MAIKRON

Journal of Artificial Intelligence and Computing Applications

Review article

The current role of artificial intelligence in cyberattacks: a short narrative review

Adrian Yama-Uitz[1], Miguel Cutz-Pech[1], Jose Munoz-May[1], Leonardo Romero-Pech[1], Jarol Lizama-Chan[1], and Zyon Maximo Rodriguez-Gonzalez[1]

[1]Tecnológico Nacional de México/IT de Mérida, Yucatán, Mexico

ABSTRACT

Artificial Intelligence (AI) exerts a growing impact on the field of cybersecurity, which is characterized by increasing complexity resulting from system interconnectivity, the expansion of the Internet of Things (IoT), and the rise in digital threats. While AI has been widely used to improve defense systems, its application in cyberattacks represents an emerging challenge. Malicious actors are leveraging this technology to automate offensives, adapt to defenses in real time, and exploit vulnerabilities with unprecedented efficiency, raising questions about how it is changing the dynamics of current attacks. This review examines five recent empirical studies published between 2021 and 2025 that address the offensive use of AI in cybersecurity, identifying key techniques such as automated code generation, evasion of detection mechanisms, execution of targeted attacks, and the reduced technical skill requirements for attackers. The results reveal that AI is transforming cybersecurity strategies, enabling more sophisticated and adaptive attacks that pose significant challenges to current detection tools. These threats are particularly critical in sectors such as healthcare, critical infrastructure, and public administration, where disruptions can have systemic consequences. The results point to the need to advance the development of more adaptable and scalable defense systems, as well as to identify key research gaps in areas such as explainability, regulation, and real-time detection of AI-driven threats.

Keywords: artificial intelligence, cyber attacks, cybersecurity vulnerabilities

1. Introduction

Artificial Intelligence (AI) has been recognized for its potential to optimize critical processes in areas as diverse as medicine, engineering, and finance [1, 2]. This transformation is underpinned by its ability to process large volumes of data, identify complex patterns, and optimize decision-making [2]. However, the same potential that drives these advances poses new challenges when applied to the field of cybersecurity, where its capabilities can be leveraged for different purposes [3].

AI is not just a complementary tool in cybersecurity, but a fundamental pillar to address the speed and sophistication of current and future threats [4]. Despite efforts to address this emerging trend, the scale and pace of its implementation, the speed of its advancement, and the specific methods used are not yet fully understood [5, 6]. This lack of understanding highlights the need for an analysis of how AI is redefining offensive capabilities in the field of cybersecurity.

This review seeks to examine the current use of AI in cyberattacks, with particular attention to how malicious actors employ it to automate attacks, evade detection, and exploit vulnerabilities in increasingly complex and interconnected systems. This paper does not propose defense mechanisms but rather synthesizes the current

E-mail address: le25080033@merida.tecnm.mx
https://doi.org/10.5281/zenodo.20802486
 ISSN: 3061-8843

literature on the offensive use of AI. By analyzing recent studies, it seeks to shed light on how these applications expand the reach and efficiency of cyberthreats, challenging traditional security paradigms and showing how AI improves the operational execution of modern cyberattacks with minimal human intervention.

The remainder of this article is structured as follows: Section 2 details the approach used for literature selection and analysis, describing the databases consulted, the inclusion and exclusion criteria, and the search strategies employed. Furthermore, the use of the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) model for literature selection is presented, along with the criteria applied. Subsequently, Section 3 presents the main emerging trends and techniques in the use of AI in cyberattacks, including automation methods, detection evasion, and vulnerability exploitation. Section 4 delves into the impact of these findings, assessing their implications for cybersecurity and considering the challenges posed by the use of AI in malicious activities. Finally, Section 5 summarizes the key points of the analysis, addresses potential mitigation strategies, and suggests future directions for research in this field. This framework allows for a comprehensive approach to the role of AI in cyberattacks, combining a critical review of the literature with an analysis of its practical and theoretical implications.

2. Methodology

To ensure an organized, transparent, and reproducible approach to identifying relevant literature, this review adopts the fundamental principles of the PRISMA methodology [7]. While not a systematic review in the strict sense, this framework serves as a methodological guide to structure a clear and replicable selection process, aiming to minimize potential bias in the synthesis of information. The primary objective is to identify and analyze recent relevant studies that significantly contribute to the understanding of the role of AI in the execution of cyberattacks. Through a rigorous selection process, this review ensures that the included papers are relevant, methodologically sound, and reflect the latest advances in this emerging field.

The bibliographic search was conducted using Google Scholar, selected for its broad coverage and ease of access to locate academic articles in different areas of knowledge. A specific search query was designed to ensure the relevance of the retrieved studies: ("artificial intelligence") AND ("cyberattacks") AND ("automation OR detection evasion") AND ("cybersecurity vulnerabilities OR malicious actors") AND ("emerging techniques OR cybersecurity trends"). This query aimed to collect studies from 2021 onwards related to AI-driven cyberattacks, with a particular focus on emerging threats. This search yielded a total of 135 records, which formed the initial dataset for subsequent analysis.

During the first selection phase, a thorough review of the titles and abstracts of the retrieved studies was conducted to eliminate those that did not meet the specific objectives of the review. Three main exclusion criteria were established: first, studies that, although related to technological topics, did not explicitly address the intersection between AI and cyberattacks were eliminated; second, works not published in scientific journals were discarded, ensuring the reliability and quality of the sources consulted; and third, review articles were excluded, given that the purpose of this research was to focus on primary research contributions that offered empirical data and original results. This filtering process ensured that only research of direct relevance to the study of AI-motivated cyberattacks was selected, resulting in the exclusion of 117 studies and the inclusion of 18 articles, which were evaluated for further analysis.

During the eligibility phase, the 18 shortlisted articles were thoroughly reviewed to ensure their methodological rigor, thematic relevance, and potential impact. In this second stage, 13 studies were excluded for the following reasons: one article was excluded due to language restrictions, eleven were not indexed in recognized indexes such as Journal Citation Reports (JCR), and one did not fit the thematic focus of this review. As a result, the final set consisted of five articles that fully met the established criteria. Figure 1 presents a flowchart adapted from the PRISMA model that summarizes the entire process of study identification, selection, eligibility assessment, and inclusion.

The five selected articles represent recent, high-impact contributions at the intersection of AI and cyberattacks. Table 1 presents a detailed summary of these studies, including the article title, journal of publication, and corresponding year. While this review was developed following a structured approach based on the principles of this established methodology, certain methodological limitations were identified. The exclusive use of Google Scholar as a search source could limit the scope and comprehensiveness of the review. Likewise, the focus on research published in recent years could have excluded relevant previous work. Similarly, the exclusion of non-indexed journals could have overlooked relevant contributions from emerging researchers or from specialized fields. Despite these limitations, the selection process, following a systematic approach, ensures that the final set of studies includes high-quality, relevant, and methodologically sound research.

3. Thematic Overview

This section critically organizes, analyzes, and synthesizes key findings from selected literature on the use of AI in cyberattacks, with particular attention to emerging techniques and research gaps. Rather than simply summarizing individual studies, this section identifies key patterns, trends, and areas of debate that emerged during the review. The thematic analysis focuses on four major domains: the automation and escalation of cyberattacks enabled by AI-driven orchestration; the automated generation of malicious code through large language models (LLMs); the use of adaptive AI tech-

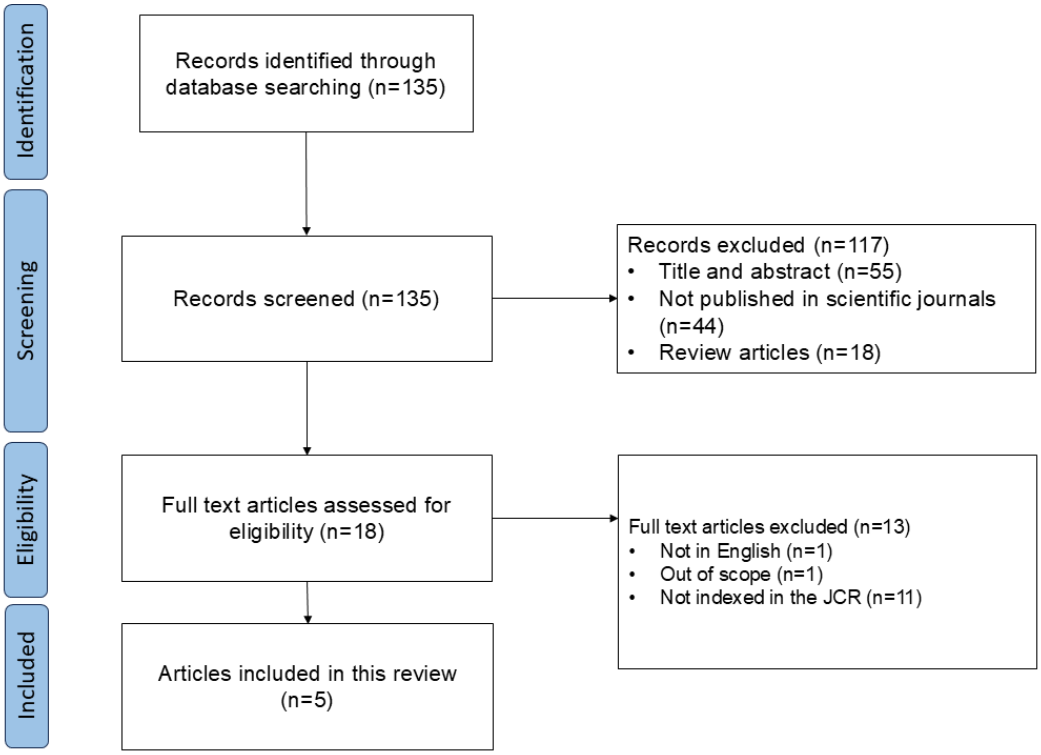

Figure 1. Flowchart of the literature identification, screening, eligibility assessment, and inclusion process.

Table 1. The final list of articles used in this review, including information for title, journal, year of publication and citation.

Title	Journal	Year	Citation
Unleashing offensive artificial intelligence: Automated attack technique code generation	Computers & Security	2024	[1]
Designing and Evaluating a Flexible and Scalable HTTP Honeypot Platform: Architecture, Implementation, and Applications	Electronics	2023	[5]
ChatGPT or Bard: Who is a better Certified Ethical Hacker?	Computers & Security	2024	[8]
Leveraging Explainable Artificial Intelligence in Real-Time Cyberattack Identification: Intrusion Detection System Approach	Applied Sciences-Basel	2023	[9]
A CNN-based automatic vulnerability detection	EURASIP Journal on Wireless Communications and Networking	2023	[10]

niques to evade intrusion detection and monitoring systems; and the challenges and opportunities for AI-Based cyberattack detection and attribution. Each of these themes is examined in depth in the following subsections, highlighting influential findings, unresolved questions, and areas where further research is necessary. A summary of these thematic domains is presented in Table 2.

3.1 Automation and escalation of AI-driven attacks

Advances in AI have made it possible to automate multiple phases of the cyberattack lifecycle, including reconnaissance, exploitation, and evasion of defense systems [8]. Generative models and adaptive architectures allow attackers to dynamically modify their attack vectors, making them difficult to detect by traditional systems [5]. Automation not only accelerates the deployment of attacks but also enables their mass replication without human intervention [1, 8, 5]. This poses a significant threat to environments that are not prepared for immediate responses to thousands of simultaneous events [5].

Current defensive capabilities do not adequately respond to this level of adaptability. While attackers integrate self-tuning and personalizing algorithms, defenses remain reactive, slow, and structurally static, as many intrusion detection systems lack simultaneous adaptability and rely on outdated datasets that do not represent evolving threat patterns [9]. Furthermore, most architectures still rely on fixed rules or non-adaptive models, making them poor at responding to dynamic, behavioral attacks generated by AI [10]. This asymmetry creates an operational gap in cybersecurity that not only compromises current defense strategies [1], but also exposes a broader research challenge: while offensive capabilities are increasingly driven by adaptive automation, existing defensive models have not yet reached a comparable level of responsiveness in concurrent environments [9, 5].

3.2 Automated generation of cyberattack techniques through large language models

AI has reached a point where it can be used not only to automate defensive tasks in cybersecurity, but also to facilitate the execution of cyberattacks [8]. The study [1] shows how generative models like ChatGPT can be prompted to produce malicious code aligned with MITRE ATT&CK techniques. These models allow the creation of specific code fragments to execute advanced attack techniques such as defense evasion, privilege escalation, or initial access, with just a textual instruction [1]. This automation considerably reduces the effort required by attackers and eliminates the need for specialized technical knowledge [1, 9], which broadens the range of users capable of carrying out sophisticated attacks.

In controlled tests, some LLMs have been shown to autonomously generate attack code with a 16% success rate, a figure that rises to 50% when guided by users with cybersecurity expertise [1]. This result highlights the tangible offensive potential of LLMs and the urgent need to develop early mitigation mechanisms against their malicious use [8].

The ability of these models to generate executable code, adapted to different operating systems and offensive tactics, also allows them to be integrated with automated breach and attack simulation platforms [9, 8]. This favors the customization of attacks according to the target environment, making their detection and response more difficult. Furthermore, the dynamic generation of variants of the same attack increases the complexity of threats, forcing defensive systems to evolve rapidly. In this context, LLMs represent a powerful tool that is transforming the offensive cybersecurity landscape, both due to their speed and operational versatility [8].

3.3 Evasion of monitoring and detection systems in cyberattacks using AI

Evading monitoring and detection systems is one of the most complex areas where attackers can leverage AI to optimize their attacks [5]. Traditionally, intrusion detection systems (IDS) and other monitoring tools have been designed to identify anomalous behavior in network traffic or unusual patterns in systems [5, 8]. However, with the use of AI, attackers can now generate adaptive attacks that instantly modify their behavior to evade defense strategies [1, 8].

One of the most advanced techniques is the use of AI algorithms to anticipate defensive responses and adapt attack strategies accordingly, enabling real-time evasion of monitoring mechanisms [1, 8]. This type of dynamic adaptation poses a critical problem for defenses: how to anticipate and detect constantly changing threats. Traditional detection systems are not designed to react so quickly or to predict changes in attack patterns [5, 9]. This creates a security gap, where attackers continue to advance while traditional defenses fall behind.

3.4 Challenges and Opportunities for AI-Based Cyberattack Detection and Attribution

The increasing sophistication of cyberattacks, driven by the integration of AI, poses unprecedented challenges in the detection and attribution of these threats. AI's ability to automate entire phases of the attack lifecycle, from reconnaissance to evasion of defenses, requires a deeper understanding of traditional security strategies [8]. Unlike conventional attacks, the dynamic adaptability inherent in AI-based offensive systems allows for modification of tactics and vectors even during execution, making them considerably more difficult to identify using systems based on static signatures or predefined rules [5]. Studies demonstrate that these adaptive strategies weaken detection models by allowing attacks to modify behavior in real time, circumventing traditional rule-based mechanisms [5].

In this scenario, where attackers are leveraging the

automatic personalization and scalability offered by AI, it is crucial to explore both the challenges inherent in detecting and tracking these new forms of evasive attacks and the opportunities that AI itself can offer to develop more advanced, predictive, and effective forensic techniques [9]. A proactive response to these emerging threats lies in the implementation of machine learning models such as convolutional neural networks (CNNs)[10], which have proven effective in automatically detecting vulnerabilities in source code. This technique represents a crucial tool for preventing exploits before they materialize [10], thus helping to reduce the gap in the AI-driven offensive-defensive cycle. However, despite recent advances, existing detection systems still have limitations in real-time attribution, especially when faced with rapidly evolving and adaptive attack patterns [1].

As previously discussed, AI has brought about a substantial shift in cybersecurity strategies, equally affecting offensive and defensive approaches [10]. While AI has improved the detection of and response to cyberthreats, it has also increased the sophistication of attacks, increasing the need for more adaptable and scalable defenses [5]. As AI systems continue to develop, addressing the implications of their use, especially in terms of privacy, accountability, and autonomy, is critical to ensuring responsible and transparent use in cybersecurity [8].

4. Discussion

This section seeks to critically interpret the previously presented thematic findings, analyzing their strategic implications for contemporary cybersecurity [10, 5]. These findings pose significant challenges for current detection tools and, at the same time, suggest the need to review the approach to defense strategies in a rapidly evolving technological environment.

AI-powered cyberattacks not only increase their destructive capacity, but also potentially change the strategic balance between attackers and defenders [10]. The reduction in human intervention required, thanks to intelligent automation, expands the scope and frequency of attacks, opening the door to actors with less technical expertise [1]. Taken together, these advancements represent a paradigm shift in cybersecurity, indicating that traditional static defense mechanisms are increasingly ineffective against highly dynamic AI-driven threats [5].

The practical implications of these findings are especially critical for sectors such as healthcare, critical infrastructure, and public administration, where a disruption can have systemic consequences [9]. The use of AI to increase the adaptability and potential personalization of cyberattacks poses a direct threat to the operational and reputational integrity of organizations, particularly given the potential misuse of sensitive data [5, 1, 8].

This review is consistent with previous studies that have documented the use of AI to increase the complexity and effectiveness of cyberattacks, as well as the limitations of traditional defense systems [8]. However, differences were identified in the methodological approaches and timeframes used. While much of the literature has focused on theoretical analyses or hypothetical scenarios, this work focuses on observable applications and techniques already implemented in actual contexts [5, 10]. The sectors analyzed also contributed to notable differences in the findings, emphasizing the importance of specific contexts in cybersecurity research.

Among the most significant methodological limitations are the exclusive use of Google Scholar as a search source, the exclusion of literature not indexed in JCR, and the focus on recent studies. These restrictions may have reduced the diversity of perspectives or excluded relevant previous references. However, although the findings presented are representative of the current state of the field, they should be interpreted within the framework of these limitations.

There is a need to expand the study of AI-driven cyberthreats to under-researched sectors such as education and logistics, and to specifically focus on the development of explainable defensive AI systems [9, 8]. Furthermore, conducting larger collaborative reviews and longitudinal empirical studies on real AI-driven cyberattacks would significantly improve understanding. Likewise, emphasis should be placed on developing and refining ethical and regulatory frameworks to guide the responsible deployment of AI technologies in cybersecurity contexts [8].

AI-powered cyberthreats have pushed the traditional boundaries of digital defense, introducing more complex, rapid, and adaptive attack scenarios [9, 10, 5]. Bridging this gap requires not only technical innovation, but also international cooperation, proactive cybersecurity policies, and the urgent development of ethical frameworks to guide the responsible use of AI in increasingly interdependent digital environments [6, 8].

5. Conclusion

AI is radically transforming the cyberattack landscape. Its ability to automate attacks, generate malicious code using LLMs, adapt in real time, and reduce reliance on human intervention has increased both the effectiveness and scope of threats. These characteristics represent a critical challenge for traditional detection systems, which fail to keep up with the pace and variability of AI-powered attacks [5].

The findings pose significant implications for cybersecurity, both practical and academic [1, 8]. AI-based offensive technologies can mimic legitimate behavior, dynamically adapt, and execute large-scale personalized attacks, with potential impact not only technical but also psychological, in terms of more effective automated persuasion, emotionally personalized attacks, and reduced user ability to identify compelling deceptions. These advances demand a rethinking of current methodologies for threat detection and assessment.

The offensive use of AI in cyberattacks represents

Table 2. Key thematic domains and insights in AI-enhanced cyberattack research as identified in the literature review, including automation and scaling of attacks, malicious content generation via LLMs, evasion of detection systems, and challenges in attribution and defense.

Thematic Research Domain	**Summary of Key Findings and Strategic Implications**
Automation and Escalation of AI-Driven Attacks	AI enables full-cycle automation of cyberattacks (e.g., reconnaissance, exploitation, evasion), allowing massive, scalable, and adaptive threats. Defensive systems remain mostly static and reactive, resulting in critical operational gaps and higher exposure to AI-augmented offensive strategies.
Automated Generation of Cyberattack Techniques via LLMs	LLMs like ChatGPT can be exploited to generate malicious code and phishing content, lowering the barrier to entry for less-skilled actors. These tools are increasingly integrated into offensive platforms like Breach and Attack Simulation (BAS) systems, improving attack customization and stealth.
Evasion of Monitoring and Detection Systems	AI-driven adversarial techniques allow real-time behavioral adaptation to evade traditional intrusion detection systems (IDS) and firewalls. These strategies significantly reduce the effectiveness of signature-based and rule-based defense models.
Challenges and Opportunities in Detection and Attribution	While AI complicates traceability and undermines conventional forensic methods, it also offers new avenues for advanced threat attribution, behavioral analytics, and proactive cyber defense strategies using predictive models.

a paradigm shift in how digital threats are configured. Understanding its scope and consequences is essential to assessing the risk posed by a technology that, while designed to drive progress, can also be used to amplify damage. Contemporary cybersecurity must recognize this inherent duality of AI: its ability to protect, but also to exploit.

Ethics Statement

This study did not involve human participants or animals and therefore did not require ethical approval.

CRediT authorship contribution statement

Adrian Yama-Uitz: Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Miguel Cutz-Pech:** Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Jose Munoz-May:** Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Leonardo Romero-Pech:** Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Jarol Lizama-Chan:** Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Zyon Maximo Rodriguez-Gonzalez:** Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT to assist with language clarity and grammar correction. All final edits and intellectual content were determined and verified by the authors.

Declaration of competing interest

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

References

[1] E. Iturbe, O. Llorente-Vazquez, A. Rego, E. Rios, and N. Toledo, "Unleashing offensive artificial intelligence: Automated attack technique code generation," *Computers & Security*, vol. 147, p. 104077, dec 2024. doi: 10.1016/j.cose.2024.104077. [Online]. Available: https://doi.org/10.1016/j.cose.2024.104077

[2] C. Colther and J. P. Doussoulin, "Artificial intelligence: Driving force in the evolution of human knowledge," *Journal of Innovation and Knowledge*, vol. 9, no. 4, p. 100625, 2024. doi: 10.1016/j.jik.2024.100625. [Online]. Available: https://doi.org/10.1016/j.jik.2024.100625

[3] A. A. Siam, M. Alazab, A. Awajan, and N. Faruqui, "A Comprehensive Review of AI's Current Impact and Future Prospects in Cybersecurity." *IEEE Access*, vol. 13, no. December 2024, pp. 14 029–14 050, 2025. doi: 10.1109/ACCESS.2025.3528114.

[4] B. Guembe, A. Azeta, S. Misra, V. C. Osamor, L. Fernandez-Sanz, and V. Pospelova, *The Emerging Threat of Ai-driven Cyber Attacks: A Review.* Taylor Francis, 2022, vol. 36, no. 1. doi: 10.1080/08839514.2022.2037254. [Online]. Available: https://doi.org/10.1080/08839514.2022.2037254

[5] M. Rabzelj, L. Š. Južnič, M. Volk, A. Kos, M. Kren, and U. Sedlar, "Designing and Evaluating a Flexible and Scalable HTTP Honeypot Platform: Architecture, Implementation, and Applications," *Electronics*, vol. 12, no. 16, p. 3480, aug 2023. doi: 10.3390/electronics12163480. [Online]. Available: https://doi.org/10.3390/electronics12163480

[6] I. H. Sarker, Y. B. Abushark, F. Alsolami, and A. I. Khan, "IntruDTree: A machine learning based cyber security intrusion detection model," *Symmetry*, vol. 12, no. 5, pp. 1–15, 2020. doi: 10.3390/SYM12050754. [Online]. Available: https://doi.org/10.3390/sym12050754

[7] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M. M. Lalu, T. Li, E. W. Loder, E. Mayo-Wilson, S. McDonald, L. A. McGuinness, L. A. Stewart, J. Thomas, A. C. Tricco, V. A. Welch, P. Whiting, and D. Moher, "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *The BMJ*, vol. 372, 2021. doi: 10.1136/bmj.n71. [Online]. Available: https://doi.org/10.1136/bmj.n71

[8] R. Raman, P. Calyam, and K. Achuthan, "ChatGPT or Bard: Who is a better Certified Ethical Hacker?" *Computers & Security*, vol. 140, p. 103804, may 2024. doi: 10.1016/j.cose.2024.103804. [Online]. Available: https://doi.org/10.1016/j.cose.2024.103804

[9] X. Larriva-Novo, C. Sánchez-Zas, V. A. Villagrá, A. Marín-Lopez, and J. Berrocal, "Leveraging Explainable Artificial Intelligence in Real-Time Cyberattack Identification: Intrusion Detection System Approach," *Applied Sciences*, vol. 13, no. 15, p. 8587, jul 2023. doi: 10.3390/app13158587. [Online]. Available: https://doi.org/10.3390/app13158587

[10] J. H. An, Z. Wang, and I. Joe, "A CNN-based automatic vulnerability detection," *EURASIP Journal on Wireless Communications and Networking*, vol. 2023, no. 1, p. 41, may 2023. doi: 10.1186/s13638-023-02255-2 . [Online]. Available: https://doi.org/10.1186/s13638-023-02255-2