Journal of *Cybersecurity and Privacy*

MDPI

*Review*

# Digital Forensics and Phishing Defense: A Literature Review and Gap Analysis

**Indah Octaviani Laleb *, John Le * and Chau Nguyen**

Institute of Cybersecurity and Cryptology, School of Computing and Information Technology, University of Wollongong, Wollongong, NSW 2522, Australia; chaun@uow.edu.au

* Correspondence: iol959@uowmail.edu.au (I.O.L.); johnle@uow.edu.au (J.L.)

**Abstract**

Phishing remains a widespread and evolving cyber threat that targets human and technical vulnerabilities across email, web, mobile, and social media. Meanwhile, digital forensics has developed into a standards-driven discipline dedicated to identifying, preserving, analysing, and presenting digital evidence. Despite overlapping goals, phishing detection research and digital forensics typically operate separately. Detection efforts emphasise classification accuracy and rapid mitigation, while forensic practices prioritise evidential integrity and incident reconstruction. The analysis suggests that incorporating forensic-quality artefacts, such as Simple Mail Transfer Protocol (SMTP) headers, Domain Name System (DNS) and Transport Layer Security (TLS) traces, memory dumps, behavioural logs, metadata, and provenance records, may support attribution analysis, interpretability, and more evidentially robust incident reporting. It covers email, network, endpoint, behavioural, and legal areas to identify common shortcomings in forensic readiness, provenance preservation, and reproducibility. Based on these insights, we propose a conceptual framework that redefines digital forensics as a proactive, ongoing capability integrated into operational phishing defences. The review highlights gaps in research, such as the limited availability and validation of AI-generated phishing datasets, privacy-aware evidence management and deanonymization risks in evidence correlation, and automated workflows for handling evidence. It also suggests future directions for integrating forensic reasoning into advanced phishing mitigation systems.

**Keywords:** attribution; conceptual framework; digital forensics; forensic artefacts; generative AI (GenAI); investigation; phishing; privacy-conscious evidence

## 1. Introduction

Phishing remains one of the most prevalent and persistent cyber threats globally [1]. The Anti-Phishing Working Group (APWG) reported over one million phishing attacks in the first quarter of 2025 alone, highlighting the scale and persistence of these campaigns [2]. Over the past two decades, phishing has evolved from opportunistic mass-email scams into a scalable socio-technical attack model driven by impersonation, deception, and rapidly shifting infrastructure [3].

Recent advances in generative artificial intelligence (GenAI) have further transformed this landscape by enabling the automated creation of highly personalised and context-aware phishing content, complicating both detection and attribution [4–6]. Recent studies further suggest that traditional phishing datasets may not adequately represent modern attack characteristics, particularly AI-generated and highly personalised phishing campaigns.



Received: 2 April 2026
Revised: 30 May 2026
Accepted: 24 June 2026
Published: 2 July 2026

Pajola et al. [7] demonstrated that realistic phishing corpus generation and cross-dataset evaluation are essential for assessing model robustness under evolving threat conditions. Similarly, Nicklas et al. [8] showed that context-augmented large language models can improve semantic understanding of sophisticated phishing content beyond conventional text-classification approaches. Forensic linguistics and AI-detection approaches have also been explored to distinguish machine-generated phishing content from human-authored communication [9]. Extensive research has proposed countermeasures such as URL blacklisting, heuristic filtering, and machine learning-based detection; however, phishing attacks continue to evade these defences by exploiting user trust, adaptive infrastructure, and cross-channel delivery mechanisms [1,10]. Although transformer-based models have significantly improved detection accuracy across datasets and settings [11], these approaches primarily optimise classification performance rather than investigative reliability.

In parallel, digital forensics has evolved into a standards-driven discipline focused on identifying, preserving, analysing, and presenting digital evidence. Frameworks such as ISO/IEC 27037 and NIST guidelines emphasise evidential integrity, reproducibility, and chain-of-custody preservation [12,13]. However, forensic processes are typically reactive, initiated after an incident, whereas phishing defence mechanisms prioritise rapid detection and mitigation. This separation creates a fundamental structural mismatch: detection systems produce probabilistic alerts, whereas forensic investigations require preserved and verifiable artefacts for reconstruction, attribution, and legal scrutiny.

Recent studies indicate the potential convergence of these domains. For example, Almarzooqi et al. [14] demonstrate the evidential value of email authentication artefacts such as SPF, DKIM, and DMARC, while Dimitriadis et al. [10] propose reasoning frameworks to improve investigative coherence. Similarly, Stoykova and Franke [15] and Ogunseyi and Adedayo [16] highlight the role of validation models and cryptographic mechanisms in strengthening evidential integrity. These studies suggest that forensic principles can enhance phishing analysis by improving transparency, attribution, and evidential defensibility. Nevertheless, the existing research largely treats phishing detection and forensic investigation as separate domains, with limited integration of evidential preservation into operational detection pipelines. This structural separation persists across both research and operational environments.

This study addresses this gap by providing a systematic integration of phishing detection and digital forensic principles into a unified analytical framework that supports both operational effectiveness and evidential accountability. Existing phishing surveys primarily focus on detection accuracy, machine learning techniques, user susceptibility, or attack taxonomies, while digital forensic surveys typically examine broader investigative methodologies without concentrating on phishing-specific evidential workflows. In contrast, this review specifically examines the convergence of phishing detection and forensic readiness, emphasising provenance preservation, chain-of-custody, cross-layer artefact correlation, and evidential accountability across phishing investigation lifecycles.

This review intentionally focuses on phishing studies that contain explicit forensic or evidential relevance, including artefact preservation, provenance tracking, integrity assurance, attribution support, or forensic-ready investigation workflows. The objective is not to comprehensively review phishing detection accuracy literature, but rather to examine how evidential and forensic principles are integrated into phishing defence and investigation ecosystems.

### *1.1. Research Question (RQ)*

In what ways can digital forensic frameworks and methodologies enhance phishing detection, investigation, and attribution across technological and human dimensions?

Table 1 compares the scope of this review with selected prior surveys in phishing and digital forensics.

**Table 1.** Comparison between existing phishing and digital forensics surveys and the scope of this review.

| Study | Primary Focus | Forensic Integration | AI Phishing Coverage | Provenance/ Chain-of-Custody | Identified Limitation |
|---|---|---|---|---|---|
| Khonji et al. [1] | Phishing detection survey | Limited | No | No | Primarily detection-centric with limited forensic discussion |
| Casino et al. [17] | General digital forensics | General DF only | No | Partial | Not specifically focused on phishing investigation workflows |
| Alkhalil et al. [18] | Phishing attack taxonomy and anatomy | No | Limited | No | Does not address forensic readiness or evidential preservation |
| This review | Integrated phishing detection and forensic readiness | Yes | Yes | Yes | Provides cross-layer evidential synthesis and conceptual forensic-ready architecture |

To operationalise this objective, the review further considers the following sub-questions:

- What forensic artefacts are most commonly utilised in phishing investigations?
- How are integrity and provenance preservation mechanisms incorporated into phishing detection workflows?
- What challenges exist in integrating forensic readiness into operational phishing defence systems?
- How do cloud environments, identity systems, and generative AI affect evidential collection and attribution?
- What privacy and governance constraints affect forensic-ready phishing defence pipelines?

To answer this question, this review synthesises prior studies, identifies conceptual and methodological gaps, and introduces a unifying framework called PhishTrace, which emphasises forensic reasoning in phishing mitigation.

### *1.2. Contributions*

This paper makes four key contributions:

(1) A forensic-layer evidence taxonomy that organises phishing-related artefacts across email, network, endpoint, behavioural, cloud identity, and legal domains, explicitly linking artefacts to investigative claims such as provenance, reconstruction, attribution, and evidential integrity;
(2) A reproducible inclusion and evidential coding framework that enables systematic comparison of prior studies based on artefact specificity, integrity controls, provenance modelling, and evaluation context;
(3) A systematic identification of critical research gaps in phishing detection and digital forensics integration, including limitations in cross-layer provenance, cloud identity artefact modelling, inter-layer correlation, automated chain-of-custody processes, and privacy-aware evidence handling;
(4) The PhishTrace framework, a conceptual forensic-ready reference architecture that illustrates how phishing detection alerts may be transformed into structured, multi-layer evidence bundles with integrated provenance, integrity preservation, and chain-of-custody support.

The remainder of this manuscript is organised as follows. Section 2 details the review methodology, including the search strategy, inclusion criteria, and forensic relevance coding framework. Section 3 presents the background concepts, including digital forensics, phishing, and the distinction between evidential investigation and operational detection. Section 4 presents the thematic synthesis of the literature and key findings. Section 5 introduces the proposed PhishTrace conceptual reference architecture. Section 6 discusses the implications of the findings, highlights research gaps, and outlines future validation directions. Finally, Section 7 presents the conclusions of the study.

## 2. Methodology

The review combines systematic search and evidential coding procedures with thematic synthesis to analyse the integration of phishing detection and digital forensic practices. This study adopts a structured integrative review informed by established systematic literature review (SLR) and survey methodologies widely applied in cybersecurity and software engineering research [19], while the objective extends beyond descriptive aggregation to evidential taxonomy development and synthesis model construction, the review protocol follows recognised principles of transparency, reproducibility, and methodological rigour, consistent with contemporary systematic review reporting frameworks such as PRISMA [20].

The review process comprised four stages: (i) search strategy formulation and database retrieval; (ii) eligibility screening based on predefined inclusion and exclusion criteria; (iii) structured data extraction and evidential coding; and (iv) thematic clustering and synthesis. Figure 1 summarises the overall structured integrative review methodology and illustrates how studies progressed from identification and screening through evidential coding and thematic synthesis.

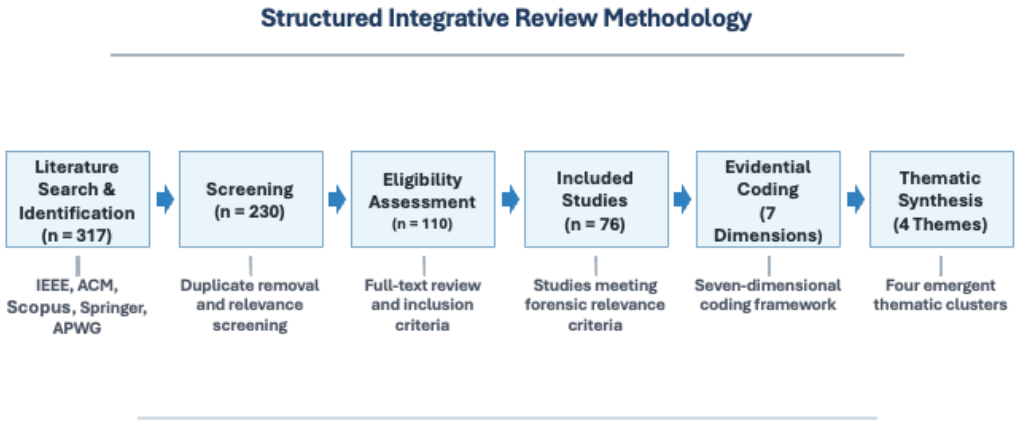


**Figure 1.** Structured integrative review methodology for phishing and digital forensics research (2004–2026).

The evidential coding stage shown in Figure 1 was designed to systematically capture forensic, evidential, and methodological characteristics across the included studies. Each paper was coded using a seven-dimensional framework encompassing phishing channels, forensic layers, artefact types, automated evidence capture, integrity controls, evaluation settings, and reported limitations.

To illustrate how the coding framework was applied in practice, Table 2 presents representative examples of coded studies across different forensic contexts. The column "Evidence Integrity Consideration" summarises how each study addressed evidential reliability, authenticity, preservation, attribution, or integrity-related concerns. These examples are intended to demonstrate the operationalisation of the coding framework rather than

provide a quantitative assessment of study quality or forensic maturity. The complete evidential coding matrix for all 76 included studies is provided in Supplementary Table S1.

**Table 2.** Illustrative evidential coding structure applied to selected studies.

| Study | Forensic Layer | Evidence Integrity Consideration | Key Evidential Focus |
|---|---|---|---|
| Almarzooqi et al. [14] | Email Forensics | Email authentication verification | SMTP headers, SPF/DKIM/DMARC analysis |
| Sikos [21] | Network evidence reconstruction | Limited integrity discussion | Packet analysis and network reconstruction |
| Zhao et al. [22] | Behavioural Forensics | Behavioural attribution support | Login behaviour analysis and account compromise attribution |
| Khan et al. [23] | Cloud Forensics | Audit-trail preservation | Cloud log collection, audit trails, and evidence preservation |
| Stoykova & Franke [15] | Legal/Evidential | Evidential reliability validation | Reliability validation and evidential integrity |
| Ogunseyi & Adedayo [16] | Legal/Evidential | Cryptographic integrity protection | Cryptographic integrity, privacy protection, and evidence assurance |

### *2.1. Search Strategy and Data Sources*

Relevant literature was retrieved from major academic databases, including IEEE Xplore, ACM Digital Library, Elsevier/Scopus, SpringerLink, and the Anti-Phishing Working Group (APWG) archives. Database searches and screening activities were conducted between January and March 2026. The review covered publications from 2004 to 2026, capturing foundational research in phishing detection and digital forensics, as well as recent developments in AI-driven phishing, forensic automation, and privacy-preserving infrastructures.

Search queries were constructed by combining core technical and evidential terms, including:

- phishing, spear-phishing, smishing, vishing
- digital forensics, email artefacts, packet capture, memory forensics
- chain of custody, attribution, forensic readiness
- automated evidence collection, SIEM, SOAR

Boolean operators and database-specific filters were applied to balance comprehensiveness and relevance. Table 3 provides example search queries used during database retrieval.

**Table 3.** Example database search queries used in the review.

| Database | Example Search Query |
|---|---|
| Scopus | (“phishing” OR “spear-phishing”) AND (“digital forensics” OR “forensic readiness”) AND (“chain of custody” OR “attribution” OR “provenance”) |
| IEEE Xplore | (“phishing detection”) AND (“email forensics” OR “network forensics”) AND (“evidence preservation”) |
| ACM Digital Library | (“AI phishing” OR “GenAI phishing”) AND (“digital evidence” OR provenance) AND (attribution OR reconstruction) |

*(phishing OR spear-phishing OR smishing OR vishing) AND (“digital forensics” OR “forensic readiness” OR attribution OR “chain of custody”) AND (email OR network OR endpoint OR cloud OR identity)*

The initial search yielded 317 records. After duplicate removal, 230 unique studies remained for title and abstract screening (see Figure 1). Following preliminary screening,

110 articles were subjected to full-text assessment based on predefined forensic and evidential inclusion criteria. Ultimately, 76 studies met all eligibility requirements and were included in the final synthesis. Of the 76 included studies, the majority focused on email and network artefacts, while substantially fewer studies examined cloud identity evidence, provenance preservation mechanisms, or automated chain-of-custody controls. Common exclusion reasons included lack of phishing relevance, absence of forensic or evidential discussion, insufficient methodological transparency, duplicate publications, and studies focused solely on classification accuracy without evidential preservation considerations. Screening and evidential coding were initially performed by the primary researcher. Inclusion decisions, thematic categorisation, and evidential classifications were subsequently reviewed during supervisory discussions to improve consistency and reduce interpretative bias. Disagreements regarding study eligibility or evidential relevance were resolved through discussion and consensus.

To provide additional insight into the characteristics of the reviewed corpus, Table 4 summarises the temporal distribution of the included studies.

**Table 4.** Temporal distribution of the studies included in the review ($n$ = 76).

| Period | Studies ($n$) | Percentage |
|---|---|---|
| 2004–2010 | 10 | 13.2% |
| 2011–2015 | 13 | 17.1% |
| 2016–2020 | 19 | 25.0% |
| 2021–2026 | 34 | 44.7% |
| Total | 76 | 100% |

Table 4 shows a steady increase in research activity over time. Nearly half of the included studies (44.7%) were published during 2021–2026, demonstrating substantial growth in scholarly attention to phishing threats and digital forensic investigations. This increase coincides with growing research interest in forensic readiness, cloud-based environments, automated evidence acquisition, privacy-preserving technologies, and AI-enabled phishing attacks. The concentration of recent studies suggests an increasing recognition of the importance of evidential preservation, attribution, and forensic-ready security practices within contemporary cybersecurity research.

### 2.2. Quantitative Summary of Evidential Coding

To complement the thematic synthesis, a descriptive analysis of the evidential coding matrix was conducted across the 76 included studies. Each included study could be assigned to multiple evidential categories where applicable; counts were derived directly from the structured coding matrix used during data extraction. Because individual studies often addressed multiple evidential dimensions simultaneously, the categories presented in Table 5 are not mutually exclusive and therefore do not sum to 100%.

**Table 5.** Descriptive summary of evidential coding categories across the included studies ($n$ = 76).

| Category | Studies ($n$) | Percentage |
|---|---|---|
| AI-/ML-based phishing analysis | 31 | 40.8% |
| Email-related artefacts (headers, messages, authentication records) | 19 | 25.0% |
| Network/web artefacts (URLs, domains, websites, traffic) | 14 | 18.4% |
| Cloud identity and log artefacts | 12 | 15.8% |
| Attribution and investigation support | 12 | 15.8% |
| Memory/endpoint artefacts | 4 | 5.3% |
| Explicit integrity or provenance controls | 1 | 1.3% |

Table 5 presents a descriptive overview of the evidential coding categories identified across the reviewed literature. Because individual studies frequently addressed multiple evidential dimensions simultaneously, the reported categories are not mutually exclusive and therefore do not sum to 100%. For example, a single study may simultaneously address email artefacts, attribution analysis, and cloud identity evidence, and is therefore counted in each applicable category. The integrity/provenance category was coded only when studies explicitly implemented or evaluated integrity-preservation mechanisms such as hashing, provenance metadata, secure timestamping, or chain-of-custody controls as part of the primary research contribution. Studies that discussed the evidential integrity conceptually, reviewed integrity requirements, or referenced chain-of-custody principles without explicit implementation or evaluation of integrity-preservation mechanisms were not included in this category.

### *2.3. Inclusion Criteria: Forensic/Evidential Link*

In addition to topical relevance to phishing or digital forensics, studies were included only if they demonstrated a clear forensic or evidential linkage. This requirement was operationalised as at least one of the following:

- Explicit discussion of forensic readiness (i.e., preparing systems and workflows to preserve investigation-ready artefacts prior to incidents);
- Automated or semi-automated artefact acquisition or triage (e.g., SIEM/SOAR playbooks, email-gateway quarantine pipelines, endpoint detection and response evidence bundles);
- Evidence integrity mechanisms such as hashing, provenance metadata, secure timestamping, or chain-of-custody logging;
- Reconstruction, attribution, or suitable for evidential review reporting processes;
- Alignment with recognised forensic standards or models (e.g., NIST or ISO guidance).

This evidential criterion distinguishes the present review from detection-centric surveys that focus solely on classification performance metrics.

### *2.4. Exclusion Criteria*

Studies were excluded if they focused exclusively on phishing classification accuracy (e.g., machine learning model comparisons) without addressing artefact preservation, provenance, investigative workflows, attribution mechanisms, or evidential and legal considerations. Studies lacking methodological transparency or clearly defined evaluation procedures were also excluded.

### *2.5. Data Extraction and Evidential Coding Framework*

Data extraction followed a structured coding protocol designed to ensure reproducibility and analytical consistency. For each included study, bibliographic metadata and technical-evidential attributes were recorded across seven dimensions:

1. Phishing channel (email, web, SMS/smishing, voice/vishing, social media).
2. Forensic layer (email, network, endpoint, behavioural, legal/evidential).
3. Artefact types (e.g., SMTP headers, DNS logs, packet captures, memory dumps, clickstreams).
4. Automated artefact capture (presence and mechanism specified).
5. Integrity controls (hashing, timestamping, provenance metadata, chain-of-custody logging).
6. Evaluation setting (laboratory, case study, operational deployment).
7. Stated limitations (e.g., encryption, cloud opacity, privacy constraints, adversarial evasion).

This coding scheme enabled systematic cross-layer comparison and structured gap identification across the selected corpus.

### 2.6. Thematic Synthesis

Following structured coding, studies were grouped into thematic clusters aligned with core forensic domains:

- Email artefact analysis: SMTP headers, authentication trails, and message-body forensics.
- Network and endpoint investigations: packet capture, memory analysis, and disk artefacts.
- Behavioural and social engineering forensics: user interaction patterns and manipulation indicators.
- Legal and evidential frameworks: chain of custody, admissibility requirements, and privacy regulation.

This thematic organisation provides the analytical foundation for Section 4, enabling progression from artefact-level analysis to integrated forensic-ready phishing defence synthesis.

### 2.7. Threats to Validity

As with any structured literature review, several limitations must be acknowledged. First, although the search strategy was designed to be comprehensive, relevant studies may have been omitted due to database coverage limitations or language restrictions. Second, the forensic/evidential inclusion criterion, while clearly defined, involves interpretative judgment in determining whether a study sufficiently demonstrates evidential linkage. Third, the rapid evolution of phishing techniques and AI-driven attack methods may result in emerging practices not yet reflected in peer-reviewed literature.

To mitigate these threats, screening and coding decisions were applied consistently using predefined criteria, and thematic synthesis was iteratively refined to ensure alignment between extracted artefacts, evidential claims, and research gaps. Nevertheless, the heterogeneity of methodologies across included studies may limit direct comparability, and future empirical validation of the proposed synthesis model is required in operational forensic environments.

## 3. Background

This section establishes the foundational concepts required for the subsequent synthesis, including digital forensics, phishing, and the distinction between evidential investigation and operational detection.

### 3.1. Background & Definitions

#### 3.1.1. Digital Forensics

Institutional organisations like INTERPOL [24], US-CERT [25], and the House of Lords [26] highlighted the importance of procedural rigour, thorough documentation, and maintaining the chain of custody. Academic viewpoints add that methodological reproducibility, reliable tools, evidential trustworthiness, and structured investigative frameworks are crucial [27,28]. Digital forensics has developed into a multidisciplinary field that involves various evidence sources and investigative methods. It continues to face cross-domain challenges, including data acquisition, maintaining evidence integrity, and ensuring legal compliance across contexts [17]. Digital forensics spans network, memory, disk, mobile, and cloud platforms, blending scientific principles with evidential law. Access control restrictions in cloud systems add to the complexity of forensic investigations [29]. Recent studies also highlight that anti-forensic tactics such as artefact wiping, log manipulation, encryption abuse, and trail obfuscation continue to undermine the reliability

and integrity of cybercrime investigations, emphasising the importance of forensic-ready evidence preservation mechanisms [30]. Its goal extends beyond detection to reconstruct events, offering defensible accounts suitable for courts or regulators. Thus, this review sees digital forensics as a lifecycle-based discipline emphasising evidence adequacy, integrity, and accountability during cyber incident investigations.

This working definition captures the current state of the art and the stance of this survey: Digital forensics is a structured, interdisciplinary process that involves collecting, preserving, analysing, and reporting digital evidence from devices and systems, with the aim of supporting investigations, maintaining evidence integrity, and enabling legal or organisational responses to cyber incidents.

#### 3.1.2. Phishing Attacks

Phishing has evolved from opportunistic email scams into a scalable socio-technical attack that relies on impersonation and deception [3]. The literature on defining phishing reflects this evolution with increasing clarity. Jakobsson and Myers [3] describe phishing as a form of social engineering where fake electronic communications are used to obtain credentials fraudulently. This approach already includes elements of scalability and automation. The APWG builds on this by depicting a crime that combines social engineering with technical tactics to steal identity and financial data, highlighting both the psychological and technical aspects of the crime [2]. In a thorough review, Lastdrager [31] examined various definitions from the literature and identified a shared understanding of phishing as "a scalable act of deception whereby impersonation is used to obtain information from a target." This definition effectively brings together historical practices and modern industrialisation by highlighting both the scalability of attacks and the importance of impersonation. The etymological and historical record also sheds more light on the evolution, with the term emerging from hacker jargon that reinterpreted "fishing" as "phishing," and the earliest cases on AOL in the 1990s shifting from exploiting weak verification to impersonating staff through spoofed emails and instant messages, laying the groundwork for a professionalised ecosystem of roles, kits, and underground markets [3]. Recent analyses mark the next stage, where generative AI enables rapid personalisation, stylistic mimicry, and cross-channel lures including smishing, vishing, and deep-fakes, developments that increase the burden on both detection and subsequent forensic reconstruction [4,9]. Against this lineage, this survey adopts the following working definition to guide scope and evaluation: Phishing is an evolving cybercrime that uses deception, impersonation, and socio-technical subterfuge to steal sensitive information or credentials.

Phishing operates as a lifecycle with distinct phases. Creation sets up the infrastructure and lures, delivery spreads these lures across email, SMS, voice, and social media platforms, exploitation captures login credentials, launches malware, or hijacks sessions, and post-exploitation makes money through scams, lateral movement, or selling access on black markets, providing a clear map for controls and evidence collection [18,32]. Within this cycle, different types of phishing fall into categories along a spectrum of targeting and technical complexity. Bulk email phishing relies on fake domains and generic verification messages [3,33], while spear-phishing and whaling target individuals or executives with detailed, personalised information [33]. Clone phishing alters legitimate messages by changing links or attachments [33], smishing and vishing use the same tricks on SMS and voice channels [32], and pharming redirects traffic through DNS manipulation [18]. The technical tools used in these categories include URL and domain deception, such as typosquatting and homographs [34], as well as DNS poisoning, SSL/TLS abuse, and fast-flux hosting that changes infrastructure to evade blocklists [2]. The social side uses fear, urgency, authority, impersonation, curiosity, and scarcity, with trust in social media plat-

forms and social proof making people more susceptible [31,35]. This mix of technical and social tools makes phishing a challenging forensic problem, as evidence can be found in human-readable content, protocol and transport traces, name-resolution events, certificate usage, and endpoint state, requiring coordinated collection from memory, disk, network, and cloud sources.

#### 3.1.3. Forensic Principles vs. Detection

Although digital forensics and security detection work together to reduce harm, they have different goals and need to be treated separately in rigorous programmes. Forensic practice focuses on admissibility through integrity, repeatability, and documentation, requiring a chain of custody and validated tools. It frames analysis as a process of reconstruction aimed at explaining and accounting for events in legal or regulatory settings [36]. Security detection prioritises timeliness through automation, model-driven classification, and preventative blocking, accepting trade-offs that may not meet evidential thresholds in raw form [33]. The two areas intersect productively when alerts, scores, and telemetry provide investigative leads for forensic acquisition and analysis. In turn, artefacts gathered through sound forensic methods, such as packet captures, DNS histories, email headers and bodies, endpoint logs, and memory artefacts feed back into refining filters and models. This strengthens recall without compromising precision [10]. The key takeaway is to design for convergence, using forensically sound collection to support detection credibility and using responsive detection to accelerate forensic triage. This ensures that rapid response is aligned with evidential sufficiency, and post-incident learning informs model features, user education, and policy hardening across the phishing lifecycle.

### *3.2. Automated Forensic Workflows*

Automation has become increasingly important in handling the growing scale and complexity of digital evidence produced by modern communication platforms. Automated pipelines can help with artefact collection, event reconstruction, and behavioural analysis. However, these methods must maintain evidential provenance and analytical transparency to ensure they remain legally admissible and scientifically reproducible.

Although this review mainly concentrates on phishing and email-related artefacts, related developments in social media forensics demonstrate how automation can be developed while maintaining evidential traceability. Arshad et al. [37] propose a multi-layer semantic automation model for online social network investigations that integrates automated evidence collection, structured data representation through ontology and RDF (Resource Description Framework), analytical operators, and investigator-oriented visualisations. Such architectures demonstrate that automated forensic workflows can support investigative scalability while maintaining traceability from derived analytical results back to original artefacts. This principle is equally critical in phishing investigations, where automated detection and analysis must remain anchored to verifiable evidential sources.

## 4. Findings

This section presents a thematic synthesis of the literature on phishing detection and digital forensics, and identifies recurring integration patterns and gaps that motivate a forensic-enabled detection framework.

### *4.1. Phishing Detection: State of the Art*

Phishing detection has advanced through a variety of methods, each grounded in distinct assumptions, yet collectively contributing to a more comprehensive defence. One of the first methods employs traditional techniques, such as blacklists, rule-based systems, and signature matching. These methods are conceptually straightforward and economically

efficient, rendering them suitable for broad deployment. However, their reactive nature renders them vulnerable to adaptive adversaries who exploit domain alterations, URL mutations, and fast-flux infrastructures. Empirical evaluations show that while static lists and handcrafted rules can block known threats, they consistently fall behind evolving campaigns, prompting a shift towards data-driven methods [18,33].

Initial comparative studies of machine learning techniques for phishing detection showed different levels of performance among classifiers like Support Vector Machines (SVMs), Naïve Bayes, and decision trees [38]. Building on these limitations, machine-learning methods have created feature-based classifiers that learn from email headers, body text, and URL attributes. Traditional algorithms, such as Support Vector Machines (SVMs), Naive Bayes, k-NN, and Random Forests, have set solid baselines for distinguishing between benign and malicious content [18,33]. Ensemble and lightweight machine-learning approaches have further improved phishing website and URL detection by combining classifiers or reducing detection overhead in fast-changing web environments [39,40]. Further refinements have expanded feature engineering to include structural and consistency checks, where mismatches between URL composition and web content have emerged as robust indicators of deception [34]. Despite their interpretability and practical auditability, these models still face ongoing challenges, particularly concept drift as adversaries adapt their tactics and the inability of handcrafted features to capture higher-order dependencies in text and markup [36]. Recent adaptive detection frameworks have attempted to address concept drift through incremental learning mechanisms. For example, Prasad and Chandra [41] introduced PhiUSIIL, which combines URL similarity analysis with incremental machine-learning models that continuously update as new phishing patterns emerge, reducing the need for complete retraining while maintaining detection effectiveness in dynamic environments.

Addressing these limitations, deep learning methods have gained prominence using hierarchical representation learning. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) identify semantic and sequential patterns in email and URL data, achieving better detection accuracy than traditional models [42]. Recent work further advances this direction by systematically evaluating both traditional machine learning and transformer-based deep learning models across multiple phishing datasets. Alhuzali et al. [11] demonstrate that models such as BERT and RoBERTa significantly outperform conventional classifiers, achieving accuracy levels approaching 99% in specific datasets. Their findings highlight the effectiveness of context-aware language models in capturing semantic patterns within phishing emails, reinforcing the shift toward deep learning-driven detection approaches. Recently, hybrid architectures, such as the ResNeXt-GRU model, have further improved performance by combining convolutional depth with sequence modelling, while addressing imbalanced datasets through the Synthetic Minority Over-sampling Technique (SMOTE), which reduces false negatives without increasing false positives [43]. However, challenges remain, as deep learning systems require significant resources, often lack the explainability needed for audit and forensic admissibility, and are limited by the scarcity of high-quality annotated datasets that reflect evolving threats [36,43]. Recent research has shown that models trained on legacy phishing datasets frequently experience substantial performance degradation when evaluated against modern phishing corpora, particularly those containing AI-generated content. Pajola et al. [7] addressed this issue through realistic phishing corpus generation and cross-dataset validation, highlighting the importance of dataset realism and evaluation diversity in phishing research.

Methods also aim to incorporate user-centred design approaches that recognise human behaviour as both a weakness and a strength in identifying threats. Measures such as awareness programmes, phishing simulations, and gamified training have been shown to

improve user vigilance towards deceptive cues, thereby reducing susceptibility on harmful content [35]. Nevertheless, social engineering attack strategies such as urgency, authority, and curiosity are still able to deceive even educated users [35]. The takeaway here is not that training does not work, but rather that it is most effective when used in concert with automated systems capable of leveraging behavioural data and contextual indicators to adjust detection thresholds dynamically.

Overall, these trends show both the progress and the ongoing challenges of detecting phishing. Traditional methods encounter difficulties in adapting, whereas machine learning is constrained by concept drift and the expression of superficial features. Deep learning raises concerns regarding interpretability and resource requirements, and user-focused approaches depend on continuous institutional support. Most importantly, detection pipelines often provide limited attribution and poor evidence handling, which hinders their usefulness in legal or forensic settings. These shortcomings highlight the need for tighter integration between detection and digital forensics, ensuring that automated classifications are both operationally effective and evidence-based [3,43].

### *4.2. Digital Forensics in Cybersecurity: Evidence-Centred Investigation*

Across the reviewed studies, phishing investigation is consistently approached as a multi-layer forensic problem spanning email, network, endpoint, cloud, and identity artefacts. However, the depth of artefact modelling and preservation varies significantly across these layers. Email artefacts are the most frequently analysed, while endpoint and cloud identity evidence are comparatively under-integrated in detection pipelines. This process uses machine learning and text mining techniques to improve the accuracy of triaging [44]. Network forensics examines packets and flows to reconstruct the delivery of attacks, for example, and maps out what has been stolen, including credentials, using tools like Wireshark and NetworkMiner for crucial visibility [21]. Endpoint forensics examines memory and disk artefacts, such as credential caches, browser histories, and injected binaries; it is particularly helpful when network traces cannot be analysed [27]. Finally, cloud forensics utilises provider logs, storage metadata, and API traces to support the reconstruction of attacker activities in service-based environments, despite challenges posed by distributed and provider-controlled infrastructures [45].

The literature indicates a growing migration of phishing activity to social media ecosystems, yet forensic tooling and enterprise visibility remain limited in these environments. Unlike corporate email systems, social platforms often restrict artefact access, creating evidential blind spots that complicate attribution and reconstruction. The threats in this area cover a broad range of account and identity-related issues, including impersonation attacks where profiles are used to exploit victims. As noted by practitioners, with social media phishing, the approach is different since it is not located within traditional corporate network perimeters and cannot be mitigated or detected with standard enterprise security tools. The ramifications of this trend are clear in high-profile instances, such as the compromise of numerous Microsoft Twitter accounts by nation-state actors in 2014, which led to the exposure of employee credentials [2]; the detection of over 3.7 million phishing attempts observed on social media during the first three months of 2018, with more than 60% originating from fake Facebook pages [2]; and the near doubling of “soshing” attacks across Facebook and Instagram by 2019, illustrating the rapid exploitation of these platforms for phishing [2]. Together, these cases demonstrate from a forensic perspective the need to expand investigations into social media ecosystems, where artefacts include posts, messages, account activity logs, and platform metadata, which are crucial for attribution, mitigation, and litigation.

Tool ecosystems range from open-source memory frameworks to commercial forensic suites and enterprise log aggregation platforms. However, the literature rarely demonstrates seamless integration of these tools into automated phishing detection workflows. Instead, most implementations position forensic tools downstream of detection systems, reinforcing the operational separation between alert generation and evidential reconstruction. Although structured processes such as the Network Forensics Development Life Cycle (NFDLC) introduce rigour and reproducibility into packet analysis and attribution [46], they are seldom embedded directly into real-time detection pipelines.

Although phishing is a primary consideration, the domain of digital forensics also covers other threats such as malware infection, ransomware dissemination, and insider abuse. Each of these leaves conflicting traces in logs, messages, and system states. Case studies show that an extensive forensic reconstruction can demonstrate targeted spear-phishing attacks against academic institutions and attribute responsibility, even with limited resources [36]. More generally, frameworks such as Fronesis illustrate that fusing forensic signals with operational pipelines can act as premonitions of incoming attacks, thereby closing the loop between observing, detecting, and responding [10].

However, there are still limitations. Forensic methods are inherently slower than frontline detection, focusing on thoroughness and traceability rather than immediacy, and often get involved only after damage has been done. Encrypted communication channels and opaque cloud infrastructures limit visibility, leading to a greater reliance on volatile memory or provider logs [45]. Within organisations, many institutions still separate detection and forensic functions, resulting in detection telemetry that lacks strong evidence and forensic findings that do not inform proactive detection. A strategic approach recommends forensic readiness: ensuring from the start that detection pipelines retain provenance, metadata, and contextual information needed for admissible analysis, while also integrating forensic findings into detection features, signatures, and user education [10,47].

### *4.3. Bridging the Gap: Detection–Forensics Convergence Patterns*

Phishing creates various artefacts across different communication layers, which can be systematically analysed using digital forensic techniques. Combining these forensic findings with automated detection systems enhances operational resilience and strengthens evidence collection. This section reviews how different forensic disciplines contribute to phishing investigation and detection.

#### 4.3.1. Email Artefact Forensics

Headers and metadata are usually the first concern during forensic analysis. By examining SMTP headers, spoofed fields, obscured routing history, and unusual timing patterns, indications of fraud can be exposed. Banday [48] demonstrates how forged From or Return-Path fields, open relays, anonymisers, and botnet traffic can be detected through structured header analysis, enabling the reconstruction of message journeys across networks. Banday [49] also discusses forensic artefacts at the architectural level, their evidential value, and tools such as Aid4Mail, FTK, and EnCase, which aid in the extraction and correlation of data. Almarzooqi et al. [14] extend this by contrasting artefacts from Gmail, Yahoo, and Livemail, focusing on SPF, DKIM, and DMARC validation, timestamp and IP inconsistencies, as well as attachment metadata. Some studies have further integrated header analysis with NLP-based content examination, combining structural and semantic artefacts into automated forensic tools capable of producing structured evidence reports. Although header analysis is widely used in phishing investigations, the literature indicates that no universally standardised methodology exists for header-first forensic processing. Altulaihan et al. [50] provide practical case studies and forensic procedures that expose

artefacts related to weaknesses in email security implementations. Rahma and Riadi [46] propose the use of the Network Forensics Development Life Cycle (NFDLC) for phishing email investigations and elaborate on a well-trained approach from evidence collection to analysis. In general, the forensic study of email artefacts can support machine-learning-based detection, extend threat intelligence, and generate forensically sound evidence suitable for legal proceedings.

#### 4.3.2. Endpoint & Network Forensics

Endpoint and network artefacts provide the strongest causal confirmation that a phishing lure progressed to compromise, as they capture execution, persistence, and post-click communication rather than intent signals alone. Memory forensics can reveal injected processes, credential dumping activity, or decrypted payload traces. The literature shows a shift from manual acquisition towards more automated, OS-aware frameworks that improve repeatability and coverage [51]. Complementing this, methods that visualise RAM dumps and apply learning-based classifiers can strengthen malware identification when traditional triage is insufficient [52], while ensemble approaches improve portability across platforms and environments [53]. On disk, artefacts such as registry changes, caches, and credential files support reconstruction of user actions and attacker persistence.

From a bridging perspective, network evidence (e.g., PCAPs, DNS, TLS, and C2 patterns) links the email or URL lure to downstream behaviour and enables campaign-level correlation. Packet and flow analysis supports delivery-chain reconstruction and can inform IDS/SIEM rule development [21]. Behavioural traffic frameworks such as *C2-Eye* further demonstrate how traffic analysis can be combined with behavioural recognition to detect command-and-control traffic using behavioural traffic analysis [54]. The core limitation is *timing and volatility*: many organisations do not retain short-window packet evidence or volatile endpoint state at the moment an alert fires, so investigations become inferential rather than evidential. This motivates detection-triggered acquisition (Section 4.3.6) and tiered retention policies that preserve just-enough telemetry for defensible reconstruction.

#### 4.3.3. User Behaviour Forensics

Behavioural artefacts provide the missing *ground truth* on whether a suspected lure led to harmful action. Clickstreams, browser traces, and login activity can validate compromise and reduce false positives by linking content-based suspicion to user interactions and authentication outcomes. Browser reconstruction methods support post-event explanations of how victims engaged with phishing pages [55]. Recent work moves beyond simple heuristics to behavioural modelling that captures temporal and relational patterns, including graph-based approaches for detecting unusual logins [22,56]. At enterprise scale, behavioural analytics and anomaly detection can detect subtle deviations consistent with credential abuse, particularly when combined with multi-factor authentication signals [57]. However, user-behaviour evidence is operationally sensitive. Bridging detection and forensics, therefore, requires governance-by-design: minimal collection, auditable access, and retention rules proportionate to incident severity. In practice, this design implies treating behavioural signals as *confirmatory* evidence to corroborate alerts and prioritise acquisition, rather than as indiscriminate surveillance.

#### 4.3.4. Attribution & Campaign Analysis

Attribution is rarely resolved by a single indicator; it becomes more defensible when multiple artefact classes are correlated across campaigns. Linguistic clustering can support authorship inference and grouping across phishing email corpora [58]. Infrastructure analysis adds a complementary axis: DNS and domain data can reveal reuse, shared hosting, and kit-based operations [59]. Surveys consistently emphasise that attribution

remains a major challenge and that practical progress comes from hybrid correlation strategies that connect content, infrastructure, and malware/tooling overlap [60]. For bridging, the main contribution of attribution work is not merely naming an actor but producing *actionable linkage*: shared infrastructure and artefact reuse can be operationalised into blocklists, sinkholing, takedown requests, and proactive hunting hypotheses, while maintaining an evidence trail suitable for later scrutiny.

#### 4.3.5. Chain of Custody & Legal Evidence

A repeated theme across the literature is that model outputs (scores, labels, explanations) carry limited legal value unless they are anchored to preserved artefacts with integrity, provenance, and reproducible procedures. Chain of custody provides this foundation by documenting who collected what, when, how, and under what controls. Validation frameworks such as RVEF aim to improve the reliability and reproducibility of digital forensic evidence [15]. Broader surveys outline how legal hurdles arise when data integrity is not demonstrably preserved across heterogeneous forensic subfields [47], and jurisprudence analyses highlight admissibility concerns when artefacts may be altered during acquisition or analysis [61]. Cryptographic mechanisms (hashing and secure timestamping) strengthen tamper resistance and evidential trust [16].

The bridging implication is straightforward: *detection becomes forensically meaningful only when evidence handling is embedded at the point of detection*. This shifts evaluation from "is the classifier accurate?" to "can the decision be reproduced and defended from preserved artefacts?"

#### 4.3.6. Auto-Collection of Artefacts: SOAR Playbooks and Forensic Readiness

The most operationally mature bridging pattern is forensic readiness triggered by detection signals. A SOAR playbook can be activated by user reporting, gateway quarantine, URL verdicts, identity anomalies, or EDR/IDS alerts, aligning with incident-handling guidance that emphasises timely and repeatable response actions [62,63]. The key principle is to treat the alert as a *collection trigger*.

Upon trigger, the playbook should acquire: (i) the raw email (RFC822/EML) and full headers [64]; (ii) authentication results (SPF/DKIM/DMARC) [65–67]; (iii) extracted indicators (URLs, domains, sender infrastructure, hashes); (iv) time-bounded network telemetry (DNS, proxy, firewall/NetFlow, and where feasible rolling PCAP); and (v) endpoint and identity context (EDR process trees, sign-in logs, MFA prompts). To preserve forensic soundness, automated acquisition should incorporate integrity and traceability controls (hashing at acquisition time, provenance metadata, and chain-of-custody logging) aligned with evidence acquisition guidance [12,13]. This converts "alert handling" into "incident record building," enabling both better investigations and systematic feedback into detection engineering.

#### 4.3.7. Cloud & Identity Forensics

As organisations adopt SaaS and cloud identity platforms, phishing increasingly targets authentication workflows rather than endpoints alone, making identity telemetry central to confirmation and forensic reconstruction. Cloud artefacts encompass sign-in logs, multi-factor authentication (MFA) events, OAuth token issuance and consent records, mailbox rule changes, and administrative API audit trails [23,45,68].

These artefacts provide verifiable evidence of whether a suspicious lure resulted in account takeover or session hijacking and enable end-to-end reconstruction of phishing incidents, from the initial lure to post-compromise actions [23,69].

The main constraint is limited retention (often weeks to a few months) and provider heterogeneity. Consequently, a bridging design should prioritise automated API-based

log collection aligned with detection triggers, ensuring that identity evidence is preserved before it expires and can be correlated with email, network, and endpoint artefacts [70].

#### 4.3.8. Privacy-Aware Forensic Retention Strategies

Integrated detection–forensics pipelines must balance evidential utility with privacy obligations because they may capture communications content, behavioural traces, and identity data. Recent studies demonstrate that linkage attacks and cross-platform data correlation can enable deanonymization even when datasets are partially anonymised, raising important concerns for privacy-aware forensic evidence management and cross-layer artefact correlation [71]. The literature supports privacy-aware retention through tiered models (retain metadata longer; retain content only for confirmed incidents), rolling capture windows for volatile artefacts, selective pseudonymisation, and cryptographic deletion after retention expiry [16,61,72]. These strategies reduce storage overhead and compliance risk while sustaining forensic readiness. Operational deployment also requires governance controls such as role-based access management, auditable access logs, retention and expiry enforcement, and selective pseudonymisation mechanisms.

From a bridging viewpoint, privacy is not a post hoc constraint but a feasibility condition: without minimisation and auditable governance, organisations will avoid collecting the very artefacts needed for defensible investigation. Therefore, operational designs should encode (i) severity-based retention, (ii) automated sensitive-data classification, and (iii) audit trails for access and handling, ensuring that evidence preservation strengthens rather than undermines legal compliance [61,73].

#### 4.3.9. Synthesis Across Layers

Across layers, the bridging requirement is consistent: evidence must be captured early, normalised, and preserved with provenance so that detection decisions can be reproduced and defended. Table 6 summarises artefact classes, techniques, and detection impacts, highlighting how a shared evidence model enables both operational response and admissible reconstruction.

Phishing detection methods have improved across various domains such as email, network, and behavioural analysis. However, a key limitation persists: most systems focus on classification accuracy and quick response, neglecting the systematic preservation of artefacts, original data provenance, and the integrity of evidence during detection. Consequently, alerts are mostly probabilistic, and vital evidence may be lost before forensic analysis. This disconnect between detection and forensic processes creates a significant gap. Figure 2 compares traditional detection systems with a forensic-enabled model, highlighting the importance of the proposed PhishTrace framework.

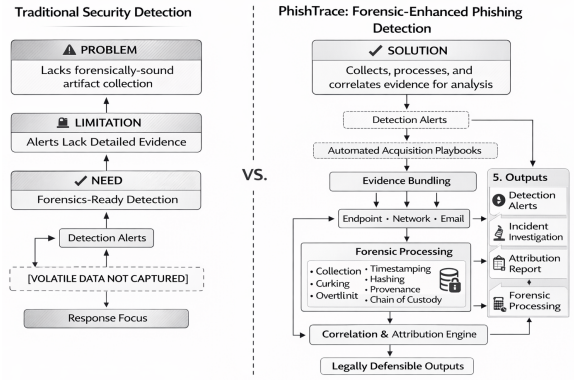


**Figure 2.** Conceptual comparison between traditional phishing detection architectures and the proposed PhishTrace forensic-ready reference architecture.

**Table 6.** Forensics-informed artefacts, evidence basis, and detection impact in phishing investigations.

| Forensic Layer | Artefacts | Techniques | Key References and Basis | Detection Impact |
| --- | --- | --- | --- | --- |
| Email Forensics | SMTP headers, SPF/DKIM/DMARC results, attachments, URLs, metadata | Header parsing, anomaly detection, WHOIS/domain lookup, macro/script extraction, forensic tools (Aid4Mail, EnCase, FTK) | Banday [48,49] [Case/Exp.]; Almarzooqi et al. [14] [Case]; Altulaihan et al. [50] [Surv./Case]; Rahma & Riadi [46] [Case] | Identifies spoofing and routing anomalies; enriches ML features and threat intelligence; produces artefacts suitable for evidential review. |
| Endpoint & Memory | RAM dumps, registry changes, caches, credential stores, binaries | Memory acquisition, manifold learning, stacked ensembles, CNNs, disk artefact analysis | Nyholm et al. [51] [Surv.]; Bozkir et al. [52] [Exp.]; Naeem et al. [53] [Exp.] | Reveals payloads; supports malware classification; provides forensic ground truth. |
| Network Forensics | PCAPs, DNS queries, TLS certificates, C2 traffic | Packet analysis, flow reconstruction, behavioural C2 frameworks | Sikos [21] [Surv.]; Haider et al. [54] [Exp.] | Detects post-phish command-and-control activity; supports IDS/SIEM rule development; may support attribution analysis. |
| User Behaviour | Clickstream, browsing logs, login metadata, device fingerprints | Browser activity reconstruction, anomaly detection, graph transformers, behavioural analytics | Oh et al. [55] [Exp.]; Gonçalves & Zanchettin [56] [Exp.]; Zhao et al. [22] [Oper.]; Ho et al. [57] [Oper.] | Provides compromise confirmation; detects compromised accounts; reduces false positives. |
| Attribution & Campaigns | Phishing kits, DNS/domain data, malware, linguistic artefacts | Authorship analysis, clustering, infrastructure correlation, campaign analysis | Seifollahi et al. [58] [Exp.]; Cova et al. [74] [Exp.]; Prasad et al. [60] [Surv.] | Links related campaigns and infrastructure; enables proactive blocking; supports threat intelligence. |
| Chain of Custody & Legal | Preserved artefacts, provenance, cryptographic hashes | Reliability frameworks (RVEF), jurisprudence, cryptographic timestamping | Stoykova & Franke [15] [Conc.]; Al-Dhaqm et al. [47] [Surv.]; Casey [61] [Conc.]; Ogunseyi & Adedayo [16] [Surv.] | Supports integrity preservation and evidential traceability; may strengthen trust in detection pipelines. |
| Cloud & Identity | IdP authentication logs, OAuth grants, mailbox rules, API audit trails, session tokens | API-based collection, behavioural analytics, risk scoring, token analysis | Alshabibi & Budookhi [45] [Surv.]; Khan et al. [23] [Surv.] | Validates actual compromise; enables attack reconstruction; reduces false positives; may support attribution analysis. |
| Privacy & Retention | Pseudonymised logs, hashed artefacts, retention metadata, audit trails, policy documents | Data minimisation, tiered retention, cryptographic deletion, selective collection | Abulaish et al. [72] [Conc.]; Casey [61] [Conc.]; Ogunseyi & Adedayo [16] [Conc.] | Supports regulatory compliance; reduces storage overhead; enables forensic readiness while limiting privacy risks. |

Evidence basis abbreviations: Exp. = Experimental; Case = Case study; Surv. = Survey; Conc. = Conceptual; Oper. = Operational.

### 4.4. Digital Forensics in the Age of Generative AI

Recent studies indicate that generative artificial intelligence (GenAI) can automatically create highly personalised phishing content that can bypass traditional detection mechanisms [4–6]. These studies converge on the observation that GenAI has fundamentally altered phishing scalability and realism, but they diverge in how forensic readiness should adapt, while detection research prioritises classifier performance against AI-generated lures, relatively few works address preservation of AI-specific artefacts such as prompt histories and model interaction traces. In addition to improving deception, AI-assisted phishing introduces new forensic artefacts such as prompt histories, model interaction traces, and synthetic linguistic signatures, which pose fresh challenges for detection, attribution, and legal admissibility. This section explores how digital forensic methods must adapt to GenAI-driven threats, emphasising evidence collection strategies, detection and prevention techniques, analysis of attack methods and human factors, and emerging open research issues.

#### 4.4.1. Forensic Methodologies for AI-Assisted Crimes

Forensic investigations increasingly need to consider artefacts generated through the use of AI tools rather than direct human actions. Interaction traces, application logs, memory artefacts, and network traffic can reveal how automated tools or generative models were used during an attack. Memory and network forensics techniques enable investigators to reconstruct execution environments, recover volatile artefacts, and correlate system activity with malicious behaviour [61,75]. These artefacts represent an emerging evidential layer for analysing AI-assisted cybercrime and require established forensic methodologies to adapt to increasingly automated attack environments [76]. Sousa-Silva [9] finds that traditional stylistic analysis may be insufficient for reliably attributing AI-generated content and instead recommends using deep linguistic features to distinguish between human and machine authors. Moreover, Alsubaei et al. [43] introduce a hybrid deep learning approach combining ResNeXt and GRU for real-time phishing detection. This method enhances classification accuracy and provides features that are useful for forensic investigations. Overall, these studies expand forensic techniques to encompass AI artefact detection and sophisticated linguistic analysis, helping ensure that findings are based on admissible forensic evidence.

#### 4.4.2. Detection & Prevention for AI-Generated Phishing

The sophistication of generative AI (GenAI) phishing requires defences that can identify malicious content while maintaining forensic integrity. Eze and Shamir [5] explore how generative AI allows attackers to create personalised phishing emails that bypass traditional detection methods. Context-aware large language models have also been investigated for phishing detection. Nicklas et al. [8] demonstrated that incorporating contextual information into open-source large language models improves semantic understanding of phishing intent and deception strategies, enabling more robust detection of sophisticated phishing content. Eze and Shamir also developed machine-learning techniques with high accuracy in distinguishing AI-generated phishing messages and stressed the need for specialised training datasets containing AI-generated malicious content for effective detection. AlEroud and Karabatis [77] demonstrate how generative adversarial networks can be used to craft phishing URLs that evade URL-based detection systems, highlighting the need for adversarially robust detection and forensic logging. Alsubaei et al. [43] combine ResNeXt and GRU architectures for real-time detection, incorporating forensic logging of key features and hashes into the process. Deshpande et al. [78] evaluate the cybersecurity risks posed by generative AI through the Cyber Kill Chain framework, highlighting how

GenAI aids multiple attack stages, including reconnaissance, lure creation, and delivery, by producing convincing phishing emails, deepfake media, and disinformation. They also suggest adaptive defence strategies to mitigate these AI-enhanced threats. Overall, these studies demonstrate that effective AI-phishing defences combine classifiers, behavioural models, and forensic logging, supporting alerts that are operationally useful and more evidentially transparent.

#### 4.4.3. Attack Mechanics & Human Factors

Generative artificial intelligence (GenAI) enhances phishing attacks by amplifying both the quality of persuasion and the specificity of targets. Schmitt and Flechais [4] describe three pillars of AI-enhanced social engineering: realistic content creation, advanced targeting and personalisation, and automated attack infrastructure, which together redefine attacker capabilities. Jabir, Le, and Nguyen [79] provide a systematic review of human factors in generative AI-enhanced phishing, demonstrating that while traditional awareness measures remain insufficient, GenAI magnifies exploitation of cognitive biases, trust cues, and attentional fatigue, thereby heightening victim susceptibility and underscoring the need for human-centred defence strategies. In parallel, AlEroud and Karabatis [77] illustrate adversarial use of generative adversarial networks (GANs) to produce phishing URLs that evade classifier detection, evidencing the evolving arms race between attackers and defenders. Together, these studies highlight that GenAI not only transforms phishing mechanics but also magnifies human vulnerability, demanding co-designed solutions that blend robust machine learning with human-centred defences.

#### 4.4.4. Synthesis: Evidence Handling & Open Problems

The convergence of generative artificial intelligence (GenAI) and phishing introduces pressing requirements for forensic science. Evidence models must expand to capture AI-specific artefacts such as prompt histories, model identifiers, and detector explanation logs. Detection pipelines require curated, legally shareable datasets of AI-generated malicious content for both training and forensic validation [79]. However, some AI-related artefacts, including prompt histories and model interaction traces, may not always be accessible in attacker-controlled or third-party generative AI environments, limiting their practical availability during real-world forensic investigations. Evaluation processes must incorporate adversarial test suites, including GAN-crafted URLs and paraphrased lures, to ensure robustness under real-world attack conditions. Security operations centres (SOCs) should integrate human-factor metrics, recognising the role of fatigue and bias in incident response [79]. Aligning forensic controls with the Cyber Kill Chain clarifies intervention points across reconnaissance, delivery, and exploitation [78]. At the same time, chain-of-custody protocols must evolve to embed cryptographic hashing and automated provenance tracking into AI-driven detection workflows, ensuring evidentiary integrity in legal contexts. Moreover, future research should address the tension between forensic preservation and privacy obligations, particularly where digital twin or behavioural baselines require processing sensitive communication data. These gaps underscore the need for a unified framework that couples forensic rigour with adaptive AI defences, bridging the divide between automated detection and legally admissible evidence generation.

## 5. Proposed Framework: PhishTrace

The proposed PhishTrace framework is presented as a conceptual forensic-ready reference architecture derived from thematic synthesis of the reviewed literature rather than as an empirically validated operational system or deployed detection platform.

PhishTrace is a conceptual forensic-ready reference architecture for phishing detection and investigation that embeds evidence readiness into standard detection processes. Conventional systems focus on classification accuracy and quick mitigation, issuing alerts based on content analysis, URL reputation, or behavioural anomalies. Although effective operationally, these systems often lack structured artefact preservation, provenance tracking, and controls for evidential integrity. Consequently, alerts are probabilistic and lack reproducibility, which can limit their evidential robustness.

PhishTrace is designed to address this limitation by redefining alerts as triggers for automated, multi-layer evidence collection. When detection tools such as email gateways, URL analysis engines, or endpoint detection systems generate alerts, they start a systematic collection of artefacts across communication and execution layers. This process includes raw email messages and authentication results, network telemetry such as DNS queries and proxy logs, endpoint artefacts such as process trees and memory traces, and cloud or identity records, including authentication events and API audit trails. This multi-layer approach is intended to facilitate causal reconstruction instead of relying solely on isolated indicators.

To ensure the reliability of evidence, PhishTrace incorporates forensic processing during acquisition. Artefacts are protected using cryptographic hashing, trusted timestamps, provenance metadata, and chain-of-custody records. These methods improve integrity, traceability, and reproducibility, aligning detection procedures with recognised digital forensic standards.

Collected artefacts are normalised and analysed through a cross-layer correlation and attribution engine. By combining evidence from email, network, endpoint, and identity sources, the system verifies compromises, maps attack progression, and detects overlaps in infrastructure or campaigns.

Unlike traditional systems that primarily generate alerts, PhishTrace conceptually supports the generation of detailed investigative artefacts such as reproducible incident records and evidence bundles that are suitable for regulatory or legal review. It includes a feedback loop that continuously enhances detection models and threat intelligence, building an integrated detection and forensics framework that converts probabilistic alerts into clear, defensible investigative results (see Figure 3).

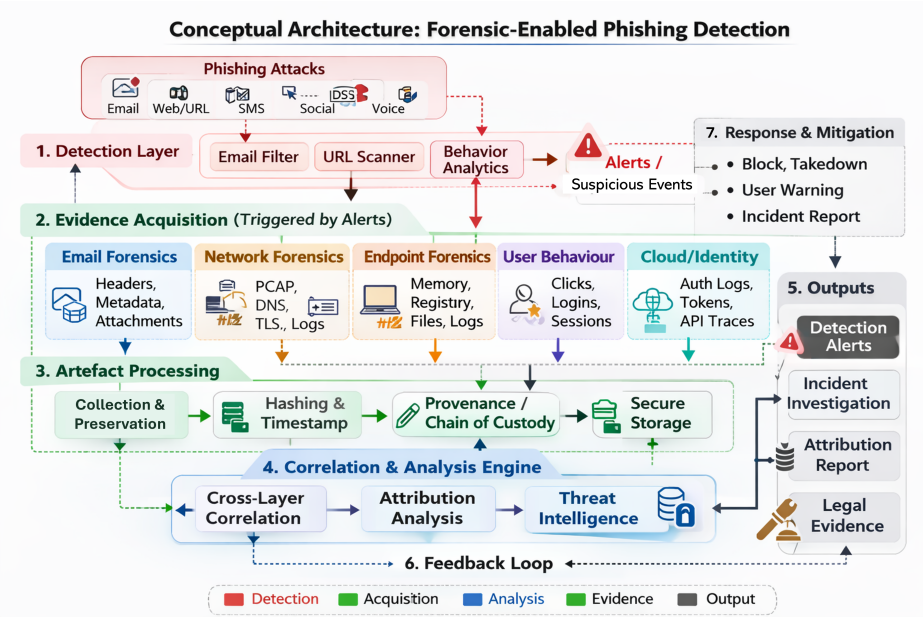


**Figure 3.** Conceptual forensic-ready reference architecture (PhishTrace) illustrating the integration of phishing detection, evidence preservation, provenance management, and cross-layer forensic analysis.

*Illustrative Phishing Investigation Workflow Using PhishTrace*

To illustrate PhishTrace's operational role as a conceptual forensic-ready reference architecture, consider a phishing incident involving a spoofed banking email sent to organisational users. The phishing email first enters the raw evidence acquisition layer, where the original `.eml` file is preserved alongside SHA-256 hashes, timestamps, and evidence identifiers. The artefact extraction layer then parses forensic-relevant metadata, including SMTP headers, SPF/DKIM/DMARC authentication results, Message-ID values, embedded URLs, sender domains, and attachment indicators. These artefacts are recorded in the provenance and chain-of-custody components to support evidential integrity and reproducibility. Subsequently, the correlation layer links extracted indicators to organisational telemetry sources such as SIEM alerts, DNS logs, authentication records, and URL reputation feeds. The resulting evidence graph enables investigators to reconstruct phishing delivery paths, identify related campaigns, and support attribution analysis. This walkthrough demonstrates how PhishTrace conceptually integrates detection-oriented telemetry with forensic evidence preservation, rather than functioning solely as a conventional phishing classifier.

Table 7 illustrates how the conceptual components of PhishTrace relate to capabilities commonly available in Security Information and Event Management (SIEM) and Security Orchestration, Automation, and Response (SOAR) platforms. The comparison is intended to demonstrate functional alignment rather than direct equivalence, while SIEM and SOAR technologies support monitoring, correlation, orchestration, and incident response activities, PhishTrace extends these capabilities by emphasising forensic readiness, evidence preservation, provenance tracking, integrity assurance, and investigative reconstruction throughout the phishing incident lifecycle.

**Table 7.** Illustrative alignment between PhishTrace components and capabilities commonly available in SIEM/SOAR platforms.

| PhishTrace Component | Related SIEM/SOAR Capability | Forensic Purpose |
|---|---|---|
| Raw Evidence Preservation | Log collection, archival storage, and event retention | Retains original phishing artefacts and associated metadata to support subsequent forensic examination and evidential review. |
| Integrity Preservation Layer | Secure logging and integrity verification mechanisms | Supports evidence authenticity and tamper detection through hashing, timestamping, and integrity controls. |
| Provenance and Chain-of-Custody Management | Audit logging and activity tracking | Provides traceability of evidence acquisition, handling, and analytical processes throughout the investigation lifecycle. |
| Cross-Layer Artefact Correlation | Event correlation and threat intelligence integration | Links email, network, endpoint, and identity artefacts to facilitate attack reconstruction and attribution. |
| Automated Evidence Acquisition | SOAR playbooks and orchestration workflows | Enables structured and repeatable collection of forensic artefacts following phishing alerts or suspicious activities. |
| Incident Evidence Bundle Generation | Case management and incident-tracking systems | Aggregates relevant artefacts and contextual information into structured records that support investigation, reporting, and review. |
| Feedback and Knowledge Integration | Detection rule refinement and threat intelligence updates | Uses investigative findings to enhance future phishing detection, response, and organisational learning. |

This illustrative scenario is intended to demonstrate how the conceptual components of PhishTrace may operate within a phishing investigation workflow. The example is not presented as empirical validation but as an operational illustration derived from forensic practices and evidence-handling principles identified throughout the reviewed literature.

## 6. Discussion

This review has explored the intersection of digital forensics and phishing detection, examining how forensic frameworks and methodologies can enhance detection, investigation, and attribution across technological and human dimensions. Several key conclusions emerge from our synthesis of literature spanning 2004 to 2026:

First, digital forensic artefacts from email headers and network traces to memory dumps and behavioural logs contain rich, underutilised signals that have the potential to improve phishing detection accuracy and investigative transparency while simultaneously supporting post-incident investigation. Our analysis reveals that current detection systems often treat these artefacts as mere classification features, neglecting their potential for evidential preservation, validation, and legal presentation.

Second, there is a clear organisational gap between detection and forensic activities. Institutions often see digital forensics as a specialised, reactive task, while detection systems prioritise speed over evidence robustness. This leads to telemetry that may be inadequate for legal processes and forensic results that do not consistently enhance detection methods. The growing use of anti-forensic techniques further complicates evidence preservation and attribution in phishing investigations [30]. The systematic evidential coding framework in this analysis reveals notable variability across studies regarding artefact specificity, provenance modelling, integrity preservation mechanisms, and reproducibility, while recurring patterns were identified across the reviewed literature, methodological heterogeneity and overlapping evidential categories limited the suitability of formal quantitative aggregation. Consequently, the findings are presented through thematic synthesis and cross-layer evidential analysis.

Future empirical evaluations of forensic-ready phishing defence frameworks should incorporate quantitative operational metrics beyond conventional detection accuracy. Relevant measures may include evidence completeness, provenance consistency, artefact recovery rates, time-to-acquisition, cross-layer correlation accuracy, forensic reconstruction success, and reproducibility of investigative outcomes. Such metrics would enable systematic assessment of how effectively forensic-ready architectures preserve evidential integrity while maintaining operational detection performance. Although many studies use phishing-related artefacts for classification, few explicitly incorporate integrity controls, secure timestamping, or chain-of-custody protocols in their detection procedures. This methodological gap exacerbates the institutional divide, making it essential to adopt a forensic readiness approach that designs detection workflows from the outset to preserve provenance, metadata, and contextual integrity.

Third, the rise of Generative AI presents both challenges and opportunities, while GenAI enables more sophisticated, personalised phishing attacks that evade traditional detection, it also creates new forensic artefacts (prompt histories, model identifiers, detector logs) and enables new detection approaches (AI-generated content identification, linguistic analysis). Forensic methodologies must evolve to capture these AI-specific traces while maintaining evidential integrity.

Fourth, successful integration requires attention to both technical and human dimensions. Technical solutions must address encryption, cloud infrastructure, and adversarial evasion, while human-centred approaches must consider cognitive biases, training effectiveness, and organisational silos. The PhishTrace framework, which views digital forensics

as an ongoing, proactive capability rather than just a reactive process, provides a conceptual model for this integration.

This review's limitations include its dependence on published literature, which may not fully reflect current operational practices, and the rapidly changing nature of phishing techniques and forensic tools. Future research should focus on filling several key gaps: developing forensic-ready detection pipelines with built-in evidence preservation, creating realistic and shareable AI-generated phishing datasets, balancing forensic preservation with privacy requirements, and establishing standardised validation frameworks for forensic tools in phishing scenarios.

Unlike prior phishing surveys that primarily focus on detection accuracy or attack taxonomy, this review emphasises forensic readiness, evidential preservation, provenance modelling, and cross-layer investigative reconstruction. These directions promote a shift toward proactive forensic readiness integrated into detection systems. The merging of digital forensics and phishing defence represents more than just technical integration. It indicates a maturation of cybersecurity practices. In this framework, detection, investigation, and attribution form a continuous cycle. This cycle is supported by evidence that can withstand both technical and legal scrutiny. As phishing advances, especially through AI-enhanced methods, a forensic-aware approach will be crucial for creating resilient and accountable security ecosystems. Future work should evaluate the proposed PhishTrace framework through prototype implementations, controlled phishing case studies, operational SIEM/SOAR integration, and cross-layer forensic reconstruction experiments using real-world phishing datasets. One challenge is that many phishing datasets provide only cleaned text or extracted features, not the original raw email files, while preprocessing can improve machine learning efficiency, it often removes important forensic artefacts, including SMTP headers, routing metadata, authentication results, timestamps, MIME structures, and attachments. This limits evidential reconstruction, provenance analysis, and reproducibility in phishing investigations. Future forensic-ready datasets should therefore preserve raw email artefacts and maintain integrity-preserving evidence collection procedures.

It is important to note that the PhishTrace framework is derived from thematic synthesis of the reviewed literature and is presented as a conceptual forensic-ready reference architecture. The framework has not yet been empirically validated through operational deployment, prototype implementation, or controlled experimental evaluation. Consequently, the observations presented in this section should be interpreted as implications of the literature synthesis and directions for future research rather than validated performance claims.

Future phishing defence systems should integrate realistic cross-dataset evaluation, context-aware semantic analysis, and adaptive learning mechanisms. Recent studies have demonstrated that static benchmark datasets may overestimate detection performance, while incremental learning approaches provide practical mechanisms for adapting to evolving phishing tactics without requiring complete model retraining [7,8,41]. These developments further support the need for forensic-ready phishing defence architectures capable of maintaining both operational effectiveness and evidential accountability under rapidly evolving threat conditions.

## 7. Conclusions

This review examines the role of digital forensics in strengthening phishing defences, synthesising research from 2004 to 2026 to assess how evidential techniques improve detection, investigation, and attribution. It reveals that despite progress in machine learning and AI detection, many systems remain primarily alert-focused and only partially forensic-ready. These detection tools often prioritise classification accuracy and quick response times, while elements like evidential preservation, provenance modelling, and chain-of-

custody are inconsistently applied. This inconsistency may hinder data recovery, reduce attribution confidence, and complicate evidential review in complex scenarios.

To address these challenges, the study introduces a forensic-layer taxonomy and a severity-based gap framework, highlighting common shortcomings in provenance preservation, cross-layer artefact correlation, cloud and identity artefact modelling, automated evidence collection, privacy-aware retention, and GenAI-related artefact capture. Results indicate that forensic readiness should be integrated into phishing defence strategies from the outset, rather than as an afterthought.

The paper also presents PhishTrace as a conceptual forensic-ready reference architecture. By embedding provenance controls, cryptographic integrity mechanisms, and cross-domain correlation into detection processes, PhishTrace conceptually illustrates how detection and forensic analysis may be linked throughout a unified evidential lifecycle.

As phishing threats increasingly utilise generative AI and cloud identity systems, improving evidential robustness alongside detection will become crucial. Continual empirical validation of forensic-ready detection models in real-world environments represents a key future research direction.

**Supplementary Materials:** The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/jcp6040116/s1: Table S1: Complete evidential coding matrix of the included studies.

**Author Contributions:** Conceptualisation, I.O.L.; methodology, I.O.L.; formal analysis, I.O.L.; investigation, I.O.L.; writing—original draft preparation, I.O.L.; writing—review and editing, I.O.L.; visualisation, I.O.L.; supervision, J.L. and C.N. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by LPDP (Indonesia Endowment Fund for Education), grant number not applicable.

**Data Availability Statement:** The evidential coding matrix generated during this review is provided as Supplementary Table S1. No additional empirical datasets were created or analysed.

**Acknowledgments:** The author acknowledges the support of LPDP (Indonesia Endowment Fund for Education).

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Khonji, M.; Iraqi, Y.; Jones, A. Phishing detection: A literature survey. *IEEE Commun. Surv. Tutor.* **2013**, *15*, 2091–2121. [CrossRef]
2. Anti-Phishing Working Group (APWG). *Phishing Activity Trends Report, 1st Quarter 2025: Activity January–March 2025*; Anti-Phishing Working Group (APWG): Cambridge, MA, USA, 2025. Available online: https://docs.apwg.org/reports/apwg_trends_report_q1_2025.pdf (accessed on 20 February 2026).
3. Jakobsson, M.; Myers, S. *Phishing and Countermeasures: Understanding the Increasing Problem of Electronic Identity Theft*; Wiley-Interscience: Hoboken, NJ, USA, 2007. [CrossRef]
4. Schmitt, M.; Flechais, I. Digital deception: Generative artificial intelligence in social engineering and phishing. *Artif. Intell. Rev.* **2024**, *57*, 324. [CrossRef]
5. Eze, C.S.; Shamir, L. Analysis and Prevention of AI-Based Phishing Email Attacks. *Electronics* **2024**, *13*, 1839. [CrossRef]
6. Sajid, S.; Fazil, A.W.; Hakimi, M. AI-based phishing attacks on university networks: A systematic literature review and defense framework. *J. Adv. Comput. Knowl. Algorithms* **2026**, *3*, 62–75. [CrossRef]
7. Pajola, L.; Caripoti, E.; Banzer, S.; Pizzi, S.; Conti, M.; Apruzzese, G. E-PhishGEN: Unlocking Novel Research in Phishing Email Detection. In Proceedings of the 18th ACM Workshop on Artificial Intelligence and Security (AISec), Taipei, Taiwan, 13–17 October 2025; pp. 64–76. [CrossRef]
8. Nicklas, F.; Ventulett, N.; Conrad, J. Enhancing Phishing Email Detection with Context-Augmented Open Large Language Models. In Proceedings of the Upper-Rhine Artificial Intelligence Symposium, Offenburg, Germany, 13–14 November 2024; pp. 159–166. [CrossRef]

9. Sousa-Silva, R. Fighting cyber-malice: A forensic linguistics approach to detecting AI-generated malicious texts. In Proceedings of the First International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security, Lancaster, UK, 29–30 July 2024; pp. 164–174. Available online: https://aclanthology.org/2024.nlpaics-1.19.pdf (accessed on 10 February 2026).
10. Dimitriadis, A.; Lontzetidis, E.; Kulvatunyou, B.; Ivezic, N.; Gritzalis, D.; Mavridis, I. Fronesis: Digital Forensics-Based Early Detection of Ongoing Cyber-Attacks. *IEEE Access* **2023**, *11*, 728–743. [CrossRef]
11. Alhuzali, A.; Alloqmani, A.; Aljabri, M.; Alharbi, F. In-Depth Analysis of Phishing Email Detection: Evaluating the Performance of Machine Learning and Deep Learning Models Across Multiple Datasets. *Appl. Sci.* **2025**, *15*, 3396. [CrossRef]
12. *ISO/IEC 27037:2012*; Information Technology—Security Techniques—Guidelines for Identification, Collection, Acquisition and Preservation of Digital Evidence. International Organization for Standardization/International Electrotechnical Commission (ISO/IEC): Geneva, Switzerland, 2012. Available online: https://www.iso.org/standard/44381.html (accessed on 9 February 2026).
13. *NIST SP 800-86*; Guide to Integrating Forensic Techniques into Incident Response. National Institute of Standards and Technology: Gaithersburg, MD, USA, 2006. [CrossRef]
14. Almarzooqi, A.K.A.K.; Alyammahi, E.A.O.S.; Almarzooqi, A.K.A.N.; AlAlawi, M.H.K.N.; Ikuesan, R.A. A Comparative Analysis of Email Artifacts from Gmail, Yahoo Mail, and Live Mail for Email Forensics. In *Information Systems and Technological Advances for Sustainable Development*; Springer: Cham, Switzerland, 2024; pp. 383–398. [CrossRef]
15. Stoykova, R.; Franke, K. Reliability validation enabling framework (RVEF) for digital forensics in criminal investigations. *Forensic Sci. Int. Digit. Investig.* **2023**, *45*, 301554. [CrossRef]
16. Ogunseyi, T.B.; Adedayo, O.M. Cryptographic Techniques for Data Privacy in Digital Forensics. *IEEE Access* **2023**, *11*, 142392–142410. [CrossRef]
17. Casino, F.; Dasaklis, T.K.; Spathoulas, G.P.; Anagnostopoulos, M.; Ghosal, A.; Borocz, I.; Solanas, A.; Conti, M.; Patsakis, C. Research Trends, Challenges, and Emerging Topics in Digital Forensics: A Review of Reviews. *IEEE Access* **2022**, *10*, 25464–25493. [CrossRef]
18. Alkhalil, Z.; Hewage, C.; Nawaf, L.; Khan, I. Phishing attacks: A recent comprehensive study and a new anatomy. *Front. Comput. Sci.* **2021**, *3*, 563060. [CrossRef]
19. Wohlin, C.; Mendes, E.; Felizardo, K.R.; Kalinowski, M. Guidelines for the search strategy to update systematic literature reviews in software engineering. *Inf. Softw. Technol.* **2020**, *127*, 106366. [CrossRef]
20. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [CrossRef] [PubMed]
21. Sikos, L.F. Packet analysis for network forensics: A comprehensive survey. *Forensic Sci. Int. Digit. Investig.* **2020**, *32*, 200892. [CrossRef]
22. Zhao, J.; Yang, C.; Wu, D.; Cao, Y.; Liu, Y.; Cui, X.; Liu, Q. Detecting Compromised Email Accounts via Login Behavior Characterization. *Cybersecurity* **2023**, *6*, 36. [CrossRef]
23. Khan, S.; Gani, A.; Wahab, A.W.A.; Bagiwa, M.A.; Shiraz, M.; Khan, S.U.; Buyya, R.; Zomaya, A.Y. Cloud log forensics: Foundations, state of the art, and future directions. *ACM Comput. Surv.* **2016**, *49*, 1–42. [CrossRef]
24. INTERPOL. *INTERPOL Global Guidelines for Digital Forensics Laboratories*; INTERPOL Global Complex for Innovation: Singapore, 2019. Available online: https://www.interpol.int/content/download/13501/file/INTERPOL_DFL_GlobalGuidelinesDigitalForensics (accessed on 11 February 2026).
25. US-CERT. *Computer Forensics*; United States Computer Emergency Readiness Team (US-CERT): Washington, DC, USA, 2008. Available online: https://www.cisa.gov/ (accessed on 9 February 2026).
26. House of Lords. *Forensic Science and the Justice System*; Authority of the House of Lords: London, UK, 2019. Available online: https://publications.parliament.uk/pa/ld201719/ldselect/ldsctech/333/333.pdf (accessed on 9 February 2026).
27. Atlam, H.F.; El-Din Hemdan, E.; Alenezi, A.; Alassafi, M.O.; Wills, G.B. Internet of things forensics: A review. *Internet Things* **2020**, *11*, 100220. [CrossRef]
28. Karie, N.M.; Kebande, V.R.; Venter, H.S. Digital forensics and judicial decision-making. *Forensic Sci. Int. Digit. Investig.* **2019**, *30*, 200–212.
29. Charanya, R.; Aramudhan, M. Survey on access control issues in cloud computing. In Proceedings of the 2016 International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS), Pudukkottai, India, 24–26 February 2016; pp. 1–4. [CrossRef]
30. Surakanti, S.; Goundar, S.; Dwight, J. Countering anti-forensic tactics in cybercrime investigations—A systematic literature review. *Int. J. Inf. Secur.* **2025**, *24*, 210. [CrossRef]
31. Lastdrager, E.E. Achieving a consensual definition of phishing based on a systematic review of the literature. *Crime Sci.* **2014**, *3*, 9. [CrossRef]

32. Chiew, K.L.; Yong, K.S.C.; Tan, C.L. A survey of phishing attacks: Their types, vectors and technical approaches. *Expert Syst. Appl.* **2018**, *106*, 1–20. [CrossRef]
33. Almomani, A.; Gupta, B.B.; Atawneh, S.; Meulenberg, A.; Almomani, E. A survey of phishing email filtering techniques. *IEEE Commun. Surv. Tutor.* **2013**, *15*, 2070–2090. [CrossRef]
34. Azeez, N.A.; Salaudeen, B.B.; Misra, S.; Damaševičius, R.; Maskeliūnas, R. Identifying phishing attacks in communication networks using URL consistency features. *Int. J. Electron. Secur. Digit. Forensics* **2020**, *12*, 200–213. [CrossRef]
35. Butavicius, M.; Parsons, K.; Pattinson, M.; McCormac, A. Breaching the Human Firewall: Social engineering in Phishing and Spear-Phishing Emails. *arXiv* **2016**, arXiv:1606.00887. [CrossRef]
36. Al-Ghamdi, N.; Alsubait, T. Digital forensics and machine learning to fraudulent email prediction. In Proceedings of the 2022 Fifth National Conference of Saudi Computers Colleges (NCCC), Makkah, Saudi Arabia, 17–18 December 2022; pp. 99–106. [CrossRef]
37. Arshad, H.; Abdullah, S.; Alawida, M.; Alabdulatif, A.; Abiodun, O.I.; Riaz, O. A Multi-Layer Semantic Approach for Digital Forensics Automation for Online Social Networks. *Sensors* **2022**, *22*, 1115. [CrossRef] [PubMed]
38. Abu-Nimeh, S.; Nappa, D.; Wang, X.; Nair, S. A comparison of machine learning techniques for phishing detection. In Proceedings of the Anti-Phishing Working Groups 2nd Annual eCrime Researchers Summit, Pittsburgh, PA, USA, 4–5 October 2007; pp. 60–69. [CrossRef]
39. Basit, A.; Zafar, M.; Javed, A.R.; Jalil, Z. A Novel Ensemble Machine Learning Method to Detect Phishing Attack. In *Proceedings of the 2020 IEEE 23rd International Multitopic Conference (INMIC), Bahawalpur, Pakistan, 5–7 November 2020*; IEEE: Piscataway, NJ, USA, 2020; pp. 1–5. [CrossRef]
40. Gu, C. A Lightweight Phishing Website Detection Algorithm by Machine Learning. In *Proceedings of the 2021 International Conference on Signal Processing and Machine Learning (CONF-SPML), Stanford, CA, USA, 14 November 2021*; IEEE: Piscataway, NJ, USA, 2021; pp. 245–249. [CrossRef]
41. Prasad, A.; Chandra, S. PhiUSIIL: A diverse security profile empowered phishing URL detection framework based on similarity index and incremental learning. *Comput. Secur.* **2024**, *136*, 103545. [CrossRef]
42. Bouijij, H.; Berqia, A.; Saliah-Hassan, H. Phishing URL classification using extra-tree and DNN. In Proceedings of the 2022 10th International Symposium on Digital Forensics and Security (ISDFS), Istanbul, Turkey, 6–7 June 2022; pp. 1–6. [CrossRef]
43. Alsubaei, F.S.; Almazroi, A.A.; Ayub, N. Enhancing Phishing Detection: A Novel Hybrid Deep Learning Framework for Cybercrime Forensics. *IEEE Access* **2024**, *12*, 8373–8389. [CrossRef]
44. Morovati, K.; Kadam, S.S. Detection of phishing emails with email forensic analysis and machine learning techniques. *Int. J. Cyber-Secur. Digit. Forensics* **2019**, *8*, 98–108. [CrossRef]
45. Alshabibi, M.M.; Budookhi, A.K.; Hafizur Rahman, M.M. Forensic investigation, challenges, and issues of cloud data: A systematic literature review. *Computers* **2024**, *13*, 213. [CrossRef]
46. Zakiyaturrahma; Riadi, I. Email forensic from phishing attack using network forensics development life cycle method. *Int. J. Comput. Appl.* **2022**, *183*, 36–42. [CrossRef]
47. Al-Dhaqm, A.; Razak, S.; Othman, S.H.; Choo, K.-K.R.; Glisson, W.B.; Ali, A. CDBFIP: Common Database Forensic Investigation Processes for Internet of Things. *IEEE Access* **2017**, *5*, 24401–24416. [CrossRef]
48. Banday, M.T. Technology Corner: Analysing E-Mail Headers for Forensic Investigation. *J. Digit. Forensics Secur. Law* **2011**, *6*, 5. [CrossRef]
49. Banday, M.T. Techniques and Tools for Forensic Investigation of E-mail. *Int. J. Netw. Secur. Its Appl. (IJNSA)* **2011**, *3*, 227–241. [CrossRef]
50. Altulaihan, E.; Alismail, A.; Rahman, M.M.H.; Ibrahim, A.A. Email security issues, tools, and techniques used in investigation. *Sustainability* **2023**, *15*, 10612. [CrossRef]
51. Nyholm, H.; Monteith, K.; Lyles, S.; Gallegos, M.; DeSantis, M.; Donaldson, J.; Taylor, C. The Evolution of Volatile Memory Forensics. *J. Cybersecur. Priv.* **2022**, *2*, 556–572. [CrossRef]
52. Bozkir, A.S.; Tahillioglu, E.; Aydos, M.; Kara, I. Catch them alive: A malware detection approach through memory forensics, manifold learning and computer vision. *Comput. Secur.* **2021**, *103*, 102166. [CrossRef]
53. Naeem, H.; Dong, S.; Falana, O.J.; Ullah, F. Development of a deep stacked ensemble with process-based volatile memory forensics for platform-independent malware detection and classification. *Expert Syst. Appl.* **2023**, *223*, 119952. [CrossRef]
54. Haider, R.Z.; Aslam, B.; Abbas, H.; Iqbal, Z. C2-Eye: Framework for Detecting Command and Control (C2) Connection of Supply Chain Attacks. *Int. J. Inf. Secur.* **2024**, *23*, 2531–2545. [CrossRef]
55. Oh, J.; Lee, S.; Hwang, H. Forensic detection of timestamp manipulation for digital forensic investigation. *IEEE Access* **2024**, *12*, 72544–72565. [CrossRef]
56. Gonçalves, L.; Zanchettin, C. Detecting abnormal logins by discovering anomalous links via graph transformers. *Comput. Secur.* **2024**, *144*, 103944. [CrossRef]

57. 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 *Proceedings of the 28th USENIX Security Symposium (USENIX Security 19)*; USENIX Association: Santa Clara, CA, USA, 2019; pp. 1273–1290. Available online: https://www.usenix.org/conference/usenixsecurity19/presentation/ho (accessed on 9 January 2026).
58. Seifollahi, S.; Bagirov, A.; Layton, R.; Gondal, I. Optimization based clustering algorithms for authorship analysis of phishing emails. *Neural Process. Lett.* **2017**, *46*, 411–425. [CrossRef]
59. Zhang, Y.; Hong, J.; Cranor, L.F. Cantina: A content-based approach to detecting phishing websites. In Proceedings of the 16th International World Wide Web Conference (WWW), Banff, AB, Canada, 8–12 May 2007; pp. 639–648. [CrossRef]
60. Prasad, N.; Diro, A.; Warren, M.; Fernando, M. A survey of cyber threat attribution: Challenges, techniques, and future directions. *Comput. Secur.* **2025**, *157*, 104606. [CrossRef]
61. Casey, E. *Digital Evidence and Computer Crime: Forensic Science, Computers, and the Internet*, 3rd ed.; Academic Press: Cambridge, MA, USA, 2011; ISBN 978-0-12-374268-1.
62. Cichonski, P.; Millar, T.; Grance, T.; Scarfone, K. *Computer Security Incident Handling Guide*; NIST Special Publication 800-61 Revision 2; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2012. [CrossRef]
63. *ISO/IEC 27035-1:2016*; Information Technology–Security Techniques–Information Security Incident Management–Part 1: Principles of Incident Management. International Organization for Standardization/International Electrotechnical Commission (ISO/IEC): Geneva, Switzerland, 2016. Available online: https://www.iso.org/standard/60803.html (accessed on 9 February 2026).
64. *RFC 5322*; Internet Message Format. RFC Editor: Wilmington, DE, USA, 2008. [CrossRef]
65. *RFC 7208*; Sender Policy Framework (SPF) for Authorizing Use of Domains in Email, Version 1. RFC Editor: Wilmington, DE, USA, 2014. [CrossRef]
66. *RFC 6376*; DomainKeys Identified Mail (DKIM) Signatures. RFC Editor: Wilmington, DE, USA, 2011. [CrossRef]
67. *RFC 7489*; Domain-Based Message Authentication, Reporting, and Conformance (DMARC). RFC Editor: Wilmington, DE, USA, 2015. [CrossRef]
68. Choo, K.-K.R.; Herman, M.; Iorga, M.; Martini, B. Cloud forensics: State-of-the-art and future directions. *Digit. Investig.* **2016**, *18*, 77–78. [CrossRef]
69. Sheng, S.; Holbrook, M.; Kumaraguru, P.; Cranor, L.F.; Downs, J. Who Falls for Phish? A Demographic Analysis of Phishing Susceptibility and Effectiveness of Interventions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*; ACM: New York, NY, USA, 2010; pp. 373–382. [CrossRef]
70. Martini, B.; Choo, K.-K.R. Cloud storage forensics: OwnCloud as a case study. *Digit. Investig.* **2014**, *10*, 287–299. [CrossRef]
71. Wong, W.; Alomari, Z.; Liu, Y.; Jura, L. Linkage deanonymization risks, data-matching and privacy: A case study. In Proceedings of the 2024 8th International Conference on Cryptography, Security and Privacy (CSP 2024), Osaka, Japan, 20–22 April 2024; pp. 10–16. [CrossRef]
72. Abulaish, M.; Haldar, N.A.; Jahiruddin. P2DF: A privacy-preserving digital forensics framework. *Int. J. Digit. Crime Forensics* **2021**, *13*, 15. [CrossRef]
73. Rowlingson, R. A Ten Step Process for Forensic Readiness. *Int. J. Digit. Evid.* **2004**, *2*, 1–28.
74. Cova, M.; Kruegel, C.; Vigna, G. Detection and analysis of drive-by-download attacks and malicious infrastructures. In Proceedings of the 19th International World Wide Web Conference (WWW), Raleigh, NC, USA, 26–30 April 2010; pp. 281–290. [CrossRef]
75. Ligh, M.H.; Case, A.; Levy, J.; Walters, A. *The Art of Memory Forensics: Detecting Malware and Threats in Windows, Linux, and Mac Memory*; Wiley: Indianapolis, IN, USA, 2014.
76. Garfinkel, S. Digital forensics research: The next 10 years. *Digit. Investig.* **2010**, *7*, S64–S73. [CrossRef]
77. AlEroud, A.; Karabatis, G. Bypassing detection of URL-based phishing attacks using generative adversarial deep neural networks. In *Proceedings of the 6th ACM International Workshop on Security and Privacy Analytics (IWSPA), New Orleans, LA, USA, 18 March 2020*; ACM: New York, NY, USA, 2020; pp. 53–60. [CrossRef]
78. Deshpande, A.S.; Gupta, S. GenAI in the Cyber Kill Chain: A Comprehensive Review of Risks, Threat Operative Strategies and Adaptive Defense Approaches. In *Proceedings of the 2023 IEEE International Conference on ICT in Business Industry & Government (ICTBIG), Indore, India, 8–9 December 2023*; IEEE: Piscataway, NJ, USA, 2023; pp. 1–5. [CrossRef]
79. Jabir, R.; Le, J.; Nguyen, C. Phishing Attacks in the Age of Generative Artificial Intelligence: A Systematic Review of Human Factors. *AI* **2025**, *6*, 174. [CrossRef]