

Review

A Comprehensive Survey of Artificial Intelligence Applications in Cyber Security: Taxonomy, Challenges, and Future Directions

Shahid Alam ^{1,*}, Ehab Alnfwawy ¹, Amina Jameel ², Sana Qadir ³, Zahida Parveen ¹, Basirah Noor ³, Anushya Arol ⁴ and Imran Chaudhry ⁵

¹ Department of Information Security, College of Computer Science and Engineering, University of Ha'il, Ha'il 81481, Saudi Arabia; alnfwawy@gmail.com (E.A.); z.malik@uoh.edu.sa (Z.P.)

² Department of Computer Engineering, Center of Excellence in Artificial Intelligence, Bahria University, Islamabad 44000, Pakistan; amina@bahria.edu.pk

³ Department of Information Security, Faculty of Computing, National University of Sciences and Technology, Islamabad 44000, Pakistan; sana.qadir@seecs.edu.pk (S.Q.); bnoor.msis19seecs@seecs.edu.pk (B.N.)

⁴ Department of Artificial Intelligence and Data Science, College of Computer Science and Engineering, University of Ha'il, Ha'il 81481, Saudi Arabia; anushya.arol@uoh.edu.sa

⁵ Department of Industrial Engineering, College of Engineering, University of Ha'il, Ha'il 81481, Saudi Arabia; i.chaudhry@uoh.edu.sa

* Correspondence: sha.alam@uoh.edu.sa

Abstract

The increasing complexity and scale of cyber threats demand intelligent and adaptive defense mechanisms that extend beyond traditional approaches. Artificial Intelligence (AI) has emerged as a key enabler for enhancing cyber security through automated detection, analysis, and response. This paper presents a comprehensive survey of AI applications in cyber security across five major domains: malware detection, intrusion detection, phishing and spam detection, botnet detection, and cyber forensics. A systematic methodology based on data and methodological triangulation is employed to analyze 75 studies published between 2021 and 2025. The paper introduces a multi-layer taxonomy that maps cyber threats to application domains, analysis methods, AI approaches, and their associated capabilities and limitations. In addition, a cross-domain meta-analysis is conducted to identify recurring trends and assess the adoption of AI across different cyber security scenarios. The analysis reveals that deep learning and transformer-based models dominate data-intensive domains such as intrusion detection and malware analysis, whereas traditional machine learning techniques remain effective in structured and resource-constrained settings, particularly for phishing detection. Key challenges include dataset limitations, limited explainability, adversarial vulnerabilities, and computational constraints. Unlike existing surveys that focus on specific techniques or individual cyber security domains, this work provides a unified, application-oriented perspective on AI-driven cyber security. It further highlights emerging trends, open challenges, and future research directions toward more robust, scalable, and trustworthy cyber security systems.



Academic Editor: Ulrich Kerzel

Received: 7 May 2026

Revised: 23 June 2026

Accepted: 24 June 2026

Published: 30 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: artificial intelligence (AI); cyber security; malware detection; intrusion detection systems; phishing detection; spam detection; botnet detection; cyber forensics; machine learning; deep learning; explainable AI

1. Introduction

The field of Artificial Intelligence (AI) [1] was formally introduced in 1955 through the pioneering proposal by McCarthy et al. [2], which outlined foundational goals such as auto-

mated reasoning, natural language understanding, learning systems, and self-improvement. Since then, AI has evolved into a broad interdisciplinary domain encompassing machine learning (ML), deep learning (DL), neural networks, and statistical learning techniques. In particular, recent advances in DL have significantly accelerated the development of intelligent systems capable of complex pattern recognition and decision-making [1]. Despite these advancements, modern AI systems continue to face critical challenges, including a lack of transparency, susceptibility to adversarial manipulation, privacy concerns, algorithmic bias, and ethical limitations.

In parallel, the rapid expansion of digital ecosystems driven by cloud computing, the Internet of Things (IoT), and large-scale interconnected networks has fundamentally transformed modern society. However, this transformation has also significantly increased the cyber attack surface, making *Cyber Security* a critical requirement for governments, organizations, and individuals. As cyber threats continue to evolve in complexity and scale, traditional rule-based security mechanisms are no longer sufficient to provide effective protection.

To address these challenges, AI has emerged as a key enabler in cyber security, supporting automated threat detection, intelligent analysis, and adaptive response mechanisms. AI-driven techniques are now widely applied across several critical domains, including malware detection, intrusion detection systems (IDS), phishing and spam detection, botnet detection, and cyber forensics. These applications demonstrate the strong potential of AI to enhance both proactive and reactive cyber defense capabilities.

However, despite this progress, several fundamental challenges remain unresolved in the existing literature. Current AI-based cyber security systems often suffer from data scarcity and imbalance, limited generalization across diverse attack scenarios, and constraints in real-time deployment. More importantly, the black-box nature of many AI models raises significant concerns regarding interpretability, trust, and accountability, particularly in high-stakes environments such as cyber forensics and legal investigations. In addition, AI systems remain vulnerable to adversarial machine learning (AML) attacks, where carefully crafted inputs can mislead models and cause incorrect threat classification, highlighting the need for more robust and secure AI frameworks.

Although several survey studies [3–7] exist in this domain, most of them either focus on specific AI techniques or are limited to narrow cyber security subfields such as intrusion detection or malware analysis. Furthermore, recent developments in emerging paradigms such as explainable AI (XAI) and generative AI (GAI) are not comprehensively addressed in many existing surveys. There is also a lack of unified, application-oriented frameworks that systematically map cyber threats to AI-driven security solutions across multiple domains.

To address these limitations, this paper presents a comprehensive and systematic survey of AI applications in cyber security, focusing on literature published between 2021 and 2025. A rigorous multi-stage methodology based on systematic literature collection, keyword filtering, duplication removal, and thematic classification is employed to ensure the reliability and reproducibility of the review process. The selected studies are organized into an application-oriented taxonomy covering five major domains: malware detection, intrusion detection, phishing and spam detection, botnet detection, and cyber forensics.

The main contributions of this study are threefold. First, it proposes a structured taxonomy of AI applications in cyber security based on a systematic threat-to-domain mapping framework. Second, it presents a comprehensive review of recent literature published between 2021 and 2025, synthesizing key techniques, trends, and domain-specific developments. Third, it identifies major challenges, research gaps, and emerging research directions, including adversarial robustness, explainability, and generative AI integration, to support future research in this area.

The increasing global interest in both AI and cyber security, as illustrated by the search trend analysis in Figure 1, further underscores the importance and timeliness of this research area. This motivates the need for a unified, up-to-date, and application-oriented survey that can support researchers and practitioners in developing more robust and trustworthy AI-driven cyber security systems. To achieve this objective, the study seeks to answer the following research questions:

RQ1. How is Artificial Intelligence currently applied across different cyber security domains, including malware detection, intrusion detection, phishing and spam detection, botnet detection, and cyber forensics?

RQ2. What AI techniques and models are most commonly used across these cyber security domains, and how do their applicability and reported performance differ across use cases?

RQ3. What key challenges, limitations, and research gaps in the application of AI to cyber security have been identified in recent literature (2021–2025)?

RQ4. What emerging AI paradigms and future research directions, including generative AI and explainable AI, can enhance the effectiveness, robustness, and trustworthiness of cyber security systems?

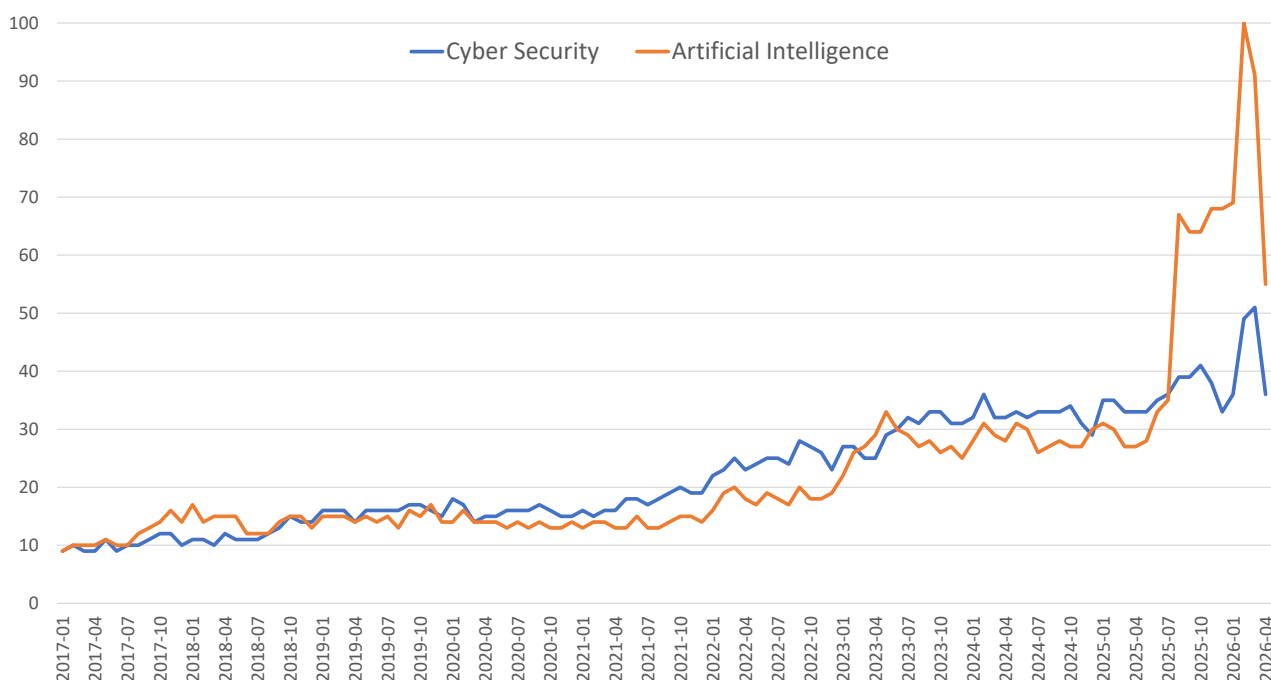


Figure 1. Monthly Google search trends [8] over the past ten years for the terms *Artificial Intelligence* and *Cyber Security* are presented. The y-axis represents the relative frequency of searches for each term. A value of 100 indicates peak popularity, while a value of 50 signifies that the term is half as popular. A score of 0 means there was insufficient data available for that term.

Organization of the Paper

The remainder of the paper is structured as follows. Section 2 outlines the methodology designed and used to conduct the survey presented in this paper. Section 3 offers a comparison between our work and previous studies that have examined the use and integration of AI in Cyber Security. Section 4 presents essential background information by reviewing fundamental concepts in both AI and Cyber Security. Section 5 presents the formalization of the taxonomy introduced in the paper. Section 6 provides cross-domain meta-analysis of the studies surveyed in this paper. Section 7 delivers an in-depth review of past research across several critical domains within Cyber Security. Section 8 explores the major challenges highlighted in existing literature and their impact on AI applications in

the field of Cyber Security. Section 9 identifies current research gaps and emerging trends that must be addressed to further strengthen AI's contribution to Cyber Security. Section 10 addresses the potential threats to the validity of the study presented in this paper. Lastly, Section 11 wraps up the paper with concluding remarks.

2. Research Methodology

Figure 2 displays the flow chart that describes the research methodology implemented in the survey of this paper. This comprehensive flow chart serves as a strategic blueprint, guiding the systematic execution of the five distinct surveys and reviews that underpin this research. Each flowchart component reflects a carefully considered step in the methodology, ensuring a thorough and organized approach to data collection and analysis throughout the study.

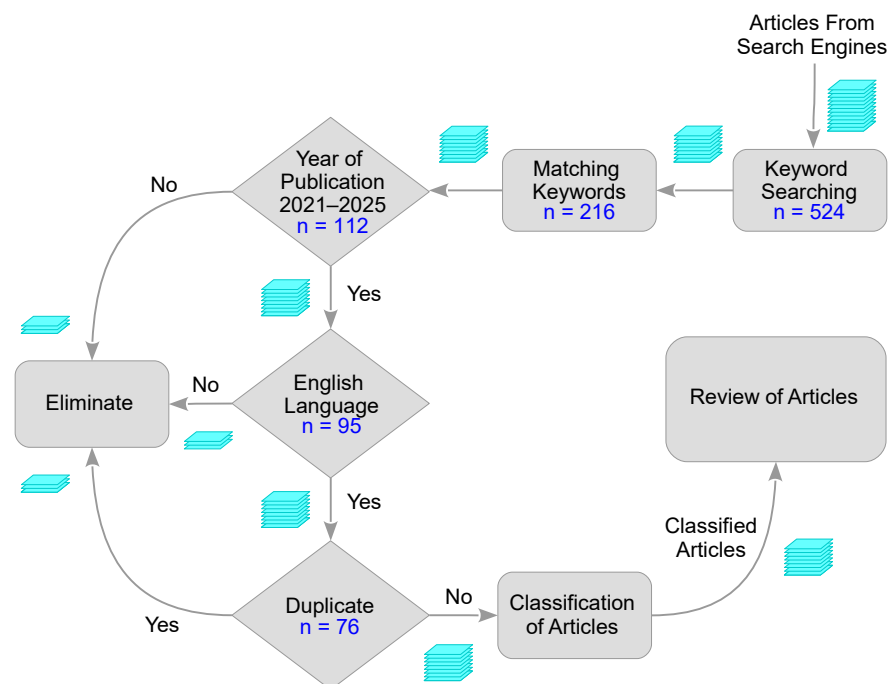


Figure 2. Flowchart depicting the research methodology and process used for the survey conducted in this paper. where n corresponds to the number of articles remaining following each screening phase.

To strengthen the credibility of the study presented in this paper, we rigorously applied the principle of *triangulation* [9,10]. This approach entails using at least three distinct methods to thoroughly verify our procedures, data, and findings, ensuring a robust and reliable outcome. We employed scite (scite.ai), an online AI-driven research tool, alongside Google Scholar searches and manual verification of data from various sources, thereby enabling data triangulation. Additionally, three reviewers were involved in validating and confirming the procedures and methods adopted in this study, representing methodological triangulation. Any disagreements regarding article eligibility, classification, or thematic categorization were discussed among the reviewers until consensus was reached.

To ensure a comprehensive and diverse collection of literature related to the study presented in this article, we collected articles from various search engines. To help filter and identify articles that specifically address AI and Cyber Security, we employ **Keyword Searching**: (“Artificial Intelligence” OR “Machine Learning” OR “Deep Learning” OR “Transformer” OR “Large Language Model”) AND (“Cyber Security” OR “Cybersecurity” OR “Malware Detection” OR “Intrusion Detection” OR “Spam and Phishing Detection” OR “Botnet Detection” OR “Cyber Forensics”). As part of the triangulation process, we

used <https://www.scite.ai> (accessed on 5 May 2026), *Google Scholar Search*, and *Manual Confirmation* of the articles collected from different sources. Then we refined our search further by **Matching Keywords**. Various keywords used were: *malware*, *intrusion*, *phishing*, *spam*, and *botnet*, etc. We check the articles by the **Year of Publication** from 2021–2025, i.e., articles published in the last five years. The next step is to perform **English Language** check. If the answer to any of these checks is no, the article is directed to the **Eliminate** stage; otherwise, the article goes to the next stage to check for the **Duplicate**. If the article is duplicated, it is sent to the elimination stage. Table 1 summarizes the quantitative filtering statistics of this article screening stage.

Table 1. Article screening stage.

Screening Stage	Number of Articles	Item	Description
Initial search results	524	Databases	Google Scholar, scite.ai, IEEE Xplore, ACM Digital Library, ScienceDirect, SpringerLink, Scopus, Web of Science
After keyword matching	216	Keywords	AI, ML, DL, Malware, IDS, Spam/Phishing, Botnet, Forensics
After year filtering	112	Period	2021–2025
After language filtering	95	Language	English
After duplicate removal	75	Reviewers	3

The specific inclusion and exclusion criteria used in this study are as follows: **Inclusion Criteria:** (i) articles published between 2021 and 2025; (ii) English-language publications; (iii) studies focusing on AI applications in cyber security; (iv) peer-reviewed journal or conference papers. **Exclusion Criteria:** (i) duplicate records; (ii) non-English publications; (iii) studies outside the scope of AI and cyber security; (iv) articles lacking sufficient technical content relevant to the objectives of this survey.

Through this process, we successfully identified a total of 75 articles for in-depth classification and review. These articles are thoroughly examined in Sections 3 and 7 of this paper, providing valuable insights into the current landscape of research in the field.

After finding the unique articles, we proceed to the **Classification of Articles**. This process systematically groups the articles according to their main focus. As noted before for this paper, our primary focus is on five areas of application of Cyber Security. Therefore, we group these articles into five groups: *Malware detection*; *Intrusion detection*; *Phishing and Spam detection*; *Botnet detection*; and *Cyber forensics*. The reason for dividing them into these groups is to provide a structural approach to analyzing the role and effect of AI on Cyber Security. After classifying the articles, the next pivotal step is the **Review of Articles**. This phase involves a thorough examination of each group, seeking to reveal valuable insights, patterns, and findings. The detailed review of classified articles is vital to synthesize results, uncover gaps in the existing literature, and deepen our understanding of the nuances within each thematic category.

In the essence of *methodological triangulation*, we engaged three reviewers to verify and affirm the procedures and methods employed in this study, as illustrated in the flow chart in Figure 2.

In summary, this systematic journey through the research methodology—ranging from collecting articles and refining them with keyword searches to the careful classification and review stages—provides a comprehensive understanding of the studies discussed in this paper. The insights gained from this systematic process contribute to a deeper understanding of the subject, paving the way for meaningful discussions, valuable implications, and potential advancements in the field of AI applied to Cyber Security.

3. Related Work

In this section, we compare our study with previous studies that have surveyed/ reviewed the application and integration of AI in Cyber Security. There are recent review works that cover only the implications of specific AI techniques/models in Cyber Security, such as [7] present an extensive review of the role and application of XAI in Cyber Security; Sai et al. [6] discuss in detail real-life applications of Generative AI in Cyber Security. We do not include them here because of their specificity.

We selected three recent studies/surveys published in the years 2024 and 2025. One [3] is published in Elsevier Information Fusion, and the other two [4,5] are published in IEEE Access. Table 2 illustrates a comparison between these previous surveys and the survey conducted in this article.

Table 2. A comparison of our survey and review with other relevant scholarly works discussed in Section 3.

Survey	Year	Major Domains Covered	Taxonomy	Challenges with Solutions	Future Directions	Advanced AI (GPT, BERT, LLM, XAI)
Our Survey [3] [4] [5]	2026	Yes	Yes	Yes	Yes	Yes
	2025	Yes	No	Yes	Yes	No
	2024	Yes	No	Yes	No	Yes
	2024	Yes	No	No	Yes	No

The study presented in [3] is the most recent and relevant to the research work carried out in this paper. It is a comprehensive study done on this subject. The study provides a broad overview of AI applications in cyber security, with comparatively less emphasis on detailed analyses of specific cyber security domains. The survey reviews a selected set of representative studies across multiple cyber security domains. A detailed review of current and previous studies involving AI and Cyber Security is missing. For example, there are only two papers specifically reviewed on malware analysis and detection, and also two papers on cyber forensics. Malware analysis and detection, and cyber forensics analysis are the two most important applications of Cyber Security. Most recent AI models, such as generative pretrained transformers (GPT), are not included in the study.

Ref. [4] presents a comprehensive survey of AI and machine learning techniques applied to Cyber Security. The paper places greater emphasis on explaining AI techniques than on discussing Cyber Security itself. While the challenges of applying AI in Cyber Security are addressed, potential solutions to mitigate these challenges are not provided. Furthermore, future research directions are not elaborated upon in detail.

Ref. [5] presents a comprehensive review of AI-driven techniques used to detect different cyber attacks. These attacks include malware attacks, spam, network intrusions, etc. Similarly to the study presented in [3]. The survey places greater emphasis on AI techniques and methodologies than on domain-specific cyber security applications. A detailed background on AI techniques is provided, but the discussion of individual cyber security application domains is comparatively concise. They divide AI into three areas: machine learning, DL, and metaheuristic algorithms. In addition, advanced AI techniques and models, such as GPT and XAI, are not included in the study. A detailed list of challenges is presented, but the solutions to these challenges are not discussed in detail.

Unlike existing surveys that focus primarily on AI/ML techniques or isolated cyber security domains, this study introduces a unified, application-oriented taxonomy of Artificial Intelligence in cyber security. It systematically maps cyber security threats to key defense domains, including malware detection, intrusion detection, phishing and spam detection, botnet detection, and cyber forensics. In addition, it provides an up-to-date systematic review (2021–2025) and incorporates emerging paradigms such as XAI and GAI.

This integrated perspective bridges the gap between AI methods and real-world cyber security applications, enabling clearer identification of research gaps and future directions.

Furthermore, a growing body of literature has emerged around Trustworthy AI, which extends beyond individual topics such as explainability and adversarial robustness. Trustworthy AI frameworks emphasize transparency, accountability, robustness, reliability, safety, and human oversight as essential requirements for deploying AI systems in high-stakes domains [11]. Recent studies [11,12] have highlighted the importance of these dimensions for developing secure and dependable AI-enabled systems. However, compared with existing cyber security surveys, broader trustworthiness considerations are often discussed only in a fragmented manner. In this survey, these aspects are reflected through the discussion of explainability, adversarial robustness, reliability, and deployment considerations across different cyber security applications, thereby providing a more holistic perspective on the trustworthy adoption of AI in cyber security.

4. Background

This section provides the necessary foundational context by reviewing the core concepts of AI and Cyber Security. It outlines the key definitions, methodologies, applications, and challenges pertinent to understanding the implications of AI within the Cyber Security domain, drawing upon recent literature and established principles.

4.1. Artificial Intelligence (AI)

AI encompasses the theory and application of computer systems designed to perform tasks that traditionally require human cognitive abilities, such as learning, problem-solving, perception, and decision-making [1,13]. Over the past decade, AI has evolved significantly, driven by advancements in computational power and data availability, leading to its widespread adoption across various sectors [14]. The global AI market's substantial growth projections underscore its transformative potential [7].

Key areas, as depicted in Figure 3 within AI, include ML and its subfield, DL. ML algorithms enable systems to learn patterns from data without explicit programming [15]. In Cyber Security, traditional ML techniques such as Support Vector Machines (SVM) and Random Forests (RF) have demonstrated considerable success in specific applications like intrusion detection in Internet of Things (IoT) environments [16]. DL employs artificial neural networks with multiple layers (deep architectures) to model intricate patterns, particularly effective with large, complex datasets [17]. Techniques like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Deep Belief Networks (DBNs) have shown potential to enhance threat detection accuracy, scalability, and performance [18]. Furthermore, hybrid approaches combining methods like transformer-based models and federated learning are being investigated to tackle sophisticated threats, including Advanced Persistent Threats (APTs), and to address challenges in emerging technological landscapes like Industry 5.0 and 5G networks.

However, the increasing reliance on ML models, particularly in security-critical applications, has highlighted their vulnerability to AML attacks [19]. AML involves manipulating AI systems through maliciously designed inputs. Key attack types include evasion attacks, where adversaries craft inputs (e.g., slightly modified malware files or network packets) at inference time to cause misclassification by a deployed model, and poisoning attacks, where attackers inject malicious data into the training set to compromise the learning process and degrade model performance or insert backdoors. The existence of such attacks underscores the need for robust defences and thorough evaluation of AI models deployed in adversarial environments like Cyber Security [20].

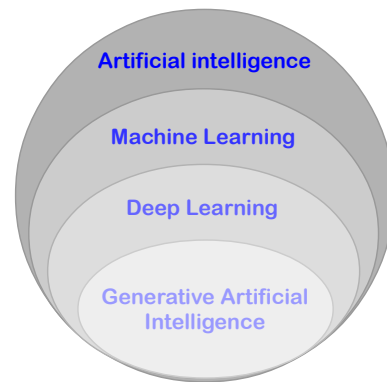


Figure 3. Key areas of AI.

Recent advancements have also led to the rise of powerful Large Language Models (LLMs), such as Generative Pre-trained Transformers (GPT), and broader Generative AI (GAI) techniques [21]. These models exhibit remarkable capabilities in understanding and generating human-like text, code, images, and other data. In Cyber Security, they present a dual-use potential: defensively, they can assist in tasks like analysing security reports, generating realistic honeypot data, or aiding in code vulnerability analysis; offensively, they could be misused to craft highly convincing phishing emails, generate polymorphic malware variants, or automate reconnaissance tasks [22]. The security implications and ethical considerations surrounding these advanced AI techniques are active areas of research and debate. The emerging open source alternatives, such as DeepSeek, have demonstrated significant potential in cyber security tasks. However, compared with GPT-based systems, cyber security-specific studies involving DeepSeek remain relatively limited, highlighting an important direction for future research.

Despite their power, many advanced AI models, especially DL systems, often function as *black boxes*, where the internal logic leading to a specific output is not readily understandable [7,23]. This opacity presents a significant barrier to adoption in high-stakes fields like Cyber Security, where transparency and trustworthiness are critical [24]. Explainable Artificial Intelligence (XAI) has emerged to address this challenge, focusing on developing techniques and models that offer understandable justifications for their outputs while preserving high performance levels [25]. The core principles of XAI emphasize providing meaningful, accurate explanations and defining the operational limits of the AI system [26]. Frameworks such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) represent post-hoc approaches to interpreting complex models [7]. However, the application of XAI in Cyber Security introduces a unique dilemma: increased transparency might inadvertently reveal vulnerabilities that could be exploited by malicious actors [24].

4.2. Cyber Security

Cyber Security encompasses the strategies, technologies, and practices employed to safeguard digital assets—including networks, computers, devices, software, and data—from unauthorized access, disruption, modification, or destruction [27]. The fundamental goal is to maintain the Confidentiality, Integrity, and Availability (commonly known as the CIA triad) of information systems in the face of diverse cyber threats [28]. The domain is extensive, covering network security, information security, application security, operational security, cloud security, and the protection of IoT ecosystems [29]. As societies become increasingly reliant on interconnected digital infrastructures, the importance of robust Cyber Security measures has grown exponentially, driving significant expansion in the global Cyber Security market.

The field of cyber security involves a continuous battle between defenders and attackers. Malicious actors continuously devise new methods to exploit vulnerabilities in technology and human behavior [30]. The threat landscape includes a wide spectrum of attacks, such as various forms of malware (viruses, ransomware, spyware), phishing campaigns designed to deceive users into revealing sensitive information, distributed denial-of-service (DDoS) attacks aimed at overwhelming systems, and sophisticated, targeted attacks like APTs and zero-day exploits [31,32].

The top 10 Cyber Security threats are listed in Table 3. AI is being applied across several key and critical application areas of Cyber Security. These key areas are discussed below.

Table 3. Top 10 Cyber Security threats [33].

Threat	Description
Malware	Malware , short for malicious software, refers to any program or code created to disrupt, damage, or gain unauthorized access to computers, networks, or data.
DDos Attacks	A DDoS (Distributed Denial of Service) attack is a harmful effort to interrupt the regular functioning of a targeted server, service, or network by bombarding it with excessive Internet traffic originating from numerous sources (e.g., botnet).
Phishing Attacks	Phishing attacks involve social engineering tactics in which attackers deceive individuals into disclosing sensitive information, such as passwords, credit card details, or login credentials, by posing as a trustworthy or legitimate entity.
Ransomware	Ransomware is a form of malicious software (malware) that locks a victim's system or encrypts their files, making the data unusable. The attacker typically demands a ransom, often in cryptocurrency, in exchange for restoring access.
APTs	APTs (Advanced Persistent Threats) are carefully orchestrated, covert, and prolonged cyberattacks carried out (usually by botnet) by highly resourced adversaries intended to steal confidential data, perform espionage, or compromise essential systems.
XSS	XSS (Cross-Site Scripting) is a web security flaw in which attackers embed malicious scripts (malware) into content from legitimate websites. When users access the affected site, the malicious code runs in their browsers, typically without their awareness.
Credential Stuffing	Credential stuffing is a cyberattack method where attackers employ automated tools (e.g., phishing) to input stolen username and password combinations, often obtained from data breaches, across various websites, taking advantage of the common practice of password reuse.
Social Engineering Attacks	Social engineering attacks are manipulative tactics (e.g., through spam , phishing , or malware) used by attackers to trick individuals into revealing confidential information or performing actions that compromise security. Instead of exploiting technical vulnerabilities, these attacks exploit human psychology and trust. Cyber forensics is often used to track these attacks.
Cryptojacking	Cryptojacking is a cyberattack (e.g., installing malware) where attackers secretly use someone else's computer or device to mine cryptocurrency without their knowledge or consent. The victim's resources (CPU, GPU, battery) get drained, slowing down their device.
Insider Threats	Insider threats occur when trusted individuals within an organization, employees, contractors, or partners, intentionally or unintentionally (e.g., through spam and phishing), cause harm to the organization's data, systems, or operations.

Malware Detection: Malware detection is the process of identifying malicious software (e.g., viruses, ransomware, spyware) designed to disrupt, damage, or gain unauthorized access to computer systems. AI techniques analyze static features (e.g., byte sequences, file structure) and dynamic behaviors (e.g., API calls, network communications) of software to classify it as benign or malicious [18]. This approach improves the detection rates for known malware families and, crucially, enables the identification of novel, zero-day malware variants that evade traditional signature matching [34]. Methods range from analyzing system call sequences to using DL for image-based malware classification [7].

Intrusion Detection Systems (IDS): An Intrusion Detection System is a security tool that monitors network or system activities for signs of malicious incidents or policy

violations. AI-powered IDS analyzes network traffic data (e.g., packet headers, flow statistics) or host-based logs (e.g., system calls, process activity) to detect unauthorized access or malicious activities within a network or on an endpoint [16]. DL models like CNNs and RNNs are particularly adept at capturing complex temporal and spatial patterns indicative of sophisticated intrusion attempts [35]. This enables the identification of stealthy, multi-stage attacks that might otherwise go unnoticed, offering a critical layer of defense against unauthorized access to sensitive data and network infrastructure.

Spam and Phishing Detection: This application focuses on identifying and filtering fraudulent communications (phishing) designed to steal sensitive information and unsolicited bulk messages (spam). AI classifiers scrutinize various elements of electronic communications, including email content, sender information, embedded URLs, and website characteristics, to differentiate legitimate messages from phishing attacks or unsolicited spam [34]. NLP and DL models enhance the ability to understand context, identify social engineering cues, and counter evolving attacker tactics. This is vital for protecting users from credential theft, financial fraud, and malware distribution by automatically blocking malicious messages that are specifically crafted to deceive human recipients.

Botnet Detection: Botnet detection is the process of identifying a network of compromised devices (bots) that are remotely controlled by an attacker to carry out coordinated malicious activities. AI contributes by analyzing network traffic for characteristic patterns of botnet Command-and-Control (C&C) communication, identifying algorithmically generated domain names (DGAs), or monitoring device behavior (especially in IoT contexts) for signs of compromise and participation in a botnet. Graph Neural Networks (GNNs) are also effective in modeling the network structure and communication relationships to pinpoint botnet clusters. Identifying and neutralizing botnets is essential for preventing large-scale, automated attacks such as Distributed Denial-of-Service (DDoS), massive spam campaigns, and click fraud, thereby safeguarding the stability and integrity of the broader internet.

Cyber Forensics: Cyber forensics involves the collection, examination, and preservation of digital evidence in a manner that is legally admissible to investigate cybercrimes. AI applications in digital forensics are emerging to help investigators analyze vast amounts of digital evidence more efficiently [36]. Potential uses include automating the identification of relevant artifacts, classifying file types, detecting data manipulation or anomalies, and identifying tampered multimedia content like deepfakes. However, the *black-box* nature of many AI models presents challenges regarding the admissibility and explainability of AI-derived evidence in legal contexts [37].

The integration of AI has marked a paradigm shift in Cyber Security capabilities, offering enhanced methods for defense. AI-driven systems leverage ML and DL to analyze vast quantities of data, identifying subtle patterns and anomalies that may indicate malicious activity often missed by traditional signature-based or rule-based systems.

Despite these benefits, deploying AI in Cyber Security presents considerable challenges. Effective AI models require vast amounts of high-quality data, and managing this data securely and efficiently is complex. AI systems must continuously adapt to the rapidly evolving tactics of cybercriminals, requiring ongoing training and model updates. Furthermore, AI models themselves can become targets; adversarial attacks aim to deceive AI classifiers, while data poisoning attacks can corrupt the training data [38]. The inherent lack of interpretability in many AI models remains a significant hurdle, impacting trust and the ability of security professionals to validate AI-driven decisions. Overcoming these technical and conceptual challenges, alongside addressing ethical considerations, is crucial for maximizing AI's contribution to Cyber Security [39]. Future research directions emphasize the development of more advanced, robust, and explainable AI methods specifically designed for the dynamic and adversarial nature of the Cyber Security domain [40].

5. Taxonomy of AI Applications in Cyber Security

To deliver a more organized and comprehensive perspective on the role of AI in cyber security, this study introduces a multi-dimensional taxonomy that goes beyond traditional flat classification schemes. Rather than merely linking cyber threats to specific detection domains, the proposed framework integrates several analytical dimensions, including threat attributes, analytical approaches, AI methodologies, operational functions, as well as associated constraints and limitations. This layered structure facilitates a more thorough and systematic assessment of how AI-based solutions are implemented across a wide range of cyber security contexts.

5.1. Taxonomy Design Principles

The taxonomy is built upon the categorization of common cyber threats outlined in Table 3, encompassing malware, distributed denial-of-service (DDoS) attacks, phishing, ransomware, APTs, XSS, credential stuffing, social engineering, cryptojacking, and insider threats. Based on their functional behavior and corresponding defense needs, these threats are systematically assigned to five key application domains: *malware detection*, *intrusion detection*, *phishing and spam detection*, *botnet detection*, and *cyber forensics*.

To overcome the constraints associated with shallow classification frameworks, we introduce a well-defined and structured mapping function:

$$f : \mathcal{T} \rightarrow \mathcal{D} \rightarrow \mathcal{M} \rightarrow \mathcal{C}$$

where:

$\mathcal{T}$ represents the set of cyber security threats,

$\mathcal{D}$ denotes application domains,

$\mathcal{M}$ corresponds to AI methodologies, and

$\mathcal{C}$ captures capabilities, constraints and limitations.

This formalization, shown in Figure 4, is integrated within the classification stage of the research methodology (Figure 2), ensuring that the taxonomy is systematically derived rather than heuristically defined.

Studies were assigned to application domains according to their primary cyber security objective as reported by the original authors. When multiple AI techniques were employed, classification was based on the dominant methodology or model architecture most directly associated with the main reported outcomes.

5.2. Multi-Level Taxonomy Structure

1. *Threat Layer ($\mathcal{T}$)*—This layer captures the range of cyber threats listed in Table 2, encompassing diverse behaviors such as static versus dynamic actions, network-oriented versus host-based intrusions, and human-focused versus system-focused exploitation patterns.
2. *Application Domain Layer ($\mathcal{D}$)*—In this layer, threats are categorized into five principal domains: Malware Detection, Intrusion Detection, Phishing and Spam Detection, Botnet Detection, and Cyber Forensics. These groupings offer a functional perspective on cyber security defense strategies by aligning threats with their respective operational contexts.
3. *Analysis and Methodology Layer ($\mathcal{M}$)*—Each domain is further defined by the analytical techniques it employs. These include static analysis, which focuses on signature-based and code-level examination; dynamic analysis, involving runtime behavior tracking; network traffic analysis, emphasizing packet- and flow-level monitoring; content and semantic analysis, used for text and URL evaluation in phishing detection; and forensic

analysis, which supports evidence reconstruction and attribution. The selection of these methods influences the appropriateness of different AI models.

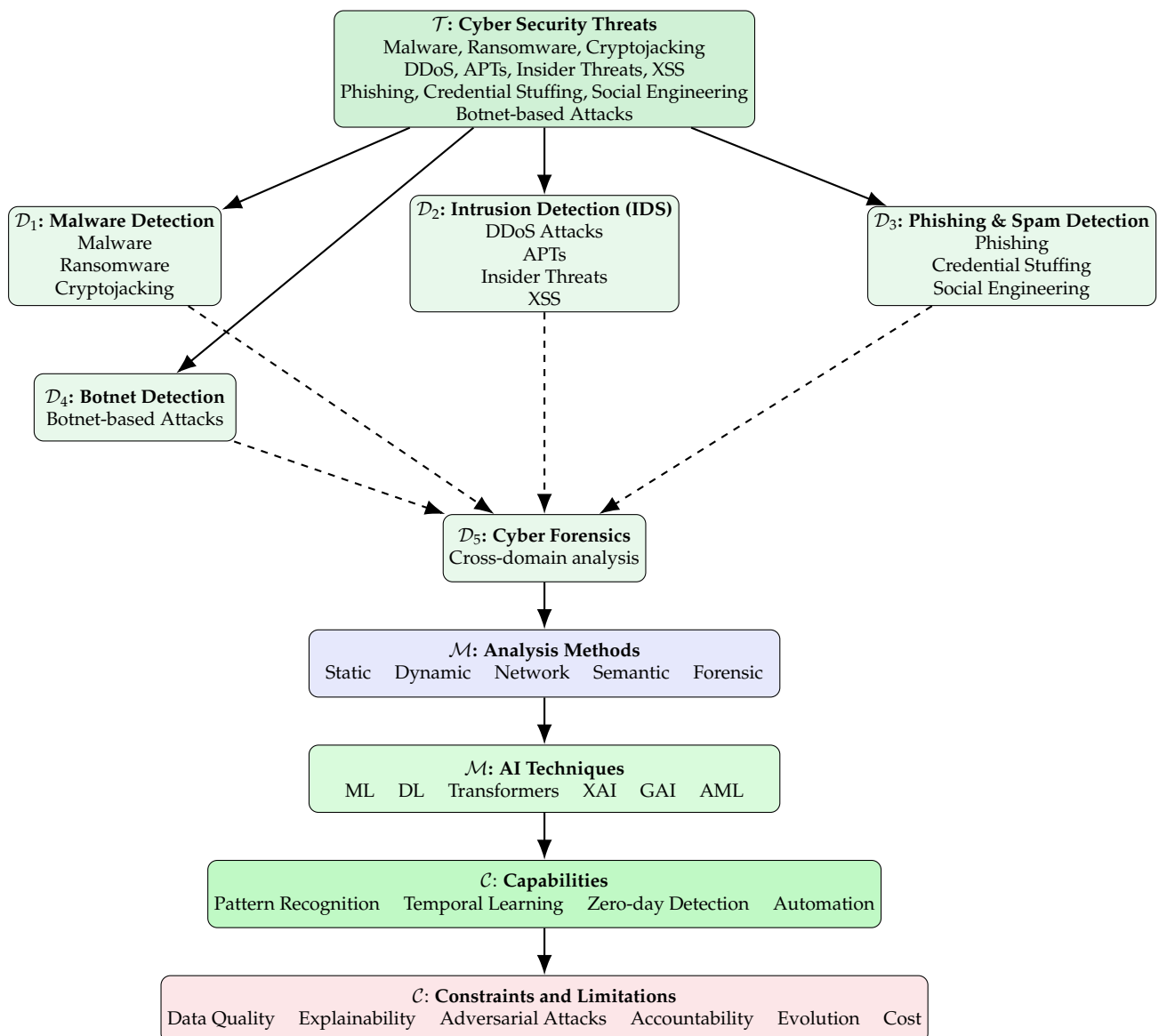


Figure 4. Multi-layer taxonomy of AI in cyber security incorporating threats, domains, analysis methods, AI techniques, capabilities, constraints, and limitations. Solid arrows $\Rightarrow$ primary taxonomy hierarchy and Dashed arrows $\Rightarrow$ cross-domain relationships.

4. *AI Techniques Layer (M)*—A variety of AI approaches are utilized across the domains, including traditional ML methods such as Support Vector Machines, Random Forests, and Decision Trees; DL models like CNNs, LSTMs, and RNNs; Transformer-based architectures that leverage attention mechanisms for sequence processing; XAI techniques to enhance transparency and trust; and GAI, which is increasingly relevant for both offensive and defensive applications. The choice of approach depends heavily on the characteristics of the threat and the form of the data.
5. *Capability, Constraint and Limitation Layer (C)*—A distinguishing aspect of this taxonomy is the explicit consideration of system performance factors. Key capabilities include the ability to identify complex patterns in high-dimensional data, model temporal relationships, detect zero-day attacks, and enable automated prediction and response. At the same time, several limitations are acknowledged, such as reliance

on large and potentially imbalanced datasets, limited interpretability of black-box models, susceptibility to adversarial manipulation, and significant computational demands that may hinder real-time deployment. This layer supports comprehensive cross-domain evaluation.

5.3. Cross-Domain Observations

The multi-layered taxonomy demonstrates that no single AI method is universally superior across all scenarios. For example, DL models perform particularly well in intrusion detection because of their strength in capturing temporal dependencies, while conventional ML approaches continue to be effective in phishing detection due to the presence of well-structured feature spaces. In addition, cyber forensics demands interpretable models to support trust and legal validity, underscoring the critical role of XAI.

5.4. Significance of the Proposed Taxonomy

By combining threat categorization, analytical approaches, AI methodologies, and performance-related attributes within a single cohesive framework, the proposed taxonomy offers:

- A well-organized and holistic perspective on the role of AI in cyber security
- A basis for conducting comparative evaluations and meta-level analyses
- A systematic way to uncover research gaps and existing limitations
- Practical guidance for choosing suitable AI techniques for particular threat contexts

Overall, this enriched taxonomy overcomes the shortcomings of prior survey models and responds to the growing demand for more in-depth, application-focused analysis in AI-driven cyber security studies. The following subsections further elaborate on the definitions and scope of each domain within $\mathcal{D}$.

6. Cross-Domain Meta-Analysis of AI Techniques in Cyber Security

In this section, we present a cross-domain meta-analysis of the studies reviewed in Section 7 to identify recurring patterns, evaluate the effectiveness of AI techniques across different cyber security domains, and synthesize common strengths, limitations, and research gaps. Unlike conventional survey discussions that summarize individual works, this analysis is structured in alignment with the proposed multi-layer taxonomy and is supported by the unified taxonomy Table 4 and the heatmap-based comparative visualization Table 5.

To improve consistency in the meta-analysis, the AI techniques reported in Tables 4 and 5 are grouped into three categories: AI paradigms (e.g., ML, DL, XAI, and AML), model architectures (e.g., Transformers and GNNs), and representative model families (e.g., BERT, GPT, and other LLMs). These categories are used to summarize research trends rather than compare individual models directly. Studies were categorized according to the primary AI technique reported by the authors. When multiple techniques were employed, the dominant model architecture or methodological paradigm was used for classification.

Table 4. Unified Multi-Layer Taxonomy and Meta-Analysis of AI in Cyber Security. Only advanced AI techniques are listed.

Threat	Domain	Analysis Method	AI Paradigms and Models	Key Capabilities	Constraints and Limitations
Malware, Ransomware, Cryptojacking	Malware Detection	Static, Dynamic, Hybrid Analysis	AML, XAI, GPT, BERT, LLM	Pattern recognition, behavior modeling, zero-day detection	Obfuscation sensitivity, dataset bias, adversarial attacks

Table 4. Cont.

Threat	Domain	Analysis Method	AI Paradigms and Models	Key Capabilities	Constraints and Limitations
DDoS, APTs, Insider Threats, XSS	Intrusion Detection	Network Traffic, Temporal Analysis	AML, XAI, GPT, BERT, LLM	Temporal dependency learning, anomaly detection	High computational cost, false positives, real-time constraints, lack of explainability in some models
Phishing, Credential Stuffing, Social Engineering	Phishing and Spam Detection	Content, URL, Semantic Analysis	ML, XAI, DNN, BERT	Fast classification, semantic understanding	Limited generalization, adversarial text manipulation
Botnet-based Attacks	Botnet Detection	Traffic Pattern, Graph Analysis	AML, XAI, GPT, BERT, LLM	Distributed pattern detection, coordination analysis	Complex data structures, scalability issues
All Threat Categories	Cyber Forensics	Forensic, Behavioral Analysis	XAI, GPT, BERT, LLM	Interpretability, traceability, legal compliance	Performance trade-offs, high complexity

Table 5. Heatmap-Based Meta-Analysis of AI Techniques Across Cyber Security Domains. ML and DL denote broad learning paradigms; Transformer and GNN denote neural architectures; XAI denotes explainability-oriented methods applied to AI models. High, Medium, and Low indicate the relative prevalence of techniques within the reviewed studies of each domain and are intended to reflect adoption trends rather than comparative performance. Legend: ● High ◐ Medium ○ Low.

Domain	ML	DL	Transformer	GNN	XAI	Temporal Modeling	Real-Time Feasibility
Malware Detection	◐	●	●	○	◐	◐	◐
Intrusion Detection	◐	●	●	◐	◐	●	○
Phishing & Spam Detection	●	◐	◐	○	◐	○	●
Botnet Detection	◐	●	◐	◐	◐	◐	○
Cyber Forensics	◐	●	●	○	◐	○	○

6.1. Performance Trends Across Cyber Security Domains

A consistent trend across studies indicates that the effectiveness of AI techniques depends heavily on the cyber security task and the characteristics of the underlying data. Many reviewed works employ deep learning models, particularly for malware and intrusion detection, where high-dimensional and sequential data are common. CNNs are frequently used for malware classification, especially when binaries are converted into image-like representations, whereas recurrent models such as LSTMs are widely applied in intrusion detection systems due to their ability to capture temporal patterns in network traffic and system logs.

In contrast, traditional ML techniques, including SVM, RF, and DT, remain highly competitive in phishing and spam detection tasks. This is primarily attributed to the structured nature of the feature space and the relatively lower computational requirements, making these models suitable for real-time and resource-constrained environments.

Transformer-based models, including Bidirectional Encoder Representations from Transformers (BERT) and other large language models, are increasingly being adopted in domains involving sequential and textual data, such as phishing detection and log-based intrusion detection. These models demonstrate strong capabilities in capturing contextual relationships and long-range dependencies; however, their computational complexity presents challenges for real-time deployment.

Furthermore, graph-based models, particularly GNNs, are predominantly utilized in botnet detection scenarios, where network entities and their interactions can be naturally represented as graphs. These models are effective in identifying coordinated and distributed attack patterns, as also reflected in the heatmap visualization in Table 5.

6.2. Model Suitability Across Threat Scenarios

The distribution of AI techniques across threat scenarios reveals that model effectiveness is closely tied to the structure of the data and the behavior of the attack. Pattern-based threats, such as malware binaries and phishing emails, are effectively addressed using CNNs and ensemble ML models due to their ability to extract discriminative features from structured inputs. In contrast, sequential and time-dependent attacks, including intrusion detection and network traffic analysis, benefit significantly from recurrent and transformer-based architectures, which are designed to model temporal dependencies and contextual relationships.

For network-based and distributed threats, such as botnet attacks, graph-based approaches provide a natural and effective modeling framework by capturing relationships and communication patterns between entities. Additionally, complex and multi-stage attacks, such as Advanced Persistent Threats (APTs), often require hybrid and ensemble models that combine multiple AI techniques to address their multi-faceted characteristics.

These observations, summarized in Table 4, reinforce that no single AI model is universally optimal across all cyber security domains. Instead, optimal performance is achieved through alignment between the model architecture, data characteristics, and threat behavior.

6.3. Role of Explainable and Emerging AI Techniques

XAI techniques, such as LIME and SHAP, are increasingly incorporated across multiple domains, particularly in malware detection, intrusion detection, and botnet analysis. Their primary role is to enhance model transparency and support trust in AI-driven decisions. However, a key observation is that XAI is predominantly applied as a post-hoc explanation tool rather than being integrated into the model development lifecycle. Very few studies leverage explainability for feature selection, model optimization, or performance improvement.

Emerging AI paradigms, including GAI and LLMs, are gaining attention in cyber security research. These models show promise in applications such as anomaly detection, automated feature generation, and log analysis. Nevertheless, their practical adoption remains limited due to challenges related to computational cost, robustness, and security risks.

Moreover, the integration of XAI with these emerging paradigms remains largely unexplored. The inherent black-box nature of transformer-based and generative models introduces additional challenges in high-stakes domains such as cyber forensics, where interpretability, accountability, and legal admissibility are critical.

6.4. Common Limitations Across Studies

Despite the promising performance reported across domains, several recurring limitations are consistently observed. First, dataset quality and representativeness remain a fundamental concern. A large proportion of studies rely on publicly available benchmark datasets, which are often outdated, imbalanced, or lack representation of zero-day and polymorphic attacks. This raises concerns regarding the generalizability of reported results to real-world environments.

Second, computational overhead is a significant limitation, particularly for DL, transformer-based, and graph-based models. While these approaches achieve high detection accuracy, their resource requirements hinder deployment in real-time and resource-constrained environments.

Third, adversarial robustness is insufficiently addressed. Although some studies incorporate adversarial training or evaluation, many do not consider adversarial threats, potentially leading to overestimation of model reliability.

Fourth, explainability is often treated as an optional enhancement rather than a core requirement. Even when XAI methods are applied, they are rarely used to iteratively improve model performance or reduce complexity.

Finally, reproducibility remains a challenge due to incomplete reporting of experimental configurations, including preprocessing steps, hyperparameters, and evaluation protocols, limiting the ability to compare results across studies.

6.5. Cross-Domain Research Gaps

Synthesizing findings across all domains reveals several critical research gaps. First, there is a lack of standardized benchmarking frameworks, which limits fair comparison across studies. Second, transfer learning and cross-domain adaptation remain underexplored, despite the potential for shared patterns across different cyber security tasks.

Third, real-time and edge-deployable AI solutions are insufficiently addressed, with most studies prioritizing accuracy over efficiency. Fourth, the integration of explainability with emerging AI paradigms, particularly transformers and generative models, is largely absent.

Fifth, adversarial defense mechanisms remain fragmented and domain-specific, with no unified framework addressing robustness across multiple threat categories. Finally, human-AI collaboration in cyber security operations remains an open research area, particularly in understanding how analysts interact with AI-driven insights and how trust is established in operational environments.

6.6. Cross-Domain Synthesis

The meta-analysis reveals three dominant trends across the reviewed literature: (1) a clear shift from traditional ML toward DL and transformer-based models in data-intensive domains such as intrusion detection and malware analysis; (2) the continued relevance of lightweight ML techniques in structured and resource-constrained scenarios such as phishing detection; and (3) the growing importance of explainability and robustness as critical requirements for real-world deployment.

These findings indicate that the effectiveness of AI in cyber security is not determined by the model alone, but by the alignment between data characteristics, threat behavior, and analytical methodology. Future research should therefore focus on developing robust, efficient, and explainable AI systems, supported by standardized evaluation frameworks and validated in real-world environments.

7. AI and Cyber Security

This section reviews previous research conducted in the various key areas of Cyber Security based on the taxonomy developed in Section 5. As mentioned in Section 2, we identified 75 articles for review. Out of these 75, three refer to the related works, and the distribution of the rest of the 72 articles, reviewed in this section, is shown in Figure 5.

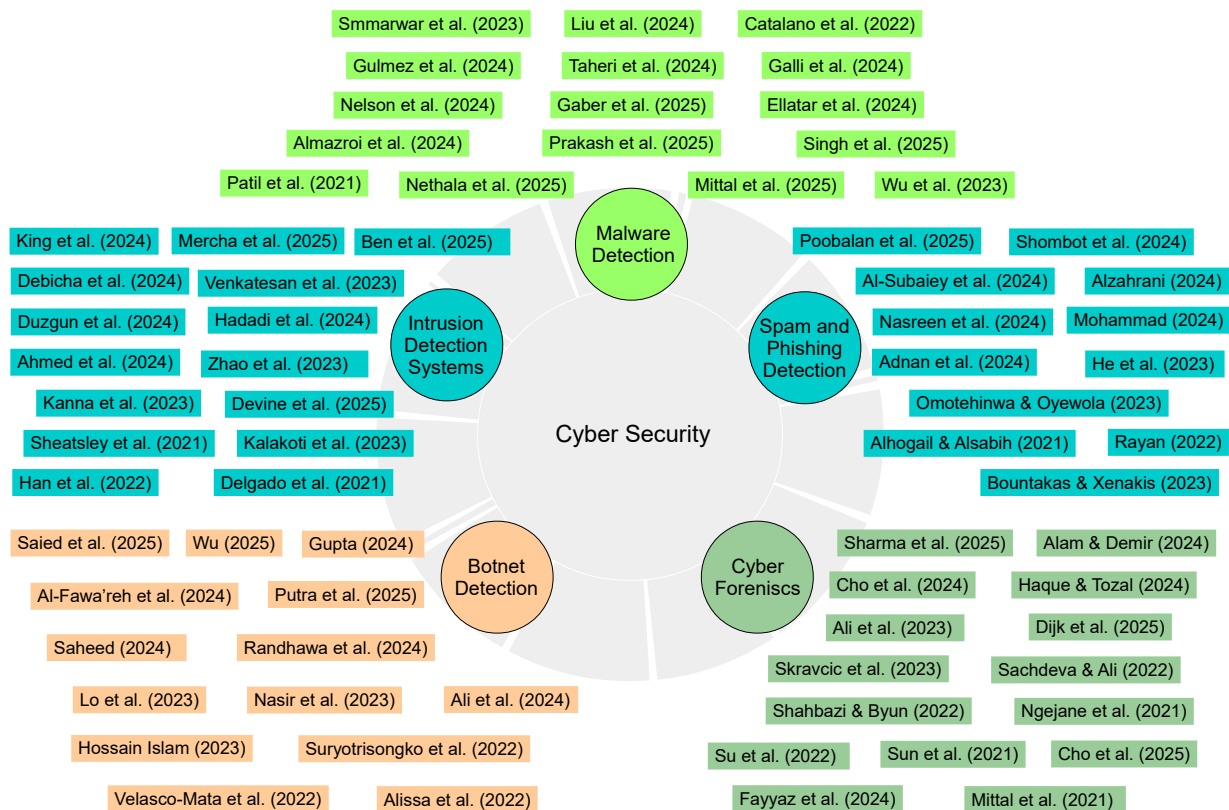


Figure 5. Distribution of the 72 articles [41–112] reviewed in this paper.

7.1. AI and Malware Detection

This section reviews sixteen studies on AI-based malware detection, with emphasis on recent advances (2024–2025) using transformers, deep learning, adversarial learning, and explainable AI.

Ref. [41] proposed a transformer-based approach for zero-day ransomware detection using dynamic binary instrumentation (Pekaboo) to capture assembly-level behavior. Ref. [42] introduced an ensemble CNN framework for Android malware detection, while ref. [43] developed Argus for early detection of fileless malware using runtime memory analysis.

IoT-focused methods include DL ensembles with XAI [44], and adversarially robust detection using A2PM-generated samples [45]. Ref. [46] combined transformer models with feature selection (Random Forest) and optimization techniques for IoT malware detection.

Explainable and interpretable approaches include SHAP-based malicious PDF detection [47], API-call-based behavioral malware detection using LIME, SHAP, and LRP [48], and the xRan ransomware detection framework using sandbox-extracted features [49].

Additional work explores LLM-assisted model generation for malware classifiers [50], and deep learning frameworks combining CNN, LSTM, and wavelet features for IIoT malware detection [51]. Graph-based approaches using function call graphs and graph2vec embeddings are proposed in [52].

Adversarial robustness is studied through CNN vulnerability to polymorphic attacks [53], adversarial training using FGSM [54], and defense mechanisms against functionality-preserving and data poisoning attacks [55,56].

Summary

Table 6 and Figure 6 summarize the reviewed studies, highlighting the use of advanced AI methods such as GPT, BERT, and other LLMs in recent 2024 research. A key research

focus is defending against AML attacks. Many approaches improve malware detection by training on adversarial examples, but these often rely on costly methods like A2PM, limiting real-time use. While effective against some attacks, they remain less reliable against polymorphic threats, reducing accuracy. Although XAI has been used for model interpretation, it has not been widely applied to enhance malware classification performance. One study [57] improved detection and feature reduction using XAI, but only in cyber forensics, indicating a gap in malware detection research. Studies use diverse datasets due to the lack of a standard benchmark, making comparisons difficult. The rapid emergence of new malware variants further complicates the creation of a universally accepted evaluation dataset.

Table 6. Summary of the research works discussed in Section 7.1 (AI and Malware Detection).

Article	AI Models	XAI	Datasets	AML	Key Features	Research Gaps
[41]	BERT; XLNet; GPT-2	–	Peekaboo	–	Ransomware detection; Binary instrumentation; Assembly embeddings.	Prone to obfuscations; Compute-intensive; May not be suitable for real-time.
[42]	Ensemble-CNNs	–	Drebin; AMD	–	Ensemble of CNNs; 2D images; Feature fusion	Compute-intensive; May not be suitable for real-time.
[43]	BERT; MLP; Llama	–	TRAM; MITRE; Others	–	Fileless malware detection; Llama model for feature generation.	Fine-tuning computational resources. Difficult to explain predictions.
[44]	CNN; LSTM; ANN; DT	LIME	IoT-23	–	IoT malware detection; Ensemble model; LIME to explain predictions.	Only used one XAI technique; Compute-intensive; May not be suitable for real-time.
[45]	A2PM; Deep RCNet; BiRAE	–	IoT-23; Bot-IoT	Yes	IoT malware detection; Provide defense against adversarial attacks; Adversarial training.	Compute-intensive; May not be suitable for real-time.
[47]	SMOTE; LightGBM	SHAP	Evasive-PDFMal2022	–	Malicious PDF detection; Tree-based ensemble models. SHAP for explaining.	Only specific file types (PDF). May miss embedded malicious code.
[50]	ChatGPT; CNN; Other ML models	–	CIC-AndMal2020	–	Android malware detection; ChatGPT to generate AI models' code.	Need some manual modification of the code; Dependent on a high-quality dataset.
[48]	LSTM; GRU	LIME SHAP LRP	Mal-API-2019; Alibaba Cloud	–	Windows malware detection; Validation of explanations of LSTM and GRU models.	Explanations not used to improve performance; Not tested against adversarial attacks.
[55]	CNN	–	BODMA	Yes	Provide defense against adversarial attacks; Adversarial training.	Dependent on meta info; No defense against certain program changes.
[56]	DNN	–	Drebin; Contagio; Genome	Yes	Provide defense against adversarial attacks; Adversarial training.	Poor accuracy against some adversarial attacks; Compute-intensive.
[49]	CNN	LIME SHAP	VirusShare; Sorel-20M; VX Heaven; ISOT-Ransomware	–	Ransomware detection; API calls; Mutexes; Detailed analysis of explanations;	Explanations not used for optimizing (such as reducing the number of features and time, etc.) classification;
[46]	LLM; SHO	SHAP	IoT-23	–	IoT malware detection; SHO to optimize LLM; SHAP for model comprehension.	Difficulty in constrained environments; Easy to manipulate.
[51]	CNN; LSTM	–	Microsoft BIG-2015; Maling	–	IoT malware detection; D3WT for feature extraction	D3WT is compute intensive; May not be suitable for real-time.
[52]	MLP; RF; KNN; SVM; LR	–	VirusTotal	–	IoT malware detection; FCG for feature extraction; graph2vec	May not detect obfuscated IoT malware because of using FCGs.
[53]	CNN	–	VirusShare	Yes	Polymorphic attacks; Adversarial training.	Poor accuracy (defense) against polymorphic attacks.
[54]	CNN	–	MaleVis	Yes	Defense against adversarial attacks; Adversarial training.	Only tested against one type of adversarial attack, FGSM.

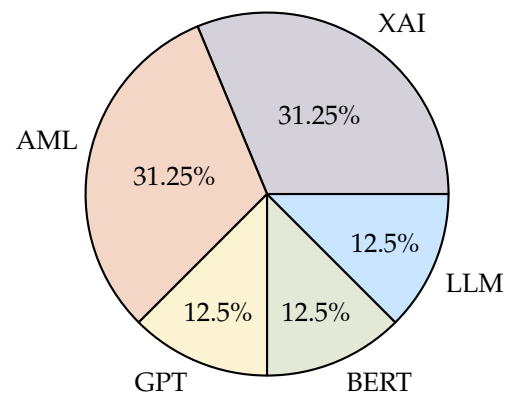


Figure 6. Distribution of the works using advanced AI techniques for malware detection, also summarized in Table 6.

7.2. AI and Intrusion Detection

Transformer-based IDS [58,59] use self-attention to capture long-range dependencies in network traffic, improving feature learning, but often lack real-time evaluation, scalability analysis, and full multi-class coverage. Hybrid and explainable models, including LSTM ensembles [60], CNN-LSTM [61], and multimodal fusion [62], achieve strong performance yet typically omit Transformer integration and real-world deployment validation.

Adversarial studies [63–65] highlight IDS vulnerabilities but rarely offer unified defenses or end-to-end integration. Graph-based methods [66] improve structural detection but lack robustness and real-time applicability. GAN-based approaches [67] address class imbalance via synthetic data but do not support streaming or Transformer-based architectures.

Explainable AI methods [68–70] enhance interpretability but remain limited in operational deployment and adversarial robustness. Federated learning [71] enables privacy-preserving IDS but is not tailored for UAVs or adversarial settings. LLM-based approaches [72] improve log analysis under noise but face scalability and network-level limitations.

Summary

Table 7 and Figure 7 summarize the reviewed works, highlighting applied AI models, use of XAI and AML, utilized datasets, and key features with research gaps. Breakthroughs include transformer-based models (Transformer, GPT, and LLMs) and GNNs, offering enhanced detection of complex and zero-day attacks. XAI applications are gaining traction for improving interpretability but remain underused for performance optimization. Adversarial defenses primarily involve adversarial training with relatively simple attack methods; robustness against more sophisticated evasion and poisoning attacks is still limited. Dataset diversity and standardization remain challenging, hindering reproducibility and fair benchmarking.

Table 7. Summary of the research works discussed in Section 7.2 (AI and Intrusion Detection).

Article	AI Models	XAI	Datasets	AML	Key Features	Research Gaps
[58]	Transformer	–	UNSW-NB15; NSL-KDD	–	No feature selection requirements; self-attention for feature importance.	Lack of multi-class attack classification; absence of hybrid architectures.
[60]	Ensemble LSTM	SHAP	UNSW-NB15; NSL-KDD	–	Explainable ensemble of multiple LSTM models; meta-learner based decision fusion; temporal learning.	Lack of transformer-based learning; no real-time deployment analysis; absence of adversarial robustness testing.

Table 7. Cont.

Article	AI Models	XAI	Datasets	AML	Key Features	Research Gaps
[63]	DNN	–	NSL-KDD	Yes	Adversarial training; improved robustness.	Lack of multi-class intrusion analysis; computational overhead not fully evaluated.
[66]	GNN	–	LANL Logs	–	Temporal graph-based IDS; scalable distributed graph processing;	No adversarial robustness analysis; real-time deployment complexity.
[62]	MLP; Transformer	–	CIC-IDS2017; UNSWNB15	–	Dual-modal learning framework; DL architecture for fusion; improved representation learning.	Limited real-time evaluation; fusion strategy is relatively simple; scalability concerns; lack of XAI.
[72]	LLM; ML Models	–	LOGEVOL; SynHDFS; SYNEVOL	–	LLM-enhanced log understanding; hybrid LLM + ML pipeline; robustness to noisy data.	Scalability concerns with LLM usage; no real-time streaming evaluation; limited focus on network traffic.
[68]	Traditional ML Models	SHAP LIME	CIC-IDS2017	–	Hybrid ensemble IDS; feature fusion and selection strategy.	Limited dataset diversity; limited real-time deployment analysis.
[67]	GAN	–	CIC-IDS2017	–	Multiple GAN architectures; synthetic attack traffic generation; enhancement of minority attack detection.	Lack of transformer-based architectures; limited real-time deployment analysis.
[61]	CNN; LSTM	–	NSL-KDD; ISCX-IDS; UNSWNB15	–	Automated feature extraction; deep hierarchical feature representation. multi-class IDS.	No XAI integration; limited real-time deployment analysis; scalability on large data not fully explored.
[69]	LSTM	SHAP LIME IG	SOC of TalTech	–	Comparative study of XAI methods; improved transparency in DL IDS; practical SOC-oriented evaluation.	Limited real-time deployment analysis; no adversarial robustness analysis.
[65]	Traditional ML Models	–	NSL-KDD	Yes	Defense against poisoning attacks; evaluation of multiple ML models.	Lack of DL models; limited dataset diversity.
[59]	Time-Aware Transformer	–	ISCX2012; CIC-IDS2017	–	N-gram based feature extraction; Attention mechanism for feature importance.	No real-time streaming evaluation; no advanced ensemble learning; no adversarial robustness analysis.
[71]	Federated Learning	–	CIC-IoT2023	–	Global model aggregation; privacy-preserving IDS.	Limited UAV and edge mobility consideration; no real-time evaluation.
[70]	CNN	Grad-CAM	NSL-KDD; UNSWNB15	–	Improving accuracy via visualization feedback; heatmap interpretation.	No adversarial robustness evaluation; limited real-time evaluation.
[64]	JSMA; HSG	–	NSL-KDD; MNIST	Yes	Focus on adversarial vulnerability; domain-constrained attacks.	Lack of defense mechanisms; absence of real-time evaluation.

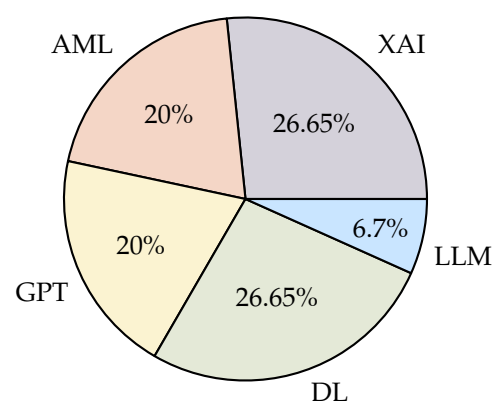


Figure 7. Distribution of works leveraging advanced AI techniques for intrusion detection summarized in Table 7.

7.3. AI and Spam and Phishing Detection

This section reviews twelve studies on AI-based spam and phishing detection, focusing on recent work in both domains.

Ref. [73] developed a high-performance phishing email classifier using a merged dataset of 82,500 emails with extensive preprocessing and applied XAI (LIME) for interpretability. The model was deployed as a real-time web application. Ref. [74] evaluated SVM-based phishing detection using polynomial and RBF kernels. Ref. [75] combined Grey Wolf Optimization for feature selection with BERT embeddings and DL models (LSTM, BiLSTM) for spam detection.

Ref. [76] proposed an explainable stacking ensemble model using SVM, LR, Gradient Boosting, and DT with RF as a meta-classifier, supported by SHAP for interpretability. Ref. [77] introduced a lifelong learning-based spam classifier achieving high adaptability, while ref. [78] used stacking ensembles combining multiple ML models for improved accuracy.

For phishing websites, ref. [79] proposed a Tiny-BERT-based stacking model that extracts semantic features from URLs using a lightweight transformer. Ref. [80,81] focused on ensemble and hybrid feature-based spam detection approaches.

Earlier work includes a hybrid bagging model combining DT and RF with feature selection via CFS [82], and GCN-based phishing detection using NLP and graph modeling of email corpora [83], which achieved high accuracy by transforming text classification into node classification.

Summary

Table 8 and Figure 8 summarize research on AI-based spam and phishing detection, highlighting the use of Tiny-BERT, GCN, and ensemble methods, which achieve high accuracy with low false positives. XAI techniques like LIME are used to improve transparency and user trust, though most models still emphasize performance over explainability and real-time deployment. Despite strong accuracy, challenges persist, including poor generalization to new threats, reliance on static datasets, high computational costs, and limited real-time integration, leaving key gaps for future research.

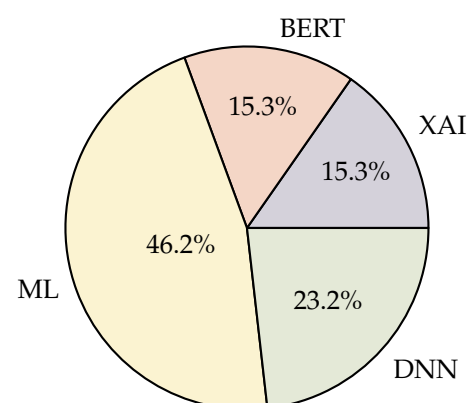


Figure 8. Distribution of works leveraging advanced AI techniques for spam and phishing detection summarized in Table 8.

Table 8. Summary of the research works discussed in Section 7.3 (AI and Spam and Phishing).

Article	AI Models	XAI	Datasets	AML	Key Features	Research Gaps
[84]	DNN-BiLSTM	–	ENRON	–	TF-IDF; DNN-BiLSTM	Lack dataset diversity; Compute intensive

Table 8. Cont.

Article	AI Models	XAI	Datasets	AML	Key Features	Research Gaps
[74]	SVM	–	Website Phishing	–	Allow users to input features for classification	Limited to website phishing
[73]	SVM; MNB; RF	LIME	Six public datasets	–	TF-IDF; Web deployment via Flask with user feedback	DL to improve; Lack of scalability
[77]	NB; SVM	–	ENRON	–	Early Drift Detection for dataset partitioning	Dependency on NB; Missing virtual concept drift test
[78]	LR; DT; GNB; KNN; AB	–	Spam-Assassin; ENRON	–	Stacking ensemble method	Vulnerable to evolving spam technique
[75]	GWO-BERT; DNN	–	Ling-Spam	–	GWO for feature selection with BERT embeddings	Outdated dataset; May fail to detect modern spam; Compute intensive
[76]	SVM; LR; GB; DT; RF	SHAP	SMS Spam Collection	–	Ensemble learning; XAI for interpreting results	Requires updates for new spam patterns
[79]	BERT	–	Phishing Site URLs	–	Tiny-BERT stacking; high accuracy and stability	Potential biases or imbalances in the dataset.
[80]	RF; XGBoost	–	ENRON	–	RF; XGBoost; Optimal tuning of hyperparameters	Lack dataset diversity; Compute intensive
[81]	DT; KNN; MLP	–	ENRON; SpamAssassin	–	Ensemble learning with hybrid features; Stacking; Voting	Highly imbalanced data; Focuses on phishing and not generic spam
[82]	DT; RF	–	Ling Spam	–	Correlation analysis; hybrid bagging	Poor adaptive learning and robustness to adversarial conditions
[83]	GCN; NLP	–	CLAIR Fraud Email	–	TF-IDF; text to a node classification problem; Graph input to a GCN	Lack dataset diversity; Compute intensive

7.4. AI and Botnet Detection

AI-driven botnet detection has attracted extensive research, with many studies proposing advanced ML, DL, graph-based, and explainable approaches. This section reviews fourteen selected works, highlighting key contributions and limitations.

Ref. [85] proposed a privacy-preserving federated learning framework that removes sensitive information while retaining botnet-relevant features using adversarial learning. Ref. [86] introduced a graph-based forensic model with adaptive edge weighting and an ensemble of ML classifiers. Ref. [87] enhanced RF-based IoT botnet detection using XAI techniques such as LIME and SHAP.

DL approaches include LSTM-Autoencoder and MLP-based hybrid models for IoT botnet detection [88], and a multi-stage framework using RFE, LightGBM, and RF with SHAP interpretability [89]. Ref. [90] combined dimensionality reduction, XAI, and CNN transfer learning optimization, while ref. [91] improved robustness using GAN-based adversarial training.

Efficient and lightweight methods include a real-time Decision Tree model using minimal traffic features [92] and hybrid feature selection with Extra Trees classifiers [93]. Ref. [94] evaluated signature-based IDS tools for IoT botnet detection across multiple datasets. Graph-based methods such as XG-BoT integrate GNNs with explainability tools [95]. Additional studies address IoT botnet detection using ML [96] and DGA botnet detection using RF with XAI and OSINT-based threat intelligence sharing [97].

Summary

Table 9 and Figure 9 summarize the reviewed studies, highlighting the use of advanced AI techniques such as GPT, BERT, and other LLMs. AI has shown strong effectiveness in botnet detection, particularly in IoT security and network analysis. Traditional ML performs well in specific cases, while deep learning—especially hybrid models and GNNs—handles complex data and advanced attacks. Emerging methods like reinforcement learning, federated learning, and GANs are also being explored, with XAI improving transparency

and trust. However, challenges remain, including limited data for new threats, trade-offs between interpretability and performance, high computational costs, and evolving adversarial attacks. Future work should focus on scalable, adaptive, and explainable systems, along with stronger defenses and improved threat intelligence sharing.

Table 9. Summary of the research works discussed in Section 7.4 (AI and Botnet Detection).

Article	AI Models	XAI	Datasets	AML	Key Features	Research Gaps
[87]	Random Forest	SHAP	N-BaIoT (IoT)	Yes	XAI for RF-based IoT botnet detection; Enhanced interpretability and trust.	Lack of balance between interpretability and security, e.g., XAI revealing features.
[85]	CNN-BiLSTM; FL	–	ISCX-2014; CTU-13	Yes	Privacy-enhanced framework; Mutual info for optimization.	High complexity; potential performance trade-off for privacy.
[86]	Ensemble ML (DT, RF, etc.)	–	CTU-13; NCC-2	–	Graph-based classification for forensics; Adaptive edge weighting;	Lack of defense against low-intensity and simultaneous attacks.
[88]	LSTM; MLP	–	N-BaIoT; UNSW-NB15	–	Focus on zero-day; synergistic fusion of LSTM for sequential analysis.	Lacks zero-day attacks testing; Not very efficient for low-resource devices.
[89]	LightGBM RF	SHAP	NCC-2	–	Multi-layer framework; RFE for feature selection.	Lack of diverse datasets; lack of adaptability to IoT.
[98]	DRL	–	MedB IoT; N-BaIoT	Yes	Attention reward optimization; SoHo platform focus.	Vulnerability to adversarial attacks; DRL convergence is time-consuming.
[90]	Ensemble CNN	SHAP	CAN-Hacking; N-BaIoT	Yes	Barnacle Mating Optimizer for hyperparameter optimization.	Not tested on diverse real-world IoT zero-day datasets.
[91]	GAN; DRL	–	ISCX2014 CICIDS2017 CICIDS2018	Yes	Semantic-aware sample generation; Evasion awareness.	Increased training time compared to base GANs.
[95]	Deep GNN	GNE; SMs	CTU-13	–	Explainable GNN platform for forensics; Addresses GNN over-smoothing.	General challenges of GNN explainability; lack of real-time performance.
[93]	Extra Trees Ensemble	SHAP	N-BaIoT; BoT-IoT; CTU-13; ISCX;	–	Hybrid feature selection pipeline; high accuracy across varied datasets.	No real-time performance analysis; targeted adversarial attacks not tested.
[94]	Signature-based IDS	–	ISOT; IoT23; BoTIoT	–	Custom signature; collaborative framework.	Inherent weakness against zero-day attacks.
[92]	Decision Tree	–	CTU-13	–	Ultra-fast analysis; Minimal feature set for speed.	Accuracy vs. speed trade-off; lack of testing.
[96]	DT; XGBoost; LR	–	UNSW-NB 15	–	SMOTE for class imbalance; feature engineering and selection.	Cross-dataset validation; lack of generalizability.
[97]	RF	An-chor LIME SHAP	Logs; Alexa; Netlab 360 DGA	Yes	DGA botnet detection platform; Blends XAI with OSINT for CTI sharing.	High time for some features; limited robustness against MaskDGA attack.

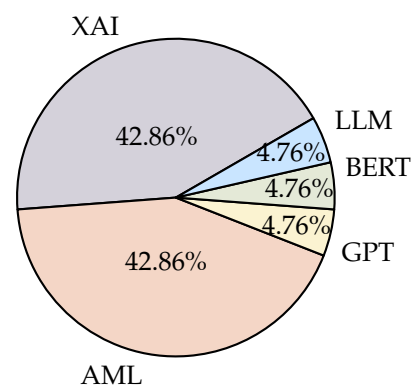


Figure 9. Distribution of the works using advanced AI techniques for Botnet detection, also summarized in Table 9.

7.5. AI and Cyber Forensics

AI significantly enhances cyber forensics by accelerating data analysis, anomaly detection, and large-scale search, enabling faster and more efficient investigations as data volumes increase. Although much of the existing work focuses on image forensics [113], AI also shows strong potential in network forensics. It reduces investigation time, supports learning from past cases, and helps analysts adapt to evolving cyber threats. This section reviews sixteen selected studies, summarizing their key contributions and limitations.

Early work by [99] introduced a Digital Forensic Process Model (DFPM) using ML to classify predatory chats, emphasizing interpretability for legal use. Ref. [100] developed FiFTy for file type identification from memory fragments, while ref. [101] proposed an NLP-based platform for analyzing digital communications, though limited in encrypted environments. In 2022, ref. [36] explored XAI for assisting forensic analysts in handling large datasets and generating explainable reports. Other contributions include DDoS detection using ML [102], NLP-blockchain integration for social media forensics [103], and the MADIK multi-agent system for automated investigation tasks [114]. Additional studies proposed Byte2Vec for file fragment classification [104] and CQSS-GA-SVM for audio forgery detection [105].

In 2023, ByteRCNN combined CNN and RNN models for file type identification [106], while MHRI improved JPEG image recovery [107]. More recent work includes SIFT, which applies XAI for feature selection in file fragment classification [57], hybrid forensic analysis of vehicle systems using ML and LLMs [108], and LLM-based author profiling [109]. In 2025, research has expanded to forensic frameworks for conversational AI systems [110], dataset design principles for Cyber Security [115], password recovery tools [111], LLM-based forensic assistants [112], and broader integration of LLMs across forensic workflows [116].

Overall, these studies demonstrate growing adoption of AI and LLMs in cyber forensics, but also highlight ongoing challenges in robustness, interpretability, and real-world deployment.

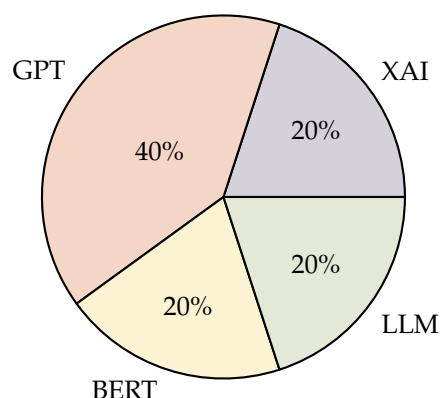
Unlike other cyber security domains that primarily emphasize detection accuracy, cyber forensics imposes additional requirements related to interpretability, auditability, chain of custody, and legal admissibility of evidence. Consequently, forensic AI systems must not only produce accurate results but also provide transparent and reproducible reasoning that can withstand legal and investigative scrutiny. While several reviewed studies explore explainable AI and forensic process frameworks, limited empirical evidence currently exists regarding the legal admissibility, auditability, and operational deployment of AI-assisted forensic findings.

Summary

Table 10 and Figure 10 summarize the reviewed studies. In cyber forensics, AI automates tasks such as data analysis, pattern recognition, and decision support. Earlier work focused on file identification and reconstruction, while recent studies use LLMs for advanced tasks like author profiling, password discovery, and classification, with some exploring full-process applications. XAI methods such as SHAP and LIME improve transparency and interpretability. Key challenges include the lack of high-quality, diverse datasets, with existing data often outdated, biased, and limited by privacy concerns. Additionally, issues such as LLM bias, hallucinations, limited interpretability, auditability requirements, and questions regarding legal admissibility may affect the reliability and practical adoption of AI-assisted forensic systems.

Table 10. Summary of the research works discussed in Section 7.5 (AI and Cyber Forensics).

Article	AI Models	XAI	Datasets	AML	Key Features	Research Gaps
[112]	GPT-4 LLaMA	–	Custom (10,000 Q&A)	–	ForensicLLM; provides offline services	Prone to hallucination; GPT bias
[111]	PassGPT DistilBERT CodeBERT XGBoost	–	Custom (RockYou, MyHeritage, carved data)	–	Able to find human-like passwords; multi-model fusion	Low generalization; slow inference
[110]	ChatGPT Gemini Copilot Claude	–	–	–	A framework to forensically investigate conversational AI services	Tool set is not able to retrieve deleted messages or artifacts when export or cache access is restricted
[109]	Polyglot EEVE Blossom	–	NIKL Korean Corpus	–	Empirically evaluate using different fine-tuning methods; Effective and scalable polyglot-based models.	Possible bias in data
[108]	K-means GPT-4 BERT DAVINCI ChatGPT	–	Hyundai & Mitsubishi infotainment images	–	A hybrid model that combines clustering and language models had the best results	Some false positives, including hallucinations
[57]	MLP	SHAP LIME	[117]	–	File fragment classification; TPR 10% better than others	Some file fragments with embedded fragments might not be classified properly
[107]	Meta-Heuristic and GA	–	DFRWS2006, DFRWS2007 Private	–	MetaHeuristic Reassemble Images for forensic recovery of Bifragmented images	Cannot handle images with missing or non-sequential fragments
[106]	ByteRCNN	–	FiFTy	–	A competitive file fragment classification model	Decreased accuracy for diverse or compressed file formats.
[105]	Custom GA and SVM	–	Real-world CMFD	–	Audio copy-move forgeries detection	Handling high-dimensional features may increase processing time
[104]	Byte2Vec kNN	–	GovDocs1; custom	–	Better handling of high-entropy data	Lower performance on similar/compressed files; limited to graphics files
[103]	LDA; RF	–	Real-world OSN data	–	NLP and blockchain for investigations	Lack of generalization to other forensic domains or multilingual datasets
[102]	MLP	–	NSL-KDD	–	Performs attack classification in cloud network environment with high accuracy	Needs labeled data
[101]	LDA; CFR; RF	–	ENRON	–	Investigation platform for online communication	Needs labeled data; limited to clear-text
[100]	1-D CNN	–	FFT-75; GovDocs; DFRWS2006	–	Open-source tool for file-type identification using memory forensics and data carving	Performance may not generalize across domains/languages
[99]	LR; XGBoost; MLP; BiLSTM	–	PAN-12	–	Investigation process when investigating online predatory chats	Lack of generalization; data not public; imbalanced data

**Figure 10.** Distribution of the works using advanced AI techniques for cyber forensics, also summarized in Table 10.

8. Challenges

The integration of AI into Cyber Security offers significant benefits but also introduces key challenges that affect performance, reliability, and ethical deployment. Addressing these issues is essential for effective AI-driven security systems.

The challenges discussed in this section are derived from both the explicit limitations reported in the reviewed studies and the cross-domain synthesis conducted in this survey. Where appropriate, broader observations are inferred from recurring trends identified across multiple cyber security domains.

8.1. Availability and Quality of Data

AI models require large, high-quality datasets for training and evaluation. However, the field lacks a standard benchmark dataset, making cross-study comparison difficult. The continuous emergence of new malware further complicates dataset standardization.

A major issue is data scarcity, particularly for rare or emerging threats such as zero-day attacks. In addition, dataset bias can lead to models that perform poorly in real-world scenarios due to limited attack diversity. Privacy regulations further restrict data collection, reducing dataset size and diversity and creating a trade-off between model performance and compliance [118].

8.2. Real-Time Processing (Cost)

Cyber threats require real-time detection, but many AI models are computationally expensive. This limits their use in resource-constrained environments such as IoT systems. High processing latency can delay threat response in large-scale networks. While lightweight models and edge computing are being explored, achieving a balance between efficiency and accuracy remains an open challenge.

8.3. Adversarial Attacks

AI systems are vulnerable to adversarial attacks where inputs are subtly modified to evade detection. For example, malware can be altered to bypass security systems while remaining malicious. This can lead to missed detections or false alarms [119]. Adversarial training is widely used to improve robustness, but remains an active research area.

8.4. Explainability and Interpretability

Many AI models operate as black boxes, limiting trust and transparency in security decisions. This is particularly critical in Cyber Security, where explanations are essential for validation and accountability. Deep learning models are especially difficult to interpret.

Explainable AI (XAI) techniques, such as SHAP and LIME, enhance model transparency and trustworthiness [7,37]. However, their use has largely been limited to model interpretation rather than directly improving classification performance through tasks such as feature selection or reduction. Recent studies [57] have demonstrated promising results in cyber forensics, suggesting potential for wider adoption across cyber security applications.

8.5. Ethical Implications (Accountability)

AI in Cyber Security raises ethical concerns related to regulation, accountability, and misuse. Many regions lack clear legal frameworks for AI-based digital investigations, limiting trust in forensic results [120]. Responsibility for AI-driven decisions is often unclear, creating accountability gaps [121].

Additionally, AI's dual-use nature enables both defensive and malicious applications, including phishing and automated attacks, raising significant ethical risks.

8.6. Continuous Adaptation (Evolution)

Cyber threats evolve rapidly, causing concept drift that reduces model effectiveness over time. AI systems must be retrained frequently with updated data, a process that is costly and computationally intensive. Managing continuous updates without performance loss remains a key challenge.

8.7. Evaluation Metrics

Standard metrics such as accuracy, precision, recall, and F1-score are widely used but insufficient in Cyber Security contexts. Imbalanced datasets and unequal costs of false positives and false negatives require more specialized evaluation strategies.

9. Future Works

Recent research has increasingly focused on advanced AI techniques such as XAI, adversarial machine learning (AML), and generative models, including GPT, BERT, and other LLMs. However, traditional ML and DNN approaches are less explored within the scope of this study.

Despite progress, several research gaps remain that require further investigation to improve AI's effectiveness in Cyber Security.

9.1. Adversarial Machine Learning

Defending against adversarial and polymorphic attacks remains a major challenge [53]. Existing systems struggle with robust detection of evolving malware. Improving adversarial training methods is essential [122], along with exploring techniques such as input preprocessing [123], gradient masking [124], and robustness optimization [125].

9.2. Explainable Artificial Intelligence (XAI)

Beyond interpretability, XAI can also enhance model performance by supporting feature selection. Methods such as SHAP and LIME help identify and remove irrelevant features, improving accuracy [49,57]. However, most applications remain limited to malware detection, with potential extensions to intrusion detection, phishing, and spam filtering.

9.3. Interdisciplinary Collaboration

Effective AI-based Cyber Security requires collaboration across disciplines, including AI researchers, security experts, legal professionals, psychologists, and policymakers. Such cooperation is essential for building robust, ethical, and legally compliant systems.

9.4. Cyber Forensics

AI-based cyber forensics, including analysis of LLM-driven systems like ChatGPT and Copilot, is becoming increasingly important [110]. However, challenges such as deepfake detection, lack of standards, and legal uncertainty remain unresolved.

Ensuring transparency through audits, human oversight, and XAI is essential for trustworthiness [126]. Integration of adversarial testing and explainability is critical for forensic reliability.

9.5. Generative Artificial Intelligence (GAI)

GAI introduces both opportunities and risks in Cyber Security [127]. While it improves detection capabilities, it can also be exploited for malicious purposes. Key challenges include a lack of interpretability, trust issues, and limited real-time applicability. Optimizing GAI models for resource-constrained environments such as edge computing remains an important future direction.

9.6. Deployment and Operational Considerations

Despite the strong detection performance reported by many AI-based cyber security solutions, deployment-related factors such as inference latency, hardware requirements, energy consumption, and scalability are often underreported. These considerations are critical for real-time and resource-constrained environments, where operational feasibility is as important as detection accuracy. Future research should therefore emphasize deployment-aware evaluation and integration with operational cyber security infrastructures, including security operation centers.

10. Threats to Validity

This survey focuses on synthesizing and categorizing recent research on AI applications in cyber security rather than quantitatively assessing study quality. Although a formal quality assessment framework was not employed, the reviewed studies were examined with respect to their datasets, evaluation methodologies, performance metrics, and practical applicability. The findings should therefore be interpreted in light of the heterogeneity of datasets and evaluation protocols across the literature.

This study is limited by its reliance on published literature, which may not fully capture the latest AI advancements due to publication delays. Given the rapid evolution of AI, emerging technologies may significantly reshape Cyber Security beyond what is reflected in current research.

The study adopts a qualitative approach, using rich data and pattern analysis to conclude. However, due to the lack of standardized comparison methods, findings should be viewed as indicative rather than definitive. Despite efforts such as triangulation [9,10], some degree of subjectivity may remain [128].

Overall, this framework enables systematic analysis of qualitative data while accounting for diverse contexts and perspectives, supporting deeper and more informed interpretation of the subject matter.

11. Conclusions

This paper presented a comprehensive survey of AI applications in cyber security, covering studies published between 2021 and 2025. A systematic multi-stage methodology was employed to ensure a rigorous and structured analysis of the literature.

A key contribution of this work is the proposed multi-layer taxonomy, which maps cyber threats to application domains, analysis methodologies, AI techniques, and their associated capabilities and limitations. This framework provides a unified and structured perspective on the application of AI across malware detection, intrusion detection, phishing and spam detection, botnet detection, and cyber forensics.

Building on this taxonomy, a cross-domain meta-analysis was conducted to identify recurring trends in the adoption and application of AI techniques. The analysis revealed three dominant patterns: (i) deep learning and transformer-based models are widely adopted in data-intensive and sequential domains, (ii) traditional machine learning techniques remain effective in structured and resource-constrained environments, and (iii) graph-based and hybrid models are increasingly used to address distributed and multi-stage attacks.

Despite these advancements, several challenges remain, including dataset limitations, limited generalization, adversarial vulnerabilities, computational overhead, and a lack of explainability. Furthermore, emerging paradigms such as XAI and GAI introduce both new opportunities and additional challenges related to transparency, robustness, and trustworthy deployment.

The survey also highlights several important research directions, including the development of standardized benchmarking frameworks, cross-domain knowledge transfer,

real-time and edge-deployable AI systems, and deeper integration of explainability into advanced AI models.

Overall, this work provides a unified taxonomy and a cross-domain analytical perspective on AI-driven cyber security. By synthesizing recent advances, challenges, and emerging trends, it offers valuable insights for researchers and practitioners seeking to develop more robust, scalable, and trustworthy cyber security systems.

Author Contributions: Conceptualization, S.A.; methodology, S.A.; funding acquisition, S.A.; writing—original draft preparation, S.A.; formal analysis, A.J., E.A., and I.C.; validation, E.A., S.Q., and A.A.; investigation, A.J., S.Q., and A.A.; resources, S.Q., B.N., and I.C.; data curation, A.J., E.A., S.Q., B.N., A.A., and I.C.; writing—review and editing, S.A., A.J., E.A., Z.P., S.Q., B.N., A.A., and I.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research is funded by the Scientific Research Deanship at the University of Ha'il, Saudi Arabia, under Grant RG-24 178.

Data Availability Statement: The original contributions presented in this study are included in the article material. Further inquiries can be directed to the corresponding author.

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

References

1. Russell, S.J.; Norvig, P. *Artificial Intelligence: A Modern Approach*; Pearson Education: London, UK, 2020.
2. McCarthy, J.; Minsky, M.L.; Rochester, N.; Shannon, C.E. A proposal for the Dartmouth summer research project on artificial intelligence, August 31, 1955. *AI Mag.* **2006**, *27*, 12.
3. Ali, S.; Wang, J.; Leung, V.C.M. AI-driven fusion with cybersecurity: Exploring current trends, advanced techniques, future directions, and policy implications for evolving paradigms—A comprehensive review. *Inf. Fusion* **2025**, *118*, 102922.
4. Ozkan-Okay, M.; Akin, E.; Aslan, Ö.; Kosunalp, S.; Iliev, T.; Stoyanov, I.; Beloev, I. A comprehensive survey: Evaluating the efficiency of artificial intelligence and machine learning techniques on cyber security solutions. *IEEE Access* **2024**, *12*, 12229–12256.
5. Salem, A.H.; Azzam, S.M.; Emam, O.; Abohany, A.A. Advancing cybersecurity: A comprehensive review of AI-driven detection techniques. *J. Big Data* **2024**, *11*, 105.
6. Sai, S.; Yashvardhan, U.; Chamola, V.; Sikdar, B. Generative AI for cyber security: Analyzing the potential of chatgpt, dall-e and other models for enhancing the security space. *IEEE Access* **2024**, *12*, 53497–53516.
7. Capuano, N.; Fenza, G.; Loia, V.; Stanzione, C. Explainable artificial intelligence in cybersecurity: A survey. *IEEE Access* **2022**, *10*, 93575–93600.
8. Google. Explore What the World Is Searching. 2026. Available online: <https://trends.google.com> (accessed on 5 May 2026).
9. Yin, R.K. *Qualitative Research from Start to Finish*; Guilford Publications: New York, NY, USA, 2015.
10. Patton, M.Q. *Qualitative Research & Evaluation Methods: Integrating Theory and Practice*; Sage Publications: Thousand Oaks, CA, USA, 2014.
11. Liu, H.; Wang, Y.; Fan, W.; Liu, X.; Li, Y.; Jain, S.; Liu, Y.; Jain, A.K.; Tang, J. Trustworthy AI: A Computational Perspective. *ACM Trans. Intell. Syst. Technol.* **2022**, *14*, 1–59. <https://doi.org/10.1145/3546872>.
12. Zhang, H.; Wu, B.; Yuan, X.; Pan, S.; Tong, H.; Pei, J. Trustworthy Graph Neural Networks: Aspects, Methods, and Trends. *Proc. IEEE* **2024**, *112*, 97–139.
13. Nilsson, N.J. *Principles of Artificial Intelligence*; Morgan Kaufmann: Burlington, MA, USA, 2014.
14. Kaplan, A.; Haenlein, M. Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Bus. Horizons* **2019**, *62*, 15–25.
15. Jordan, M.I.; Mitchell, T.M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 255–260.
16. Abdullahi, M.; Baashar, Y.; Alhussian, H.; Alwadain, A.; Aziz, N.; Capretz, L.F.; Abdulkadir, S.J. Detecting Cybersecurity Attacks in Internet of Things Using Artificial Intelligence Methods: A Systematic Literature Review. *Electronics* **2022**, *11*, 138.
17. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444.
18. Dixit, P.; Silakari, S. Deep Learning Algorithms for Cybersecurity Applications: A Technological and Status Review. *Comput. Sci. Rev.* **2021**, *39*, 100317.
19. Papernot, N.; McDaniel, P.; Goodfellow, I.; Jha, S.; Celik, Z.B.; Swami, A. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*; ACM: New York, NY, USA, 2017; pp. 506–519.

20. Carlini, N.; Wagner, D. Towards evaluating the robustness of neural networks. In *Proceedings of the 2017 IEEE Symposium on Security and Privacy (sp)*; IEEE: New York, NY, USA, 2017; pp. 39–57.
21. Bommasani, R.; Hudson, D.A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M.S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. On the opportunities and risks of foundation models. *arXiv* **2021**, arXiv:2108.07258. <https://doi.org/10.48550/arXiv.2108.07258>.
22. Weidinger, L.; Mellor, J.; Rauh, M.; Griffin, C.; Uesato, J.; Huang, P.S.; Cheng, M.; Glaese, M.; Balle, B.; Kasirzadeh, A.; et al. Ethical and social risks of harm from Language Models. *arXiv* **2021**, arXiv:2112.04359. <https://doi.org/10.48550/arXiv.2112.04359>.
23. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Benetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115.
24. Taddeo, M.; McCutcheon, T.; Floridi, L. Trusting artificial intelligence in cybersecurity is a double-edged sword. *Nat. Mach. Intell.* **2019**, *1*, 557–560.
25. Gunning, D.; Aha, D.W. DARPA’s Explainable Artificial Intelligence (XAI) Program. *AI Mag.* **2019**, *40*, 44–58.
26. Phillips, P.J.; Hahn, C.A.; Fontana, P.C.; Broniatowski, D.A.; Przybocki, M.A. *Four Principles of Explainable Artificial Intelligence*; Technical Report NISTIR 8312; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2020. <https://doi.org/10.6028/NIST.IR.8312>.
27. Von Solms, R.; Van Niekerk, J. From information security to cyber security. *Comput. Secur.* **2013**, *38*, 97–102.
28. Choo, K.K.R. The cyber threat landscape: Challenges and future research directions. *Comput. Secur.* **2011**, *30*, 719–731.
29. Sarker, I.H.; Kayes, A.S.M.; Badsha, S.; Alqahtani, H.; Watters, P.; Ng, A. Cybersecurity data science: An overview from machine learning perspective. *J. Big Data* **2020**, *7*, 41.
30. Ali, A.; Shah, M.; Foster, M.; Alraja, M. N. Cybercrime resilience in the era of advanced technologies: Evidence from the financial sector of a developing country. *Computers*. **2025**, *14*, 38.
31. ENISA. *ENISA Threat Landscape 2020—List of Top 15 Threats*; Technical Report; European Union Agency for Cybersecurity; ENISA: Attiki, Greece, 2020.
32. Bilge, L.; Dumitras, T. Before we knew it: An empirical study of zero-day attacks in the real world. In *Proceedings of the 2012 ACM Conference on Computer and Communications Security*; ACM: New York, NY, USA, 2012; pp. 833–844.
33. Albert, C. The Top 10 Cybersecurity Threats—From Malware and Phishing to Insider Threats—And Learn Practical Steps to Protect Yourself and Your Business. 2024. Available online: <https://vivattechnology.com/news/top-10-cybersecurity-threats-and-how-to-prevent-them> (accessed on 5 May 2026).
34. Sankaram, M.; Roopesh, M.; Rasetti, S.; Nishat, N. A comprehensive review of artificial intelligence applications in enhancing cybersecurity threat detection and response mechanisms. *Glob. Mainstream J. Bus. Econ. Dev. Proj.* **2024**, *3*, 1–14.
35. Buczak, A.L.; Guven, E. A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Commun. Surv. Tutor.* **2016**, *18*, 1153–1176.
36. Hall, S.W.; Sakzad, A.; Choo, K.K.R. Explainable artificial intelligence for digital forensics. *WIREs Forensic Sci.* **2022**, *4*, e1433.
37. Alam, S.; Altiparmak, Z. XAI-CF—Examining the role of explainable artificial intelligence in cyber forensics. *Eng. Appl. Artif. Intell.* **2026**, *167*, 113892.
38. Oseni, A.; Moustafa, N.; Janicke, H.; Liu, P.; Tari, Z.; Vasilakos, A. Security and Privacy for Artificial Intelligence: Opportunities and Challenges. *arXiv* **2021**, arXiv:2101.01195. <https://doi.org/10.48550/arXiv.2101.01195>.
39. Xin, Y.; Kong, L.; Liu, Z.; Chen, Y.; Li, Y.; Zhu, H.; Gao, M.; Hou, H.; Wang, C. Machine learning and deep learning methods for cybersecurity. *IEEE Access* **2018**, *6*, 35365–35381.
40. Kaur, R.; Gabrijelcic, D.; Klobučar, T. Artificial Intelligence for Cybersecurity: Literature Review and Future Research Directions. *Inf. Fusion* **2023**, *97*, 101804.
41. Gaber, M.; Ahmed, M.; Janicke, H. Zero day ransomware detection with pulse: Function classification with transformer models and assembly language. *Comput. Secur.* **2025**, *148*, 104167.
42. Nethala, S.; Chopra, P.; Kamaluddin, K.; Alam, S.; Alharbi, S.; Alsaffar, M. A Deep Learning-Based Ensemble Framework for Robust Android Malware Detection. *IEEE Access* **2025**, *13*, 46673–46696.
43. Singh, N.; Tripathy, S. Unveiling the veiled: An early stage detection of fileless malware. *Comput. Secur.* **2025**, *150*, 104231.
44. Mittal, S.; Wazid, M.; Singh, D.P.; Das, A.K.; Hossain, M.S. A deep learning ensemble approach for malware detection in Internet of Things utilizing Explainable Artificial Intelligence. *Eng. Appl. Artif. Intell.* **2025**, *139*, 109560.
45. Prakash, S.; Mohapatra, A.K. MA_BiRAE-Malware analysis and detection technique using adversarial learning and deep learning. *J. Inf. Secur. Appl.* **2025**, *90*, 104009.
46. Almazroi, A.A.; Ayub, N. Deep learning hybridization for improved malware detection in smart Internet of Things. *Sci. Rep.* **2024**, *14*, 7838.
47. Elattar, M.; Younes, A.; Gad, I.; Elkabani, I. Explainable AI model for PDFMal detection based on gradient boosting model. *Neural Comput. Appl.* **2024**, *36*, 21607–21622.

48. Galli, A.; La Gatta, V.; Moscato, V.; Postiglione, M.; Sperli, G. Explainability in AI-based behavioral malware detection systems. *Comput. Secur.* **2024**, *141*, 103842.
49. Gulmez, S.; Kakisim, A.G.; Sogukpinar, I. XRan: Explainable deep learning-based ransomware detection using dynamic analysis. *Comput. Secur.* **2024**, *139*, 103703.
50. Nelson, J.; Pavlidis, M.; Fish, A.; Kapetanakis, S.; Polatidis, N. ChatGPT-driven machine learning code generation for android malware detection. *Comput. J.* **2024**, *68*, 331–345.
51. Smmarwar, S.K.; Gupta, G.P.; Kumar, S. AI-empowered malware detection system for industrial internet of things. *Comput. Electr. Eng.* **2023**, *108*, 108731.
52. Wu, C.Y.; Ban, T.; Cheng, S.M.; Takahashi, T.; Inoue, D. IoT malware classification based on reinterpreted function-call graphs. *Comput. Secur.* **2023**, *125*, 103060.
53. Catalano, C.; Chezzi, A.; Angelelli, M.; Tommasi, F. Deceiving AI-based malware detection through polymorphic attacks. *Comput. Ind.* **2022**, *143*, 103751.
54. Patil, S.; Varadarajan, V.; Walimbe, D.; Gulechha, S.; Shenoy, S.; Raina, A.; Kotecha, K. Improving the robustness of ai-based malware detection using adversarial machine learning. *Algorithms* **2021**, *14*, 297.
55. Liu, L.; Kuang, X.; Liu, L.; Zhang, L. Defend against adversarial attacks in malware detection through attack space management. *Comput. Secur.* **2024**, *141*, 103841.
56. Taheri, R.; Shojafar, M.; Arabikhan, F.; Gegov, A. Unveiling vulnerabilities in deep learning-based malware detection: Differential privacy driven adversarial attacks. *Comput. Secur.* **2024**, *146*, 104035.
57. Alam, S.; Demir, A.K. SIFT: Sifting file types—Application of explainable artificial intelligence in cyber forensics. *Cybersecurity* **2024**, *7*, 52.
58. Mercha, E.M.; Chakir, E.M.; Erradi, M. Trans-IDS: A Transformer-Based Intrusion Detection System. In *Proceedings of the 20th International Conference on Security and Cryptography (SECRYPT 2023)*; SCITEPRESS: Setúbal, Portugal, 2023; pp. 402–409. <https://doi.org/10.5220/0012085800003555>.
59. Han, X.; Cui, S.; Liu, S.; Zhang, C.; Jiang, B. Network Intrusion Detection Based on N-gram Frequency and Time-Aware Transformer. *Comput. Secur.* **2023**, *128*, 103171. <https://doi.org/10.1016/j.cose.2023.103171>.
60. Ben Ncir, C.E.; Ben HajKacem, M.A.; Alattas, M. Enhancing Intrusion Detection Performance Using Explainable Ensemble Deep Learning. *PeerJ Comput. Sci.* **2024**, *10*, e2289. <https://doi.org/10.7717/peerj-cs.2289>.
61. Rajesh Kanna, P.; Santhi, P. Unified Deep Learning Approach for Efficient Intrusion Detection System Using Integrated Spatial–Temporal Features. *Knowl.-Based Syst.* **2021**, *226*, 107132. <https://doi.org/10.1016/j.knosys.2021.107132>.
62. Düzgün, B.; Çayır, A.; Ünal, U.; Dağ, H. Network Intrusion Detection System by Learning Jointly from Tabular and Text-Based Features. *Expert Syst.* **2024**, *41*, e13518. <https://doi.org/10.1111/exsy.13518>.
63. Debicha, I.; Debatty, T.; Dricot, J.M.; Mees, W. Adversarial Training for Deep Learning-based Intrusion Detection Systems. In *Proceedings of the Sixteenth International Conference on Systems (ICONS 2021)*; ThinkMind: Coimbatore, India, 2021; pp. 45–49. <https://doi.org/10.48550/arxiv.2104.09852>.
64. Sheatsley, R.; Papernot, N.; Weisman, M.J.; Verma, G.; McDaniel, P. Adversarial Examples for Network Intrusion Detection Systems. *J. Comput. Secur.* **2022**, *30*, 727–752. <https://doi.org/10.3233/JCS-210094>.
65. Venkatesan, S.; Sikka, H.; Izmailov, R.; Chadha, R.; Oprea, A.; De Lucia, M.J. Poisoning Attacks and Data Sanitization Mitigations for Machine Learning Models in Network Intrusion Detection Systems. In *Proceedings of the 2021 IEEE Military Communications Conference (MILCOM)*; IEEE: New York, NY, USA, 2021; pp. 874–879. <https://doi.org/10.1109/MILCOM52596.2021.9652916>.
66. King, I.J.; Huang, H.H. Euler: Detecting Network Lateral Movement via Scalable Temporal Link Prediction. *ACM Trans. Priv. Secur.* **2023**, *26*, 1–29. <https://doi.org/10.1145/3588771>.
67. Zhao, X.; Fok, K.W.; Thing, V.L.L. Enhancing Network Intrusion Detection Performance Using Generative Adversarial Networks. *Comput. Secur.* **2024**, *145*, 104005. <https://doi.org/10.1016/j.cose.2024.104005>.
68. Ahmed, U.; Jiangbin, Z.; Almogren, A.; Khan, S.; Sadiq, M.T.; Altameem, A.U.; Rehman, A. Explainable AI-Based Innovative Hybrid Ensemble Model for Intrusion Detection. *J. Cloud Comput.* **2024**, *13*, 150. <https://doi.org/10.1186/s13677-024-00712-x>.
69. Kalakoti, R.; Vaarandi, R.; Bahşi, H.; Nömm, S. Evaluating Explainable AI for Deep Learning-Based Network Intrusion Detection System Alert Classification. In *Proceedings of the 11th International Conference on Information Systems Security and Privacy (ICISSP 2025)*; SCITEPRESS: Setúbal, Portugal, 2025; pp. 47–58. <https://doi.org/10.5220/0013180700003899>.
70. Caforio, F.P.; Andresini, G.; Vessio, G.; Appice, A.; Malerba, D. Leveraging Grad-CAM to Improve the Accuracy of Network Intrusion Detection Systems. In *Proceedings of the 24th International Conference on Discovery Science (DS 2021)*; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2021; Volume 12986, pp. 385–400. https://doi.org/10.1007/978-3-030-88942-5_30.
71. Devine, M.; Ardakani, S.P.; Al-Khafajiy, M.; James, Y. Federated machine learning to enable intrusion detection systems in IoT networks. *Electronics* **2025**, *14*, 1176.

72. Hadadi, F.; Xu, Q.; Bianculli, D.; Briand, L.C. LLM Meets ML: Data-efficient Anomaly Detection on Unstable Logs. *Acm Trans. Softw. Eng. Methodol.* **2025**, 1–47. <https://doi.org/10.1145/3771283>.
73. Al-Subaiey, A.; Al-Thani, M.; Alam, N.A.; Antora, K.F.; Khandakar, A.; Zaman, S.A.U. Novel interpretable and robust web-based AI platform for phishing email detection. *Comput. Electr. Eng.* **2024**, *120*, 109625.
74. Shombot, E.S.; Dusserre, G.; Bestak, R.; Ahmed, N.B. An application for predicting phishing attacks: A case of implementing a support vector machine learning model. *Cyber Secur. Appl.* **2024**, *2*, 100036. <https://doi.org/10.1016/j.csa.2024.100036>.
75. Nasreen, G.; Khan, M.M.; Younus, M.; Zafar, B.; Hanif, M.K. Email spam detection by deep learning models using novel feature selection technique and BERT. *Egypt. Inform. J.* **2024**, *26*, 100473. <https://doi.org/10.1016/j.eij.2024.100473>.
76. Alzahrani, A. Explainable AI-based Framework for Efficient Detection of Spam from Text using an Enhanced Ensemble Technique. *Eng. Technol. Appl. Sci. Res.* **2024**, *14*, 15596–15601.
77. Mohammad, R.M.A. A lifelong spam emails classification model. *Appl. Comput. Inform.* **2024**, *20*, 35–54. <https://doi.org/10.1016/j.aci.2020.01.002>.
78. Adnan, M.; Imam, M.O.; Javed, M.F.; Murtza, I. Improving spam email classification accuracy using ensemble techniques: A stacking approach. *Int. J. Inf. Secur.* **2024**, *23*, 505–517.
79. He, D.; Lv, X.; Zhu, S.; Chan, S.; Choo, K.K.R. A Method for Detecting Phishing Websites Based on Tiny-Bert Stacking. *IEEE Internet Things J.* **2024**, *11*, 2236–2243.
80. Omotehinwa, T.O.; Oyewola, D.O. Hyperparameter Optimization of Ensemble Models for Spam Email Detection. *Appl. Sci.* **2023**, *13*, 1971.
81. Bountakas, P.; Xenakis, C. HELPHED: Hybrid Ensemble Learning Phishing Email Detection. *J. Netw. Comput. Appl.* **2023**, *210*, 103545.
82. Rayan, A. Analysis of e-mail spam detection using a novel machine learning-based hybrid bagging technique. *Comput. Intell. Neurosci.* **2022**, *2022*, 2500772.
83. Alhogail, A.; Alsabih, A. Applying machine learning and natural language processing to detect phishing email. *Comput. Secur.* **2021**, *110*, 102414.
84. Poobalan, A.; Ganapriya, K.; Kalaivani, K.; Parthiban, K. A novel and secured email classification using deep neural network with bidirectional long short-term memory. *Comput. Speech Lang.* **2025**, *89*, 101667. <https://doi.org/10.1016/j.csl.2024.101667>.
85. Wu, G.; Wang, X. A privacy-enhanced framework with deep learning for botnet detection. *Cybersecurity* **2025**, *8*, 9.
86. Putra, M.A.R.; Ahmad, T.; Hostiadi, D.P.; Ijtihadie, R.M. Ensemble Network Graph-Based Classification for Botnet Detection Using Adaptive Weighting and Feature Extraction. *IEEE Access* **2025**, *13*, 31183–31204.
87. Saied, M.; Guirguis, S. Explainable artificial intelligence for botnet detection in internet of things. *Sci. Rep.* **2025**, *15*, 7632. <https://doi.org/10.1038/s41598-025-90420-6>.
88. Ali, S.; Ghazal, R.; Qadeer, N.; Saidani, O.; Alhayan, F.; Masood, A.; Saleem, R.; Khan, M.A.; Gupta, D. A novel approach of botnet detection using hybrid deep learning for enhancing security in IoT networks. *Alex. Eng. J.* **2024**, *103*, 88–97. <https://doi.org/10.1016/j.aej.2024.05.113>.
89. Gupta, S.; Singh, B. An intelligent multi-layer framework with SHAP integration for botnet detection and classification. *Comput. Secur.* **2024**, *140*, 103783.
90. Saheed, Y.K.; Chukwuere, J.E. XAIEnsembleTL-IoV: A new explainable Artificial Intelligence ensemble transfer learning for zero-day botnet attack detection in the Internet of Vehicles. *Results Eng.* **2024**, *24*, 103171.
91. Randhawa, R.H.; Aslam, N.; Alauthman, M.; Khalid, M.; Rafiq, H. RELEVAGAN: Deep reinforcement learning based Evasion Generative Adversarial Network for botnet detection. *Future Gener. Comput. Syst.* **2024**, *150*, 294–302. <https://doi.org/10.1016/j.future.2023.09.011>.
92. Velasco-Mata, J.; González-Castro, V.; Fidalgo, E.; Alegre, E. Real-time botnet detection on large network bandwidths using machine learning. *Sci. Rep.* **2023**, *13*, 4282. <https://doi.org/10.1038/s41598-023-31260-0>.
93. Hossain, M.A.; Islam, M.S. A novel hybrid feature selection and ensemble-based machine learning approach for botnet detection. *Sci. Rep.* **2023**, *13*, 21207. <https://doi.org/10.1038/s41598-023-48230-1>.
94. Nasir, M.H.; Arshad, J.; Khan, M.M. Collaborative device-level botnet detection for internet of things. *Comput. Secur.* **2023**, *129*, 103172. <https://doi.org/10.1016/j.cose.2023.103172>.
95. Lo, W.W.; Kulatilleke, G.; Sarhan, M.; Layeghy, S.; Portmann, M. XG-BoT: An explainable deep graph neural network for botnet detection and forensics. *Internet Things* **2023**, *22*, 100747.
96. Alissa, K.; Alyas, T.; Zafar, K.; Abbas, Q.; Tabassum, N.; Sakib, S. Botnet attack detection in IoT using machine learning. *Comput. Intell. Neurosci.* **2022**, *2022*, 4515642.
97. Suryotrisongko, H.; Musashi, Y.; Tsuneda, A.; Sugitani, K. Robust Botnet DGA Detection: Blending XAI and OSINT for Cyber Threat Intelligence Sharing. *IEEE Access* **2022**, *10*, 34613–34624. <https://doi.org/10.1109/ACCESS.2022.3162588>.
98. Al-Fawa'reh, M.; Abu-Khalaf, J.; Szweczyk, P.; Kang, J.J. MalBoT-DRL: Malware Botnet Detection Using Deep Reinforcement Learning in IoT Networks. *IEEE Internet Things J.* **2024**, *11*, 9610–9629. <https://doi.org/10.1109/JIOT.2023.3324053>.

99. Ngejane, C.; Eloff, J.; Sefara, T.; Marivate, V. Digital forensics supported by machine learning for the detection of online sexual predatory chats. *Forensic Sci. Int. Digit. Investig.* **2021**, *36*, 301109. <https://doi.org/10.1016/j.fsidi.2021.301109>.
100. Mittal, G.; Korus, P.; Memon, N. FiFTy: Large-Scale File Fragment Type Identification Using Convolutional Neural Networks. *IEEE Trans. Inf. Forensics Secur.* **2021**, *16*, 28–41. <https://doi.org/10.1109/TIFS.2020.3004266>.
101. Sun, D.; Zhang, X.; Choo, K.K.R.; Hu, L.; Wang, F. NLP-based digital forensic investigation platform for online communications. *Comput. Secur.* **2021**, *104*, 102210. <https://doi.org/10.1016/j.cose.2021.102210>.
102. Sachdeva, S.; Ali, A. Machine learning with digital forensics for attack classification in cloud network environment. *Int. J. Syst. Assur. Eng. Manag.* **2022**, *13*, 156–165.
103. Shahbazi, Z.; Byun, Y.C. NLP-based digital forensic analysis for online social network based on system security. *Int. J. Environ. Res. Public Health* **2022**, *19*, 7027.
104. Haque, M.E.; Tozal, M.E. Byte embeddings for file fragment classification. *Future Gener. Comput. Syst.* **2022**, *127*, 448–461.
105. Su, Z.; Li, M.; Zhang, G.; Wu, Q.; Li, M.; Zhang, W.; Yao, X. Robust audio copy-move forgery detection using constant Q spectral Sketches and GA-SVM. *IEEE Trans. Dependable Secur. Comput.* **2022**, *20*, 4016–4031.
106. Skračić, K.; Petrović, J.; Pale, P. Bytercnn: Enhancing file fragment type identification with recurrent and convolutional neural networks. *IEEE Access* **2023**, *11*, 138176–138187.
107. Ali, R.R.; Mohamad, K.M.B.; Mostafa, S.A.; Zebari, D.A.; Jubair, M.A.; Alouane, M.T.H. A meta-heuristic method for reassemble bifragmented intertwined JPEG image files in digital forensic investigation. *IEEE Access* **2023**, *11*, 111789–111800.
108. Fayyaz, Y.; Almeahmadi, A.; El-Khatib, K. A hybrid artificial intelligence framework for enhancing digital forensic investigations of infotainment systems. *Forensic Sci. Int. Digit. Investig.* **2024**, *49*, 301751. <https://doi.org/10.1016/j.fsidi.2024.301751>.
109. Cho, S.H.; Kim, D.; Kwon, H.C.; Kim, M. Exploring the potential of large language models for author profiling tasks in digital text forensics. *Forensic Sci. Int. Digit. Investig.* **2024**, *50*, 301814. <https://doi.org/10.1016/j.fsidi.2024.301814>.
110. Cho, K.; Park, Y.; Kim, J.; Kim, B.; Jeong, D. Conversational AI forensics: A case study on ChatGPT, Gemini, Copilot, and Claude. *Forensic Sci. Int. Digit. Investig.* **2025**, *52*, 301855. <https://doi.org/10.1016/j.fsidi.2024.301855>.
111. Van Dijk, R.; Van de Wetering, J.; Argentini, R.; Gorka, L.; Van Luenen, A.F.; Minnema, S.; Rijgersberg, E.; Ugen, M.; Ádám Mann, Z.; Geradts, Z. PaSSwOrdVib3s!: AI-assisted password recognition for digital forensic investigations. *Forensic Sci. Int. Digit. Investig.* **2025**, *52*, 301870. <https://doi.org/10.1016/j.fsidi.2025.301870>.
112. Sharma, B.; Ghawaly, J.; McCleary, K.; Webb, A.M.; Baggili, I. ForensicLLM: A local large language model for digital forensics. *Forensic Sci. Int. Digit. Investig.* **2025**, *52*, 301872. <https://doi.org/10.1016/j.fsidi.2025.301872>.
113. Qadir, S.; Noor, B. Applications of Machine Learning in Digital Forensics. In *Proceedings of the 2021 International Conference on Digital Futures and Transformative Technologies (ICoDT2)*; IEEE: New York, NY, USA, 2021; pp. 1–8. <https://doi.org/10.1109/ICoDT252288.2021.9441543>.
114. Dunsin, D.; Ghanem, M.C.; Ouazzane, K. The use of artificial intelligence in digital forensics and incident response (DFIR) in a constrained environment. In *Proceedings of the International Conference on Digital Forensics and Security of Cloud Computing*; IEEE: New York, NY, USA, 2022.
115. Göbel, T.; Breitingner, F.; Baier, H. Optimising data set creation in the cybersecurity landscape with a special focus on digital forensics: Principles, characteristics, and use cases. *Forensic Sci. Int. Digit. Investig.* **2025**, *52*, 301882. <https://doi.org/10.1016/j.fsidi.2025.301882>.
116. Wickramasekara, A.; Breitingner, F.; Scanlon, M. Exploring the potential of large language models for improving digital forensic investigation efficiency. *Forensic Sci. Int. Digit. Investig.* **2025**, *52*, 301859. <https://doi.org/10.1016/j.fsidi.2024.301859>.
117. Garfinkel, S.L.; McCarrin, M. Hash-based carving: Searching media for complete files and file fragments with sector hashing and hashdb. *Digit. Investig.* **2015**, *14*, S95–S105. <https://doi.org/10.1016/j.diin.2015.05.001>.
118. Sontan, A.D.; Samuel, S.V. The intersection of Artificial Intelligence and cybersecurity: Challenges and opportunities. *World J. Adv. Res. Rev.* **2024**, *21*, 1720–1736.
119. Pooyandeh, M.; Han, K.J.; Sohn, I. Cybersecurity in the AI-Based Metaverse: A Survey. *Appl. Sci.* **2022**, *12*, 12993.
120. Dunsin, D.; Ghanem, M.C.; Ouazzane, K.; Vassilev, V. A comprehensive analysis of the role of artificial intelligence and machine learning in modern digital forensics and incident response. *Forensic Sci. Int. Digit. Investig.* **2024**, *48*, 301675.
121. González-Arencia, M.; Ordoñez-Erazo, H.; González-Sanabria, J.S. Explainable Artificial Intelligence as an Ethical Principle. *Ingeniería* **2024**, *29*, e21583.
122. Goodfellow, I.; McDaniel, P.; Papernot, N. Making machine learning robust against adversarial inputs. *Commun. ACM* **2018**, *61*, 56–66.
123. Xie, C.; Zhang, Z.; Zhou, Y.; Bai, S.; Wang, J.; Ren, Z.; Yuille, A.L. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2019; pp. 2730–2739.
124. Athalye, A.; Carlini, N.; Wagner, D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the International Conference on Machine Learning*; PMLR: Cambridge, MA, USA, 2018; pp. 274–283.

125. Guo, Y.; Stutz, D.; Schiele, B. Improving robustness of vision transformers by reducing sensitivity to patch corruptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2023; pp. 4108–4118.
126. Mökander, J.; Schuett, J.; Kirk, H.R.; Floridi, L. Auditing large language models: A three-layered approach. *AI Ethics* **2023**, *4*, 1085–1115.
127. Scanlon, M.; Breiting, F.; Hargreaves, C.; Hilgert, J.N.; Sheppard, J. ChatGPT for digital forensic investigation: The good, the bad, and the unknown. *Forensic Sci. Int. Digit. Investig.* **2023**, *46*, 301609.
128. Baghrmian, M.; Carter, J.A. *Relativism*. New Problems of Philosophy. Routledge: New York, USA, 2020.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.