

In addressing **RQ2**, the survey finds that phishing attacks cannot be adequately described by a single classification dimension. Instead, phishing attacks are better characterized by a multi-dimensional approach that considers the attack delivery medium, the level of victim interaction, technical sophistication, and the use of automation or AI-based techniques. The comprehensive taxonomy introduced in Section 4 provides a systematic classification of these attacks into coherent categories based on their execution mechanisms and

Table 10 Comparison of centralized and federated phishing detection paradigms with respect to privacy, communication overhead, scalability, threat models, robustness mechanisms, and evaluation constraints.

Dimension	Centralized Detection	Federated Detection
Data privacy	Requires aggregation of Raw data in a central repository, increasing privacy and security risks.	Data remains local and only model updates are shared, reducing privacy exposure
Regulatory compliance	Raises compliance challenges under privacy regulations (e.g., GDPR, HIPAA)	Better aligned with privacy regulations by avoiding centralized data pooling
Communication cost / data-transfer	High data-transfer overhead is primarily associated with data aggregation and centralized storage.	Relies on iterative exchange of model updates; communication overhead can be significant
Scalability	Limited by central resources and single points of failure.	Naturally scalable across distributed clients
Threat model / Attack surface	Vulnerable to centralized data breaches, insider threats, and single-point failures.	Susceptible to poisoning, backdoor, and inference attacks on model updates
Robustness mechanisms	Relies mainly on traditional regularization and centralized security controls	Employs secure aggregation, differential privacy, and robust aggregation strategies
Evaluation realism	Commonly evaluated on controlled IID datasets, which may not reflect real-world settings	Aligns more closely with real-world settings by supporting Non-IID data distributions and client heterogeneity
Zero-day attacks	Slower as they have spread globally in the organization	Faster at detecting local, emerging campaigns that are specific to one site
Non-IID challenge	CL handles this better with a single global model	FL can suffer from model divergence, where the global model struggles to satisfy everyone

target-interaction models, encompassing email, mobile, websites, technical subterfuge, wireless, AI-driven, and hybrid attack types. This taxonomy provides a structured framework for understanding the diversity of phishing techniques and helps fill gaps in prior surveys, which often limited their scope to specific phishing vectors, such as email-only or website-only attacks. In summary, this evolving phishing landscape further emphasizes the need for more targeted, adaptive, and effective defence strategies.

In response to **RQ3**, the survey proposed a structured taxonomy of phishing detection approaches. Within this taxonomy, methods are grouped into two main categories: 1) centralized methods, which rely on consolidated data processing and model training; and 2) distributed methods, which include both human- and AI-based strategies, that leverage federated or decentralised architectures to detect phishing without centralising sensitive data, and to reach broader human populations for training, awareness, and reducing vulnerability to phishing attacks.

While human-based countermeasures such as user education, awareness, and simulation-based training remain essential, they are not sufficient on their own. Even well-trained individuals continue to fall victim to sophisticated SE attacks, especially those enhanced by generative AI, advanced impersonation techniques, and highly personalized spear-phishing. The effectiveness of training also declines over time as users often forget the material they learned; moreover, daily training sessions can be costly. Many pro-

grams additionally focus narrowly on a specific set of attack types, fail to adapt quickly to evolving adversarial tactics, or assume a level of cybersecurity knowledge that users may not have. These limitations highlight the need to complement user training with adaptive, context-aware, and automated technical defences that are capable of mitigating attacks effectively when human capabilities fall short.

Across software-based defence methods, conventional approaches, such as list-based, heuristics, and similarity-based techniques, provide fast and useful baseline protection but suffer from notable limitations. Content-similarity methods compare webpage elements to trusted references and they can detect impersonation attempts, yet they require high computational resources and storage demands, depend on sophisticated image processing tools, are vulnerable to small layout or visual changes made by attackers, struggle with zero-hour phishing attacks, and often generate higher false positive rates (FPR). Heuristic approaches rely on structural, lexical, and DNS-based indicators to detect certain zero-day attempts, but their effectiveness decreases as attackers obfuscate or modify these features, and not all phishing sites share consistent characteristics, leading to increases in FPR. Likewise, list-based methods, including both blacklists and whitelists, typically offer simplicity, high accuracy, and low FPR; However, they are easily bypassed through slight URL modifications and cannot detect newly emerging or zero-day phishing domains unless the lists are continuously and rapidly updated.

Table 11 Comparative summary of representative centralized and federated learning approaches for phishing detection, highlighting datasets, methods, learning paradigms, and evaluation metrics reported in the literature.

Reference	Dataset(s)	Model / Method	Learning Paradigm	Evaluation Metrics
Butnaru et al. [104]	PhishTank, Kaggle	RF, DT, SVM, NB, MLP	Centralized	Accuracy, Precision, Recall, F1-score, ROC, AUC
Yang et al. [112]	PhishTank, DMOZ	CNN-LSTM, XGBoost	Centralized	Accuracy, FPR, FNR, Cost, Detection Time
Adane et al. [179]	Mendeley, Kaggle	RF,GB, CATB (Hybird)	Centralized	Accuracy, Precision, Recall, F1-score, Computational time
Maini et al. [177]	PhishTank, OpenPhish, moz.com, CIC-URL	Voting ensemble (RF, AdaBoost, XGBoost, etc.)	Centralized	Accuracy, Precision, Recall, F1-score, ROC, AUC, MAE, MSE
Karim et al. [178]	Kaggle	LR, SVC, DT (Hybird)	Centralized	Accuracy, Precision, Recall, F1-score, Specificity
Chanis et al. [180]	Enron, Nazario	LR, RF, GB, DT,NB (stacked)	Centralized	Accuracy, Precision, Recall, F1-score, FPR, FNR, ROC-AUC
He et al. [181]	Enron, CERT r6.2, Mon-key.org	LSTM, Bi-LSTM, XGBoost	Centralized	Accuracy, Precision, Recall, F1-score, FPR, ROC,AUC
Thapa et al. [66]	IWSPA-AP, Nazario, Enron, CSIRO, Phish-bowl	BERT, THEMIS	Federated	Accuracy, Precision, Recall, F1-score, FPR, FNR, Communication Overhead
Lobner et al. [135]	Enron (2020)	LSTM, FedAvg, LDP (DP-SGD)	Federated	Accuracy, Precision, Recall, F1-score, Privacy budget (€), Training time
Sun et al. [184]	Enron, Microsoft 365	LSTM, global word embedding, FedAvg	Federated	Accuracy, Communication efficiency
Yoon et al. [198]	KorCCVi	Personalized FL with client grouping, FedAvg	Federated	Accuracy
Khramtsova et al. [108]	PhishTank, OpenPhish, URLHaus, Kaggle	MLP-based models	Federated	Accuracy, Loss
Sakazi et al. [182]	CPURNN, PDRCNN, Khramtsova	Bi-LSTM, FedAvg	Federated	Accuracy, Precision, Recall, F1-score
Ongun et al. [195]	University HTTP logs; Mirai, Gafgyt, Data-Exfiltration Malware	FFNN, FedAvg	Federated	Precision, Recall, PR-AUC, FPR
Li et al. [201]	Putra (2023)	Federated multimodal VLM with LoRA, FedAvg	Federated	Accuracy, Precision, Recall, F1-score
Rose et al. [200]	UCI SMS Spam	CNN-LSTM, FedAvg	Federated	Accuracy, Precision, Recall, F1-score

These shortcomings motivated the adoption of ML and DL techniques, which significantly improve detection accuracy by automatically extracting complex features, adapting to evolving phishing techniques, and reducing reliance on manual rules. As demonstrated across the studies reviewed in this survey, ML/DL models show strong capabilities in identifying zero-day attacks, reducing false-positive rates, and capturing sophisticated patterns that traditional methods often miss. However, their effectiveness remains highly dependent on the availability of large, high-quality labelled datasets, appropriate model selection, and extensive hyperparameter tuning. Traditional ML additionally requires substantial manual feature engineering and the use of external process-

ing tools, making it computationally expensive, resource-intensive, and difficult to scale in real-world environments. While DL requires larger datasets and longer training times, it offers superior learning and pattern-recognition capabilities, removing the need for manual feature engineering or reliance on third-party tools.

More critically, ML/DL systems are typically trained in a centralized setting, requiring the aggregation of sensitive user data into a single repository. This introduces significant privacy, security, and compliance risks (e.g., GDPR, HIPAA), creates operational overhead, and single points of failure that adversaries may exploit. Those challenges underscore the need for distributed learning approaches such as Federated

Learning, which preserve ML/DL performance while eliminating the requirement for centralized data collection.

Regarding **RQ3.1**, which assesses the effectiveness of applying FL for phishing detection compared with a centralized approach, FL emerges as a promising paradigm for phishing detection and collaborative defence. By enabling decentralized model training without sharing or pooling data centrally, FL inherently reduces privacy and safety risks, supports compliance with regulations such as GDPR and HIPAA, and minimizes the exposure associated with centralized data aggregation. As shown in Table 8, earlier surveys have not examined FL, despite its potential to address many limitations of traditional centralized architectures.

The reviewed studies demonstrate that FL can be effectively applied across different phishing detection domains, including URL inspection, email classification, and vishing detection, while maintaining competitive and, in some cases, superior accuracy compared to centralized learning while preserving data locality and supporting secure cross-organizational knowledge sharing. A notable advantage of FL is its ability to enable participants with limited data or computational resources to improve their detection accuracy by contributing to and benefiting from a global model trained on richer, larger client data.

FL also offers important operational advantages by removing the single point of failure inherent in centralized architectures, thus enhancing resilience against data breaches or infrastructure failures. Its decentralized design supports scalability, allowing clients to join or leave the training process without fundamental system changes. These features make FL well-suited for large-scale phishing detection in distributed environments. Overall, FL provides a privacy-preserving, collaborative, and scalable framework for phishing detection capable of achieving high performance while addressing key shortcomings of centralized approaches.

Despite these advantages, the survey findings indicate that several critical technical and operational challenges must be addressed to enable the deployment of FL-based phishing detection in real-world environments. First, data heterogeneity and non-IID distributions remain major obstacles, as phishing data varies widely across clients, leading to increased divergence, slower convergence, and lower global model accuracy. Second, Communication and resource Constraints at the edge, including limited computation, memory, and unstable connectivity, further reduce training efficiency and client consistency. In addition, high communication overhead and device and network heterogeneity restrict the scalability of FL.

FL also introduces several adversarial threats, such as poisoning, backdoor attacks, malicious clients, and gradient leakage, which could expose sensitive information and compromise the global model. Protecting against these risks requires integrating defences such as secure aggregation, DP,

and HE, although these techniques introduce trade-offs in accuracy, computational cost, and system efficiency.

Another key finding of this study is that we have identified a lack of large, realistic, distributed phishing datasets suitable for federated research. As part of our work, we have compared several datasets available for phishing detection. Table 8 shows an overview of the most popular datasets used in phishing detection, and Table 11 shows a comparative summary of representative centralized and federated learning approaches for phishing detection, highlighting datasets, methods, learning paradigms, and evaluation metrics reported in the literature. As may be seen, datasets have not changed substantially in several years. Most existing datasets are centralized, static, and do not reflect the distributional characteristics present in real environments. More effort and cooperation are still needed to create such datasets.

Overall, the insights gathered in this survey highlight that effective phishing defence requires adaptive, collaborative, and context-aware solutions integrating human expertise with ML/DL and FL strategies. Future research should focus on robustness, privacy-preserving guarantees, and the development of larger, representative datasets. Key directions include enhancing FL resilience under non-IID data, resource limitations, and adversarial conditions, alongside developing lightweight models for edge deployment. Addressing these challenges is critical for building scalable, trustworthy, and privacy-aware anti-phishing systems capable of countering the evolving threat landscape. Moreover, the challenges presented show that, while FL provides a privacy-preserving and scalable framework for phishing detection, further advancements are needed to ensure its reliability, scalability, and resilience in real-world operational environments.

8 Conclusion and future works

This survey has provided an in-depth analysis of the phishing problem, evaluated existing anti-phishing solutions, and identified their limitations across traditional and modern detection approaches. It has also offered a comparative review of existing phishing detection works, clarifying this survey's distinct focus and contributions. Additionally, we introduced two novel taxonomies of phishing techniques and countermeasures, thereby enhancing understanding of both the phishing landscape and the solution space.

To our knowledge, this is the first comprehensive survey to examine phishing detection through the specific perspective of FL, a promising direction in privacy-preserving machine learning. While traditional detection methods provide foundational protection, they often fall short against such sophisticated, adaptive tactics. In contrast, FL offers a decentralized, scalable, and privacy-aware framework for phishing detection without sharing raw data. Our review

of existing centralized and federated models demonstrates promising results, but also highlights open challenges related to data heterogeneity, non-IID distributions, adversarial robustness, and communication overhead. Thus, from our work, we can conclude that FL helps to: implement optimized secure aggregation techniques to reduce the computational overhead, improve privacy-preserving techniques for adversarial attack defenses, and achieve superior performance evaluation against separated local models.

However, current federated learning models for phishing detection still experience some limitations, and they must face several critical challenges, such as data heterogeneity and instability during server-side parameter aggregation, training instability in single neural network architectures leading to mode collapse and convergence challenges, constrained expressive capability in multi-module frameworks due to excessive complexity, and model and data communication inefficiencies.

To effectively counter future phishing threats, research should focus on enhancing the resilience, adaptability, and privacy of detection systems, with a special emphasis on FL-based systems. Key directions include improving model performance under non-IID data settings, strengthening defences against adversarial attacks, and ensuring scalability in real-world deployments. Additionally, integrating human-focused measures such as training and awareness is crucial to complement technical solutions. Advancing these areas will be essential for developing a privacy-preserving, trustworthy, and adaptive anti-phishing solution.

To achieve the best results, hybrid approaches seem to be the future trend. Some modern systems now use a Hybrid Federated approach, where public data (like known blacklists) is trained centrally, while sensitive user-specific patterns are trained via Federated Learning to provide a personalized, private shield.

Author contributions ED, JGB, SA, and JC contributed to the conception and design of the work, analysis and interpretation of results, and writing and editing of the manuscript. Papers collection, analysis, and writing of a first draft of the manuscript were performed by the first author, ED.

Funding Open access funding provided by Blekinge Institute of Technology. This work has received support from the “UC3M Center for the Analysis and Modelling of Complex Systems in Engineering and Biomedicine”, funded by the Programa Estatal para Impulsar la Investigación Científico-Técnica y su Transferencia (project reference EQC2021-007184-P).

Data Availability Not applicable.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this manuscript.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Code availability Not applicable.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Ahmad, Z., Shahid Khan, A., Wai Shiang, C., Abdullah, J., Ahmad, F.: Network intrusion detection system: A systematic study of machine learning and deep learning approaches. *Transactions on Emerging Telecommunications Technologies* **32**(1), 4150 (2021)
- Alshira’h, M., Al-Fawa’reh, M.: Detecting phishing urls using machine learning lexical feature-based analysis. *Int. J. Adv. Trends Comput. Sci. Eng* **9**(4), 5828–5837 (2020)
- Jafar, M.T., Al-Fawa’reh, M., Barhoush, M., Alshira’H, M.H.: Enhanced analysis approach to detect phishing attacks during covid-19 crisis. *Cybernetics and Information Technologies* **22**(1), 60–76 (2022)
- Goenka, R., Chawla, M., Tiwari, N.: A comprehensive survey of phishing: mediums, intended targets, attack and defence techniques and a novel taxonomy. *International Journal of Information Security*, 1–30 (2023)
- Al-Omari, M., Rawashdeh, M., Qutaishat, F., Alshira’H, M., Ababneh, N.: An intelligent tree-based intrusion detection model for cyber security. *J. Netw. Syst. Manage.* **29**, 1–18 (2021)
- Al-Fawa’reh, M., Abu-Khalaf, J., Szweczyk, P., Kang, J.J.: Malbot-drl: Malware botnet detection using deep reinforcement learning in iot networks. *IEEE Internet of Things Journal* (2023)
- Agrawal, S., Sarkar, S., Aouedi, O., Yenduri, G., Piamrat, K., Alazab, M., Bhattacharya, S., Maddikunta, P.K.R., Gadekallu, T.R.: Federated learning for intrusion detection system: Concepts, challenges and future directions. *Computer Communications* (2022)
- Karacan, H., Sevri, M.: A novel data augmentation technique and deep learning model for web application security. *IEEE Access* **9**, 150781–150797 (2021)
- Al-Qahtani, A.F., Cresci, S.: The covid-19 scamdemic: A survey of phishing attacks and their countermeasures during covid-19. *IET Inf. Secur.* **16**(5), 324–345 (2022)
- Horowitz, M.: Cyber security report, 2025. Technical report, CheckPoint (2025). <https://www.checkpoint.com/security-report/?flz-category=items&flz-item=report--cyber-security-report-2025>
- Ardito, C., Di Noia, T., Di Sciascio, E., Lofù, D., Paziienza, A., Vitulano, F.: An artificial intelligence cyberattack detection system

- to improve threat reaction in e-health. In: ITASEC, pp. 270–283 (2021)
12. Wang, S., Khan, S., Xu, C., Nazir, S., Hafeez, A.: Deep learning-based efficient model development for phishing detection using random forest and blstm classifiers. *Complexity* **2020**, 1–7 (2020)
 13. Hautsalo, J.: Using Supervised Learning and Data Fusion to Detect Network Attacks (2021)
 14. López Martínez, A., Gil Pérez, M., Ruiz-Martínez, A.: A comprehensive review of the state-of-the-art on security and privacy issues in healthcare. *ACM Comput. Surv.* **55**(12), 1–38 (2023)
 15. Abdulla, R.M., Faraj, H.A., Abdullah, C.O., Amin, A.H., Rashid, T.A.: Analysis of social engineering awareness among students and lecturers. *IEEE Access* (2023)
 16. Burda, P., Allodi, L., Zannone, N.: Cognition in social engineering empirical research: a systematic literature review. *ACM Transactions on Computer-Human Interaction* **31**(2), 1–55 (2024)
 17. Goel, D., Jain, A.K.: Mobile phishing attacks and defence mechanisms: State of art and open research challenges. *computers & security* **73**, 519–544 (2018)
 18. Shombot, E.S., Dusserre, G., Bestak, R., Ahmed, N.B.: An application for predicting phishing attacks: A case of implementing a support vector machine learning model. *Cyber Security and Applications* **2**, 100036 (2024)
 19. Spoorthi, M., Hegde, R., Soumyasri, S.: Social engineering threat: Phishing detection using machine learning approach. In: 2023 IEEE 3rd Mysore Sub Section International Conference (MysuruCon), pp. 1–7 (2023). IEEE
 20. Chung, M.-H., Yang, Y., Wang, L., Cento, G., Jerath, K., Raman, A., Lie, D., Chignell, M.H.: Implementing data exfiltration defense in situ: A survey of countermeasures and human involvement. *ACM Computing Surveys* (2023)
 21. Salahdine, F., Kaabouch, N.: Social engineering attacks: A survey. *Future internet* **11**(4), 89 (2019)
 22. Khonji, M., Iraqi, Y., Jones, A.: Phishing detection: a literature survey. *IEEE Communications Surveys & Tutorials* **15**(4), 2091–2121 (2013)
 23. Aleroud, A., Zhou, L.: Phishing environments, techniques, and countermeasures: A survey. *Computers & Security* **68**, 160–196 (2017)
 24. Patil, K., Arra, S.R.: Detection of phishing and user awareness training in information security: A systematic literature review. In: 2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM), vol. 2, pp. 780–786 (2022). IEEE
 25. Alkhalil, Z., Hewage, C., Nawaf, L., Khan, I.: Phishing attacks: A recent comprehensive study and a new anatomy. *Frontiers in Computer Science* **3**, 563060 (2021)
 26. Abdillah, R., Shukur, Z., Mohd, M., Murah, T.M.Z.: Phishing classification techniques: A systematic literature review. *IEEE Access* **10**, 41574–41591 (2022)
 27. Brandqvist, J., Lieberth Nilsson, J.: Phishing detection challenges for private and organizational users: A comparative study (2023)
 28. Li, B., Wang, P., Shao, Z., Liu, A., Jiang, Y., Li, Y.: Defending byzantine attacks in ensemble federated learning: A reputation-based phishing approach. *Futur. Gener. Comput. Syst.* **147**, 136–148 (2023)
 29. McMahan, B., Moore, E., Ramage, D., Hampson, S., Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: *Artificial Intelligence and Statistics*, pp. 1273–1282 (2017). PMLR
 30. Mäntylä, M.V., Adams, B., Khomh, F., Engström, E., Petersen, K.: On rapid releases and software testing: a case study and a semi-systematic literature review. *Empir. Softw. Eng.* **20**(5), 1384–1425 (2015)
 31. Caldiera, V.R.B.G., Rombach, H.D.: The goal question metric approach. *Encyclopedia of software engineering*, 528–532 (1994)
 32. Basit, A., Zafar, M., Liu, X., Javed, A.R., Jalil, Z., Kifayat, K.: A comprehensive survey of ai-enabled phishing attacks detection techniques. *Telecommun. Syst.* **76**, 139–154 (2021)
 33. Dou, Z., Khalil, I., Khreishah, A., Al-Fuqaha, A., Guizani, M.: Systematization of knowledge (sok): A systematic review of software-based web phishing detection. *IEEE Communications Surveys & Tutorials* **19**(4), 2797–2819 (2017)
 34. Vo Quang, M., Bui Tan Hai, D., Tran Kim Ngoc, N., Ngo Duc Hoang, S., Nguyen Huu, Q., Phan The, D., Pham, V.-H.: Shark-eyes: A multimodal fusion framework for multi-view-based phishing website detection. In: *Proceedings of the 12th International Symposium on Information and Communication Technology*, pp. 793–800. IEEE, ??? (2023)
 35. Yuan, Y., Apruzzese, G., Conti, M.: Multi-spacephish: Extending the evasion-space of adversarial attacks against phishing website detectors using machine learning. *Research and Practice, Digital Threats* (2023)
 36. Hammi, B., Zeadally, S., Nebhen, J.: Security threats, countermeasures, and challenges of digital supply chains. *ACM Computing Surveys* (2023)
 37. Naqvi, B., Perova, K., Farooq, A., Makhdoom, I., Oyedeji, S., Porras, J.: Mitigation strategies against the phishing attacks: A systematic literature review. *Computers & Security*, 103387 (2023)
 38. Ribeiro, L., Guedes, I.S., Cardoso, C.S.: Which factors predict susceptibility to phishing? an empirical study. *Computers & Security* **136**, 103558 (2024)
 39. Varshney, G., Kumawat, R., Varadharajan, V., Tupakula, U., Gupta, C.: Anti-phishing: A comprehensive perspective. *Expert Syst. Appl.* **238**, 122199 (2024)
 40. Alabdian, R.: Phishing attacks survey: Types, vectors, and technical approaches. *Future internet* **12**(10), 168 (2020)
 41. Roy, S., Sharmin, N., Acosta, J.C., Kiekintveld, C., Laszka, A.: Survey and taxonomy of adversarial reconnaissance techniques. *ACM Comput. Surv.* **55**(6), 1–38 (2022)
 42. Chaki, C., Biswas, A.: Inspecting the legal aspects of the cyber crimes committed against financial institutions with special reference to the bangladesh bank cyber intrusion
 43. Albakry, S., Vaniea, K., Wolters, M.K.: What is this url's destination? empirical evaluation of users' url reading. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–12 (2020)
 44. Li, W., Manickam, S., Laghari, S.U.A., Chong, Y.-W.: Uncovering the cloak: A systematic review of techniques used to conceal phishing websites. *IEEE Access* (2023)
 45. Alharbi, A., Dong, H., Yi, X., Tari, Z., Khalil, I.: Social media identity deception detection: a survey. *ACM computing surveys (CSUR)* **54**(3), 1–35 (2021)
 46. Gao, Y., Saad, M., Oest, A., Zhang, J., Han, B., Chen, S.: Can i own your nfts? understanding the new attack surface to nfts. *IEEE Communications Magazine* (2023)
 47. Qayyum, A., Butt, M.A., Ali, H., Usman, M., Halabi, O., Al-Fuqaha, A., Abbasi, Q.H., Imran, M.A., Qadir, J.: Secure and trustworthy artificial intelligence-extended reality (ai-xr) for metaverses. *arXiv preprint arXiv:2210.13289* (2022)
 48. Adesola, H., Chen, L., Ji, Y., Kim, J.: Application of robotics process automation to the moveit attack: A case study. In: *World Congress in Computer Science, Computer Engineering & Applied Computing*, pp. 503–515 (2024). Springer
 49. Teichmann, F.M., Boticiu, S.R.: The most impactful ransomware attacks in 2023 and their business implications. *International Cybersecurity Law Review* **5**(2), 301–311 (2024)
 50. Martínez, A.L., Naghmouchi, M., Laurent, M., Garcia-Alfaro, J., Pérez, M.G., Martínez, A.R., Nespoli, P.: Empower healthcare through a self-sovereign identity infrastructure for secure electronic health data access. *arXiv preprint arXiv:2501.12229* (2025)

51. Vang, T.: Understanding the impact of ransomware on biotechnology (2025)
52. Kanter, G.P., Rekowski, J.R., Kannarkat, J.T.: Lessons from the change healthcare ransomware attack. In: JAMA Health Forum, vol. 5, pp. 242764–242764 (2024). American Medical Association
53. Keepnet Labs: Top 15 Data Breaches of 2025 and Their Financial Impacts. <https://keepnetlabs.com/blog/top-15-data-breaches>
54. Zhou, M., Feng, L., Chen, W., Li, B., Lin, C., Wang, L.: A hybrid simulated annealing-xgboost framework with feature selection for enhanced network intrusion detection. In: 2025 10th International Conference on Electronic Technology and Information Science (ICETIS), pp. 436–441 (2025). IEEE
55. Akartuna, E.A., Yeung, F.S.W., Manning, M., Bish, A.: Shifting routines and the industrialisation of scams: the impact of covid-19 on deception crimes in hong kong. Trends in Organized Crime, 1–41 (2025)
56. Pashentsev, E.: Malicious use of AI and challenges to psychological security: Future risks. RIAC (2024)
57. Cox, K.: The 5 Biggest Phishing Attacks of 2024. <https://www.memcyco.com/the-5-biggest-phishing-attacks-of-2024/> Accessed November 6, 2025
58. Alliance, C.M.: “Sept 2025: Biggest Cyber Attacks, Ransomware Attacks and Data Breaches”. November 27, 2025. <https://www.cm-alliance.com/cybersecurity-blog/sept-2025-biggest-cyber-attacks-ransomware-attacks-and-data-breaches>
59. Computer, B.: Harvard University Discloses Data Breach Affecting Alumni, Donors. November 27, 2025. <https://www.bleepingcomputer.com/news/security/harvard-university-discloses-data-breach-affecting-alumni-donors/>
60. Technology, H.U.I.: Cybersecurity Incident Notification. November 27, 2025. <https://www.huit.harvard.edu/cyberincident>
61. Desolda, G., Ferro, L.S., Marrella, A., Catarci, T., Costabile, M.F.: Human factors in phishing attacks: a systematic literature review. ACM Computing Surveys (CSUR) **54**(8), 1–35 (2021)
62. Zieni, R., Massari, L., Calzarossa, M.C.: Phishing or not phishing? a survey on the detection of phishing websites. IEEE Access **11**, 18499–18519 (2023)
63. Chanti, S., Chithralekha, T.: A literature review on classification of phishing attacks. International Journal of Advanced Technology and Engineering Exploration **9**(89), 446–476 (2022)
64. Safi, A., Singh, S.: A systematic literature review on phishing website detection techniques. Journal of King Saud University-Computer and Information Sciences (2023)
65. Abroshan, H., Devos, J., Poels, G., Laermans, E.: Phishing happens beyond technology: The effects of human behaviors and demographics on each step of a phishing process. IEEE Access **9**, 44928–44949 (2021)
66. Thapa, C., Tang, J.W., Abuadba, A., Gao, Y., Camepe, S., Nepal, S., Almashor, M., Zheng, Y.: Evaluation of federated learning in phishing email detection. Sensors **23**(9), 4346 (2023)
67. Zhuo, S., Biddle, R., Koh, Y.S., Lottridge, D., Russello, G.: Sok: Human-centered phishing susceptibility. ACM Transactions on Privacy and Security **26**(3), 1–27 (2023)
68. Jaldá, C.S., Nanda, A.K., Pitchai, R.: Spoofing e-mail detection using stacking algorithm. In: 2022 8th International Conference on Smart Structures and Systems (ICSSS), pp. 01–04 (2022). IEEE
69. Bitaab, M., Cho, H., Oest, A., Zhang, P., Sun, Z., Pourmohamad, R., Kim, D., Bao, T., Wang, R., Shoshitaishvili, Y., *et al.*: Scam pandemic: How attackers exploit public fear through phishing. In: 2020 APWG Symposium on Electronic Crime Research (eCrime), pp. 1–10 (2020). IEEE
70. Lastdrager, E.E.: Achieving a consensual definition of phishing based on a systematic review of the literature. Crime Sci. **3**, 1–10 (2014)
71. Salloum, S., Gaber, T., Vadera, S., Shaalan, K.: A systematic literature review on phishing email detection using natural language processing techniques. IEEE Access **10**, 65703–65727 (2022)
72. Merwe, A., Loock, M., Dabrowski, M.: Characteristics and responsibilities involved in a phishing attack. In: Proceedings of the 4th International Symposium on Information and Communication Technologies, pp. 249–254 (2005)
73. Kirda, E., Kruegel, C.: Protecting users against phishing attacks with antiphish. In: 29th Annual International Computer Software and Applications Conference (COMPSAC’05), vol. 1, pp. 517–524 (2005). IEEE
74. Xiang, G., Hong, J., Rose, C.P., Cranor, L.: Cantina+ a feature-rich machine learning framework for detecting phishing web sites. ACM Transactions on Information and System Security (TISSEC) **14**(2), 1–28 (2011)
75. Whittaker, C., Ryner, B., Nazif, M.: Large-scale automatic classification of phishing pages. In: Ndss, vol. 10, p. 2010 (2010)
76. Group, A.-P.W.: Phishing Activity Trends Report 2nd Quarter 2024. https://docs.apwg.org/reports/apwg_trends_report_q2_2024.pdf
77. OpenDNS, L.: PhishTank: An Anti-Phishing Site. <https://www.phishtank.com>
78. Mohammad, R.M., Thabtah, F., McCluskey, L.: Tutorial and critical analysis of phishing websites methods. Computer Science Review **17**, 1–24 (2015)
79. Ramesh, G., Krishnamurthi, I., Kumar, K.S.S.: An efficacious method for detecting phishing webpages through target domain identification. Decis. Support Syst. **61**, 12–22 (2014)
80. Do, N.Q., Selamat, A., Krejcar, O., Herrera-Viedma, E., Fujita, H.: Deep learning for phishing detection: Taxonomy, current challenges and future directions. IEEE Access **10**, 36429–36463 (2022)
81. Paturi, R., Swathi, L., Pavithra, K.S., Mounika, R., Alekhya, C.: Detection of phishing attacks using visual similarity model. In: 2022 International Conference on Applied Artificial Intelligence and Computing (ICAIC), pp. 1355–1361 (2022). IEEE
82. Thomopoulos, G.A., Lyras, D.P., Fidas, C.A.: A systematic review and research challenges on phishing cyberattacks from an electroencephalography and gaze-based perspective. Personal and Ubiquitous Computing, 1–22 (2024)
83. Chiew, K.L., Yong, K.S.C., Tan, C.L.: A survey of phishing attacks: Their types, vectors and technical approaches. Expert Syst. Appl. **106**, 1–20 (2018)
84. Yong, K.S., Chiew, K.L., Tan, C.L.: A survey of the qr code phishing: the current attacks and countermeasures. In: 2019 7th International Conference on Smart Computing & Communications (ICSCC), pp. 1–5 (2019). IEEE
85. Abroshan, H., Devos, J., Poels, G., Laermans, E.: A phishing mitigation solution using human behaviour and emotions that influence the success of phishing attacks. In: Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization, pp. 345–350 (2021)
86. Alsubaei, F.S., Almazroi, A.A., Ayub, N.: Enhancing phishing detection: A novel hybrid deep learning framework for cybercrime forensics. IEEE Access (2024)
87. Elberri, M.A., Tokeşer, Ü., Rahebi, J., Lopez-Guede, J.M.: A cyber defense system against phishing attacks with deep learning game theory and lstm-cnn with african vulture optimization algorithm (avoa). International Journal of Information Security, 1–24 (2024)
88. Fakhouri, H.N., Alawadi, S., Awaysheh, F.M., Hamad, F., Alzubi, S., AlAdwan, M.N.: An overview of using of artificial intelligence in enhancing security and privacy in mobile social networks. In: 2023 Eighth International Conference on Fog and Mobile Edge Computing (FMEC), pp. 42–51 (2023). IEEE

89. Hu, Z., Yuan, Z.: A review of data-driven approaches for malicious website detection. arXiv preprint [arXiv:2305.09084](https://arxiv.org/abs/2305.09084) (2023)
90. Lansky, J., Ali, S., Mohammadi, M., Majeed, M.K., Karim, S.H.T., Rashidi, S., Hosseinzadeh, M., Rahmani, A.M.: Deep learning-based intrusion detection systems: a systematic review. *IEEE Access* **9**, 101574–101599 (2021)
91. Alawadi, S., Alkharabsheh, K., Alkhabbas, F., Kebande, V., Awaysheh, F.M., Palomba, F.: Fedcsd: A federated learning based approach for code-smell detection. arXiv preprint [arXiv:2306.00038](https://arxiv.org/abs/2306.00038) (2023)
92. Alkhabbas, F., Alawadi, S., Ayyad, M., Spalazzese, R., Davidsson, P.: Art4fl: An agent-based architectural approach for trustworthy federated learning in the iot. In: 2023 Eighth International Conference on Fog and Mobile Edge Computing (FMEC), pp. 270–275 (2023). IEEE
93. Sirohi, D., Kumar, N., Rana, P.S., Tanwar, S., Iqbal, R., Hijjii, M.: Federated learning for 6g-enabled secure communication systems: a comprehensive survey. *Artificial Intelligence Review*, 1–93 (2023)
94. White, J., Legg, P.: Federated learning: Data privacy and cyber security in edge-based machine learning. In: *Data Protection in a Post-Pandemic Society: Laws, Regulations, Best Practices and Recent Solutions*, pp. 169–193. Springer, ??? (2023)
95. Doriguzzi-Corin, R., Siracusa, D.: Flad: adaptive federated learning for ddos attack detection. *Computers & Security* **137**, 103597 (2024)
96. Ridwan, M.A., Radzi, N.A.M., Abdullah, F., Jalil, Y.: Applications of machine learning in networking: a survey of current issues and future challenges. *IEEE access* **9**, 52523–52556 (2021)
97. Gupta, B.B., Tewari, A., Jain, A.K., Agrawal, D.P.: Fighting against phishing attacks: state of the art and future challenges. *Neural Comput. Appl.* **28**, 3629–3654 (2017)
98. Varshney, G., Misra, M., Arey, P.K.: A survey and classification of web phishing detection schemes. *Security and Communication Networks* **9**(18), 6266–6284 (2016)
99. AlEroud, A., Karabti, G.: Bypassing detection of url-based phishing attacks using generative adversarial deep neural networks. In: *Proceedings of the Sixth International Workshop on Security and Privacy Analytics*, pp. 53–60 (2020)
100. Azeez, N.A., Misra, S., Margaret, I.A., Fernandez-Sanz, L., et al.: Adopting automated whitelist approach for detecting phishing attacks. *Computers & Security* **108**, 102328 (2021)
101. Barraclough, P.A., Fehringer, G., Woodward, J.: Intelligent cyber-phishing detection for online. *computers & security* **104**, 102123 (2021)
102. Benavides-Astudillo, E., Fuertes, W., Sanchez-Gordon, S., Rodriguez-Galan, G., Martínez-Cepeda, V., Nuñez-Agurto, D.: Comparative study of deep learning algorithms in the detection of phishing attacks based on html and text obtained from web pages. In: *International Conference on Applied Technologies*, pp. 386–398 (2022). Springer
103. Bozkir, A.S., Aydos, M.: Logosense: A companion hog based logo detection scheme for phishing web page and e-mail brand recognition. *Computers & Security* **95**, 101855 (2020)
104. Butmaru, A., Mylonas, A., Pitropakis, N.: Towards lightweight url-based phishing detection. *Future internet* **13**(6), 154 (2021)
105. Chiew, K.L., Tan, C.L., Wong, K., Yong, K.S., Tiong, W.K.: A new hybrid ensemble feature selection framework for machine learning-based phishing detection system. *Inf. Sci.* **484**, 153–166 (2019)
106. Das, A., Baki, S., El Aassal, A., Verma, R., Dunbar, A.: Sok: a comprehensive reexamination of phishing research from the security perspective. *IEEE Communications Surveys & Tutorials* **22**(1), 671–708 (2019)
107. Jain, A.K., Gupta, B.B.: Phish-safe: Url features-based phishing detection system using machine learning. In: *Cyber Security: Proceedings of CSI 2015*, pp. 467–474 (2018). Springer
108. Khramtsova, E., Hammerschmidt, C., Lagraa, S., State, R.: Federated learning for cyber security: Soc collaboration for malicious url detection. In: 2020 IEEE 40th International Conference on Distributed Computing Systems (ICDCS), pp. 1316–1321 (2020). IEEE
109. Maroofi, S., Korczyński, M., Hesselman, C., Ampeau, B., Duda, A.: Comar: classification of compromised versus maliciously registered domains. In: 2020 IEEE European Symposium on Security and Privacy (EuroS&P), pp. 607–623 (2020). IEEE
110. Prakash, P., Kumar, M., Kompella, R.R., Gupta, M.: Phishnet: predictive blacklisting to detect phishing attacks. In: 2010 Proceedings IEEE INFOCOM, pp. 1–5 (2010). IEEE
111. Rao, R.S., Pais, A.R.: Two level filtering mechanism to detect phishing sites using lightweight visual similarity approach. *J. Ambient. Intell. Humaniz. Comput.* **11**(9), 3853–3872 (2020)
112. Yang, P., Zhao, G., Zeng, P.: Phishing website detection based on multidimensional features driven by deep learning. *IEEE access* **7**, 15196–15209 (2019)
113. Group, A.-P.W.: Anti-Phishing Working Group Phishing Archive. <https://www.antiphishing.org>
114. Aburrou, M., Hossain, M.A., Dahal, K., Thabtah, F.: Predicting phishing websites using classification mining techniques with experimental case studies. In: 2010 Seventh International Conference on Information Technology: New Generations, pp. 176–181 (2010). IEEE
115. Dhamija, R., Tygar, J.D.: The battle against phishing: Dynamic security skins. In: *Proceedings of the 2005 Symposium on Usable Privacy and Security*, pp. 77–88 (2005)
116. Tamal, M.: Phishing Detection Dataset. <https://data.mendeley.com/datasets/6tm2d6sz7p/1>
117. Tan, C.L.: Phishing Dataset for Machine Learning: Feature Evaluation. <https://data.mendeley.com/datasets/h3cgnj8hft/1>
118. Vrbanič, G.: Phishing Websites Dataset. <https://data.mendeley.com/datasets/72ptz43s9v/1>
119. Tamal, M.A., Islam, M.K., Bhuiyan, T., Sattar, A.: Dataset of suspicious phishing url detection. *Frontiers in Computer Science* **6**, 1308634 (2024)
120. Vrbanič, G., Fister, I., Jr., Podgorelec, V.: Datasets for phishing websites detection. *Data Brief* **33**, 106438 (2020)
121. Maci, A., Tamma, N., Coscia, A.: Deep reinforcement learning-based malicious url detection with feature selection. In: 2024 IEEE 3rd International Conference on AI in Cybersecurity (ICAIC), pp. 1–7 (2024). IEEE
122. Maturure, P., Ali, A., Gegov, A.: Hybrid machine learning model for phishing detection. In: 2024 IEEE 12th International Conference on Intelligent Systems (IS), pp. 1–7 (2024). IEEE
123. Prince, M.S.M., Hasan, A., Shah, F.M.: A new ensemble model for phishing detection based on hybrid cumulative feature selection. In: 2021 IEEE 11th IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE), pp. 7–12 (2021). IEEE
124. Rahman, S.S.M.M., Islam, T., Jabiullah, M.I.: Phishstack: evaluation of stacked generalization in phishing urls detection. *Procedia Computer Science* **167**, 2410–2418 (2020)
125. Sarma, D., Mitra, T., Bawm, R.M., Sarwar, T., Lima, F.F., Hosain, S.: Comparative analysis of machine learning algorithms for phishing website detection. In: *Inventive Computation and Information Technologies: Proceedings of ICICIT 2020*, pp. 883–896 (2021). Springer
126. Mohammad, R., McCluskey, L.: Phishing Websites. UCI Machine Learning Repository. <https://doi.org/10.24432/C51W2X> (2012)
127. Mohammad, R.M., Thabtah, F., McCluskey, L.: An assessment of features related to phishing websites using an automated tech-

- nique. In: 2012 International Conference for Internet Technology and Secured Transactions, pp. 492–497 (2012). IEEE
128. Alsariera, Y.A., Adeyemo, V.E., Balogun, A.O., Alazzawi, A.K.: Ai meta-learners and extra-trees algorithm for the detection of phishing websites. *IEEE access* **8**, 142532–142542 (2020)
 129. Babagoli, M., Aghababa, M.P., Solouk, V.: Heuristic nonlinear regression strategy for detecting phishing websites. *Soft. Comput.* **23**(12), 4315–4327 (2019)
 130. Niranjana, A., Haripriya, D., Pooja, R., Sarah, S., Deepa Shenoy, P., Venugopal, K.: Ekrv: Ensemble of knn and random committee using voting for efficient classification of phishing. In: *Progress in Advanced Computing and Intelligent Engineering: Proceedings of ICACIE 2017, Volume 1*, pp. 403–414 (2019). Springer
 131. Sindhu, S., Patil, S.P., Sreevalsan, A., Rahman, F., AN, M.S.: Phishing detection using random forest, svm and neural network with backpropagation. In: 2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE), pp. 391–394 (2020). IEEE
 132. CALO: Enron Email Dataset. <https://www.cs.cmu.edu/enron/>
 133. Klimt, B., Yang, Y.: The enron corpus: A new dataset for email classification research. In: *European Conference on Machine Learning*, pp. 217–226 (2004). Springer
 134. Georgala, K., Kosmopoulos, A., Paliouras, G.: Spam filtering: An active learning approach using incremental clustering. In: *Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14)*, pp. 1–12 (2014)
 135. Löbner, S., Gogov, B., Tesfay, W.B.: Enhancing privacy in federated learning with local differential privacy for email classification. In: *International Workshop on Data Privacy Management*, pp. 3–18 (2022). Springer
 136. Stringhini, G., Thonnard, O.: That ain't you: Blocking spearphishing through behavioral modelling. In: *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*, pp. 78–97 (2015). Springer
 137. OpenPhish: OpenPhish Phishing Feeds. https://www.openphish.com/phishing_feeds.html
 138. Asiri, S., Xiao, Y., Alzaharani, S., Li, S., Li, T.: A survey of intelligent detection designs of html url phishing attacks. *IEEE Access* **11**, 6421–6443 (2023)
 139. Maurer, M.-E.: Phishload: The Phishing Test Database. <http://www.medien.fki.lmu.de/team/max.maurer/files/phishload/>
 140. Chin, T., Xiong, K., Hu, C.: Phishlimiter: A phishing detection and mitigation approach using software-defined networking. *IEEE Access* **6**, 42516–42531 (2018)
 141. Li, W., Manickam, S., Chong, Y.-W., Leng, W., Nanda, P.: A state-of-the-art review on phishing website detection techniques. *IEEE Access* (2024)
 142. Internet, A.: Alexa Top Sites Dataset. <http://www.alexa.com/topsites>
 143. DMOZ: The Directory of the Web. <http://www.dmoz.org/>
 144. Kaggle: Malicious URLs Dataset. <https://www.kaggle.com/datasets/sid321axn/malicious-urls-dataset>
 145. Kumar, S.: Malicious And Benign URLs. <https://www.kaggle.com/siddharthkumar25/malicious-and-benign-urls>
 146. MillerSmiles: Spoof Email and Phishing Scams List. <http://www.millersmiles.co.uk/scams.php>
 147. Apache Software Foundation: pamaassassin public corpus. Accessed: 2025-02-10 (2014). <https://spamassassin.apache.org/>
 148. Alotaibi, R., Al-Turaiki, I., Alakeel, F.: Mitigating email phishing attacks using convolutional neural networks. In: 2020 3rd International Conference on Computer Applications & Information Security (ICCAIS), pp. 1–6 (2020). IEEE
 149. Nazario, J.: Nazario Phishing Corpus. Accessed: 2025-02-11 (2015). <https://monkey.org/jose/phishing/>
 150. Zhao, J., Yan, Q., Li, J., Shao, M., He, Z., Li, B.: Timiner: Automatically extracting and analyzing categorized cyber threat intelligence from social data. *Computers & Security* **95**, 101867 (2020)
 151. Longtchi, T.T., Rodriguez, R.M., Al-Shawaf, L., Atyabi, A., Xu, S.: Internet-based social engineering psychology, attacks, and defenses: A survey. *Proceedings of the IEEE* (2024)
 152. Heartfield, R., Loukas, G.: A taxonomy of attacks and a survey of defence mechanisms for semantic social engineering attacks. *ACM Computing Surveys (CSUR)* **48**(3), 1–39 (2015)
 153. Chitara, N., Coventry, L., Nicholson, J.: “it may take ages”: Understanding human-centred lateral phishing attack detection in organisations. In: *Proceedings of the 2023 European Symposium on Usable Security*, pp. 344–355 (2023)
 154. Bhadane, A., Mane, S.B.: Detecting lateral spear phishing attacks in organisations. *IET Inf. Secur.* **13**(2), 133–140 (2019)
 155. Ho, G., Cidon, A., Gavish, L., Schweighauser, M., Paxson, V., Savage, S., Voelker, G.M., Wagner, D.: Detecting and characterizing lateral phishing at scale. In: 28th USENIX Security Symposium (USENIX Security 19), pp. 1273–1290 (2019)
 156. Yamagishi, R., Fujii, S.: An analysis of susceptibility to phishing via business chat through online survey. *Journal of Information Processing* **31**, 609–619 (2023)
 157. Jain, A.K., Gupta, B.: A survey of phishing attack techniques, defence mechanisms and open research challenges. *Enterprise Information Systems* **16**(4), 527–565 (2022)
 158. Reddy, V., Rashmi, S.: Prevalent cyber attacks and defense. In: 2023 IEEE International Conference on Integrated Circuits and Communication Systems (ICICACS), pp. 1–6 (2023). IEEE
 159. Subairu, S., Alhassan, J., Abdulhamid, S., Ojieniyi, J.: A review of detection methodologies for quick response code phishing attacks. In: 2020 2nd International Conference on Computer and Information Sciences (ICCIS), pp. 1–5 (2020). IEEE
 160. Zhang, X., Klevering, G., Lei, X., Hu, Y., Xiao, L., Tu, G.-h.: The security in optical wireless communication: A survey. *ACM Computing Surveys* (2023)
 161. Longtchi, T., Rodriguez, R.M., Al-Shawaf, L., Atyabi, A., Xu, S.: Internet-based social engineering attacks, defenses and psychology: a survey. *arXiv preprint arXiv:2203.08302* (2022)
 162. Rathod, T., Jadav, N.K., Tanwar, S., Alabdulatif, A., Garg, D., Singh, A.: A comprehensive survey on social engineering attacks, countermeasures, case study, and research challenges. *Information Processing & Management* **62**(1), 103928 (2025)
 163. Wazid, M., Das, A.K., Chamola, V., Park, Y.: Uniting cyber security and machine learning: Advantages, challenges and future research. *ICT express* **8**(3), 313–321 (2022)
 164. Gupta, B.B., Arachchilage, N.A., Psannis, K.E.: Defending against phishing attacks: taxonomy of methods, current issues and future directions. *Telecommun. Syst.* **67**, 247–267 (2018)
 165. Sabir, B., Ullah, F., Babar, M.A., Gaire, R.: Machine learning for detecting data exfiltration: A review. *ACM Computing Surveys (CSUR)* **54**(3), 1–47 (2021)
 166. Muraleedhara, P., Christo, M.S., Jaya, J., Yuvasini, D.: Any blue-tooth device can be hacked. know how? *Cyber Security and Applications* **2**, 100041 (2024)
 167. Gueembe, B., Azeta, A., Misra, S., Osamor, V.C., Fernandez-Sanz, L., Pospelova, V.: The emerging threat of ai-driven cyber attacks: A review. *Appl. Artif. Intell.* **36**(1), 2037254 (2022)
 168. Rosenberg, I., Shabtai, A., Elovici, Y., Rokach, L.: Adversarial machine learning attacks and defense methods in the cyber security domain. *ACM Computing Surveys (CSUR)* **54**(5), 1–36 (2021)
 169. Gal, S., Bulgurcu, B.: Exploring factors influencing internet users' susceptibility to deepfake phishing. In: *Proceedings of the Americas Conference on Information Systems (AMCIS)* (2024). AMCIS
 170. Mirsky, Y., Demontis, A., Kotak, J., Shankar, R., Gelei, D., Yang, L., Zhang, X., Pintor, M., Lee, W., Elovici, Y., et al.: The threat of

- offensive ai to organizations. *Computers & Security* **124**, 103006 (2023)
171. Mirsky, Y., Lee, W.: The creation and detection of deepfakes: A survey. *ACM computing surveys (CSUR)* **54**(1), 1–41 (2021)
 172. Kshetri, N.: The economics of deepfakes. *Computer* **56**(8), 89–94 (2023)
 173. Mustak, M., Salminen, J., Mäntymäki, M., Rahman, A., Dwivedi, Y.K.: Deepfakes: Deceptions, mitigations, and opportunities. *J. Bus. Res.* **154**, 113368 (2023)
 174. Bahnsen, A.C., Torroledo, I., Camacho, L.D., Villegas, S.: Deep-phish: simulating malicious ai. In: 2018 APWG Symposium on Electronic Crime Research (eCrime), pp. 1–8 (2018)
 175. Yu, Y., Bian, N.: An intrusion detection method using few-shot learning. *IEEE Access* **8**, 49730–49740 (2020)
 176. Sumaiya Thaseen, I., Saira Banu, J., Lavanya, K., Rukunuddin Ghalib, M., Abhishek, K.: An integrated intrusion detection system using correlation-based attribute selection and artificial neural network. *Transactions on Emerging Telecommunications Technologies* **32**(2), 4014 (2021)
 177. Maini, A., Kakwani, N., Ranjitha, B., Shreya, M., Bharathi, R.: Improving the performance of semantic-based phishing detection system through ensemble learning method. In: 2021 IEEE Mysore Sub Section International Conference (MysuruCon), pp. 463–469 (2021). IEEE
 178. Karim, A., Shahroz, M., Mustofa, K., Belhaouari, S.B., Joga, S.R.K.: Phishing detection system through hybrid machine learning based on url. *IEEE Access* **11**, 36805–36822 (2023)
 179. Adane, K., Beyene, B., Abebe, M.: Single and hybrid-ensemble learning-based phishing website detection: examining impacts of varied nature datasets and informative feature selection technique. *Digital Threats: Research and Practice* **4**(3), 1–27 (2023)
 180. Chanis, I., Arampatzis, A.: Enhancing phishing email detection with stylometric features and classifier stacking. *Int. J. Inf. Secur.* **24**(1), 1–16 (2025)
 181. He, D., Lv, X., Xu, X., Chan, S., Choo, K.-K.R.: Double-layer detection of internal threat in enterprise systems based on deep learning. *IEEE Transactions on Information Forensics and Security* (2024)
 182. Sakazi, I., Grolman, E., Elovici, Y., Shabtai, A.: Stfl: Utilizing a semi-supervised, transfer-learning, federated-learning approach to detect phishing url attacks. In: 2024 International Joint Conference on Neural Networks (IJCNN), pp. 1–10 (2024). IEEE
 183. Serpanos, D., Xenos, G.: Vertical federated learning in malware detection for smart cities. In: 2023 IEEE International Smart Cities Conference (ISC2), pp. 1–5 (2023). IEEE
 184. Sun, Y., Chong, N., Ochiai, H.: Federated phish bowl: Lstm-based decentralized phishing email detection. In: 2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 20–25 (2022). IEEE
 185. Almomani, A., Gupta, B.B., Atawneh, S., Meulenberg, A., Almomani, E.: A survey of phishing email filtering techniques. *IEEE communications surveys & tutorials* **15**(4), 2070–2090 (2013)
 186. Gamisch, L., Pöhn, D.: A study of different awareness campaigns in a company. In: Proceedings of the 18th International Conference on Availability, Reliability and Security, pp. 1–8 (2023)
 187. Kenneth, A., Hayashi, B.B., Lionardi, J., Richie, S., Achmad, S., Junior, F.A., *et al.*: Phishing attack awareness among college students. In: 2023 3rd International Conference on Electronic and Electrical Engineering and Intelligent System (ICE3IS), pp. 344–348 (2023). IEEE
 188. Lee, J., Tang, F., Ye, P., Abbasi, F., Hay, P., Divakaran, D.M.: 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 (2021). IEEE
 189. Madleňák, M., Hubočan, S.: Phishing 2.0: Human ability to detect ai-generated content. *Transportation Research Procedia* **93**, 1125–1132 (2026)
 190. Alawadi, S., Ait-Mlouk, A., Toor, S., Hellander, A.: Toward efficient resource utilization at edge nodes in federated learning. arXiv preprint [arXiv:2309.10367](https://arxiv.org/abs/2309.10367) (2023)
 191. Li, B., Ma, S., Deng, R., Choo, K.-K.R., Yang, J.: Federated anomaly detection on system logs for the internet of things: A customizable and communication-efficient approach. *IEEE Trans. Netw. Serv. Manage.* **19**(2), 1705–1716 (2022)
 192. Serpanos, D., Xenos, G.: Federated learning in malware detection. In: 2023 IEEE 28th International Conference on Emerging Technologies and Factory Automation (ETFA), pp. 1–4 (2023). IEEE
 193. Alansary, S.A., Ayyad, S.M., Talaat, F.M., Saafan, M.M.: Emerging ai threats in cybercrime: a review of zero-day attacks via machine, deep, and federated learning. *Knowl. Inf. Syst.* **67**(11), 10951–10987 (2025)
 194. Tabrizchi, H., Aghasi, A.: Cyber security intelligent systems based on federated learning. In: Federated Cyber Intelligence: Federated Learning for Cybersecurity, pp. 75–100. Springer, ??? (2025)
 195. Ongun, T., Boboila, S., Oprea, A., Eliassi-Rad, T., Hiser, J., Davidson, J.: Celest: federated learning for globally coordinated threat detection. arXiv preprint [arXiv:2205.11459](https://arxiv.org/abs/2205.11459) (2022)
 196. Adhithya, R., Revathi, M., *et al.*: Exploring the efficacy of federated-continual learning nodes with attention-based classifier for robust web phishing detection: An empirical investigation. In: 2024 International Conference on Advancements in Power, Communication and Intelligent Systems (APCI), pp. 1–7 (2024). IEEE
 197. Kaushal, V., Sharma, S.: Fairness-driven federated learning-based spam email detection using clustering techniques. *Neural Comput. Appl.* **37**(9), 6515–6526 (2025)
 198. Yoon, J.Y., Choi, B.J.: Privacy-friendly phishing attack detection using personalized federated learning. In: International Conference on Intelligent Human Computer Interaction, pp. 460–465 (2022). Springer
 199. Elkhawas, A.I., Chen, T.M., Gashi, I.: Privacy-preserving federated learning for phishing detection. *IEEE Technology and Society Magazine* (2025)
 200. Rose, J.D., *et al.*: Next-gen phishing detection system based on federated learning integrated cnn-lstm for sms communication. In: 2024 5th International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), pp. 367–372 (2024). IEEE
 201. Li, W., Manickam, S., Chong, Y.-W.: Fedphishllm: A privacy-preserving and explainable phishing detection mechanism using federated learning and llms. *Journal of King Saud University Computer and Information Sciences* **37**(8), 252 (2025)
 202. Hossain, I., Puppala, S., Talukder, S., Alam, M.J.: Ai-in-the-loop: Privacy preserving real-time scam detection and conversational scambaiting by leveraging llms and federated learning. arXiv preprint [arXiv:2509.05362](https://arxiv.org/abs/2509.05362) (2025)
 203. Staples, D., Cao, H., Hakak, S., Cook, P.: Multilingual phishing email detection using lightweight federated learning. In: 2025 22nd Annual International Conference on Privacy, Security, and Trust (PST), pp. 1–6 (2025). IEEE
 204. Xiong, Y., Cao, J., Chen, G.: Federated learning spam detection based on fedprox and multi-level multi-feature fusion. *Informatics* **12**(3) (2025) <https://doi.org/10.3390/informatics12030093>
 205. Singh, A.: Enhancing federated learning security with reputation-based phishing defense. In: 2025 8th International Conference on Information and Computer Technologies (ICICT), pp. 273–282 (2025). IEEE

206. Jha, A., Raj, A.: Advanced phishing detection system using federated learning. *Global Journal of Applied Sciences and Technology* (2025)
207. Kumar, M.S.: Federated learning for cyber threat detection in digital banking systems. *American Journal of AI Cyber Computing Management* **5**(4), 26–37 (2025)
208. Purkait, A., Shikhare, R., Pillai, S., Pathak, P., Yadav, V.: Combating phishing with ai: Advanced strategies for modern cyber defense. In: *Innovative Computing and Communications: Proceedings of ICICC 2025*, Volume 8, pp. 397–406. Springer, ??? (2026)
209. Chandra, P., Varshney, P., Sachan, N., Singh, L.K.: Federated learning and privacy-preserving ai for cardiac signal analysis in robotic surgery. In: *Deep Learning for Cardiac Signal Analysis in Robotic Applications*, pp. 317–330. Elsevier, ??? (2026)
210. Tsantikidou, K., Sklavos, N., Symeonidis, I.: Anonymity and privacy-enhancing mechanisms for ai-based e-health systems. In: *Transformative Cloud Computing, IoT and Extended Reality: Applications for Smart Cities, Education, Healthcare, Industry and Business*, pp. 257–273. Springer, ??? (2026)
211. Bakas, S., Li, X., Shah, P., Roth, H.R.: Federated learning in healthcare: From research to real-world deployment. *Annual Review of Biomedical Engineering* **28** (2026)
212. Gwass, O.A.H., Uçan, O.N., Navarro, E.A.: Cyber-xai-block: an end-to-end cyber threat detection & fl-based risk assessment framework for iot enabled smart organization using xai and blockchain technologies. *Multimedia Tools and Applications* **84**(23), 26527–26568 (2025)
213. Stryczek, S., Gwiazdowicz, M., Gozdecki, J., Kosek-Szott, K., Rapacz, N., Rzasa, J., Szott, S., Natkaniec, M.: Cyberdart: A corporate federation system for mitigating email threats. *IEEE Access* (2024)
214. Al-Khalisy, S.H.J., Al-Kateb, G.E.: Metaguard: A federated learning approach to hybrid xgboost and meta-learning models for proactive cyber threat hunting. *Iraqi Journal for Computer Science and Mathematics* **6**(3), 27 (2025)
215. Rahaman, M., Vanna, K., Gaurav, A., Berutu, S.S., Nedjah, N., Chui, K.T., Gupta, B.B.: Federated learning for phishing detection and protection in wearable health networks. In: *2025 24th International Symposium on Communications and Information Technologies (ISCIT)*, pp. 49–54 (2025). IEEE

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.