World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

AI-powered threat detection: Opportunities and limitations in modern cyber defense

Kenechukwu Ikenna Nnaka [1, *], Paul Oluchukwu Mbamalu [2], John Cherechim Nwaigbo [3], Peter Chika Ozo-ogueji [4], Victor Ifeanyi Njoku [5] and Chijioke Cyriacus Ekechi [6]

[1] *Department of Chemical Engineering, University of Benin.*
[2] *Department of Project Management Technology, Federal University of Technology Owerri, Nigeria.*
[3] *Department of Mechanical Engineering/University of Nigeria, Nsukka.*
[4] *Department of Mathematics and Statistics (Data Science) CAS, American University, Washington, DC.*
[5] *Cybersecurity Professional, Department of Business Management,Miva Open University.*
[6] *Department of Electrical and Computer Engineering, Tennessee Technological University.*

World Journal of Advanced Research and Reviews, 2025, 27(02), 210-223

Publication history: Received on 25 June 2025; revised on 30 July 2025; accepted on 02 August 2025

Article DOI: https://doi.org/10.30574/wjarr.2025.27.2.2854

Abstract

Artificial intelligence (AI) and machine learning (ML) have become critical components of modern cybersecurity strategies, offering dynamic capabilities for detecting, analyzing, and mitigating cyber threats. This review synthesizes existing literature to explore how AI and ML technologies are being applied in cyber threat detection, focusing on their operational integration, effectiveness, and limitations. The study draws on 43 referenced sources, including peer-reviewed journal articles, technical whitepapers, vendor documentation, and authoritative blogs, to provide a comprehensive overview of the field. Findings highlight that AI enhances threat detection through real-time data analysis, reduces false positives, and uses predictive modeling and adaptive learning. These technologies enable more proactive and scalable defense mechanisms compared to traditional rule-based systems. However, challenges persist, including the opacity of black-box models, vulnerability to adversarial attacks, data quality issues, and the lack of standard evaluation frameworks. Regulatory concerns and the need for human oversight further complicate widespread deployment. The review concludes that while AI significantly augments cyber defense capabilities, it is not a standalone solution. For AI to be effectively and ethically integrated into cybersecurity, it must be transparent, explainable, and aligned with organizational and regulatory goals. The study emphasizes the importance of explainable AI, robust datasets, and interdisciplinary collaboration in shaping the next generation of secure and trustworthy AI-driven defense systems.

Keywords: AI; Machine Learning; Cybersecurity; Threat Detection; SIEM; SOAR; XDR; Anomaly Detection; Adversarial AI

1. Introduction

The advent of artificial intelligence (AI) in cybersecurity is a decisive change when it comes to identifying, examining, and protecting threats [1], [2]. Initially limited to basic rule-based enhancements in intrusion detection systems, AI integration became more relevant in the middle of the 2010s when machine learning (ML) algorithms were applied to intrusion detection to detect anomalies, classify malware, and analyze user behavior [3], [15]. This evolution accelerated between 2017 and 2022 such that security-focused platforms have begun integrating AI-powered functionality into SIEM (e.g., Splunk, IBM QRadar), SOAR (e.g., Cortex XSOAR), and more recently XDR systems which combine endpoint and network telemetry to give a panoramic view of the threat [4], [5]. These platforms use AI in order to match the events, find patterns in streams of massive amounts of data, and even allow automating the incident response [6]. The technological growth has seen adoption in different areas of the industry though the effects are not even as better-

* Corresponding author: Kenechukwu Ikenna Nnaka

established organizations enjoy sophisticated implementations. Simultaneously, a lot of small businesses experience obstacles caused by finances, specialized knowledge, and their complexity in terms of integration [7]. However, AI solutions have become even the focal point of contemporary cybersecurity today, capable of providing their scalable approach to covering the diverse and changing threat environments [8].

Cybersecurity has increased significantly in its level of complexity over the last decade and the threats have not only multiplied but also increased in nature [9]. Advanced persistent threats (APTs), APTs, zero-day exploits, fileless malware, and multi-stage ransomware attacks now routinely bypass conventional perimeter defenses [10], [11]. Conventional rule-based and signature-driven intrusion detection systems (IDS) have trouble tracking these changing methods of attack, and they tend to produce a high false-positive rate and miss new or obfuscated attacks [12]. Further, the volume of log data created in endpoints, networks, and cloud environments has made manual processing of threats too cumbersome [13]. This puts into perspective the intensity of finding smarter, more scalable and adaptive solutions, to support or boost the existence of static security frameworks [14].

1.1. Problem Statement

Despite the growing integration of AI into modern cybersecurity systems, there are still major loopholes in terms of performance, reliability, and transparency of these solutions. Although AI and machine learning have enabled an increase in the speed and scalability of detection of known threats, limitations to these abilities exist and can be attributed to imbalanced datasets, adversarial manipulation, and model interpretability, among other factors. Most AI-based systems are black boxes, giving minimal transparency about how decisions are arrived at, which is a problem because it is complicated in forensics analysis, compliance, and user assurance. Also, the use of AI may demand special skills, as well as high computational power, making a particular application less possible in numerous organizations. These loopholes demonstrate the issue that requires a critical evaluation of the existing artificial intelligence-based methods, their weaknesses in their functioning, and the new dangers that lie in excessive faith in automated detect systems.

1.2. Aim and Scope of the Review

The objective of this review is to evaluate how artificial intelligence has been applied to cyber threat detection, particularly through platforms such as SIEM, SOAR, and XDR. It aims to:

- Examine the types of AI and machine learning models used for threat detection
- Analyze the strengths and limitations of these models in real-world scenarios
- Compare leading AI-powered security tools and their practical integration
- Identify ongoing technical, operational, and ethical challenges
- Offer recommendations for improving the transparency, scalability, and effectiveness of AI-based cybersecurity systems

2. Materials and Methods

2.1. Study Design

This paper adopts a structured narrative review methodology to evaluate the application of artificial intelligence (AI) and machine learning (ML) in cyber threat detection. Unlike systematic reviews, which require meta-analytical synthesis and strict protocol registration, a structured narrative review is appropriate for emerging and multidisciplinary fields such as AI in cybersecurity, where heterogeneity in study designs, technologies, and evaluation metrics limits quantitative aggregation. The study design follows a transparent and replicable protocol, emphasizing thematic categorization, tool-based comparisons, and critical analysis of trends.

2.2. Data Sources and Search Strategy

Literature was retrieved from a combination of academic databases and authoritative industry sources to ensure a comprehensive review of both theoretical advances and real-world applications. The primary academic databases consulted included IEEE Xplore, ACM Digital Library, Elsevier ScienceDirect, SpringerLink, and Wiley Online Library. In addition, technical whitepapers, vendor documentation, and threat intelligence reports were sourced from major cybersecurity firms such as IBM, Palo Alto Networks, Microsoft, and CrowdStrike, capturing tool-specific developments and applied insights.

A structured search strategy was employed using Boolean combinations of keywords related to AI and cybersecurity. Search terms included: “artificial intelligence” or “AI” combined with “cybersecurity” or “threat detection”; “machine learning” or “deep learning” combined with terms such as “intrusion detection system,” “SIEM,” “SOAR,” or “XDR”; “anomaly detection” in conjunction with “cyber defense” or “incident response”; and finally, “adversarial AI” or “explainable AI” with “network security.” These queries were designed to capture literature at the intersection of AI technologies and cyber defense mechanisms.

Searches were limited to the period from January 2017 to July 2025 to focus on the most recent and relevant developments. Only English-language publications were considered to ensure consistency in evaluation and interpretation across all sources.

2.3. Inclusion and Exclusion Criteria

Sources were selected based on their direct relevance to the application of AI and machine learning in cyber threat detection. Eligible studies included those that examined deployed tools, real-world system architectures, or documented case studies. Priority was given to literature that offered performance analysis of AI models, discussed practical benefits and limitations, or referenced widely adopted industry frameworks such as MITRE ATT&CK or NIST SP 800-53. Only sources that were peer-reviewed or published by technically credible vendors were included to ensure academic and practical reliability.

Studies were excluded if they focused exclusively on theoretical aspects of AI without linking to cybersecurity applications, lacked methodological transparency, or were classified as duplicates, opinion pieces, or editorial commentaries. Publications that did not address detection or defense-related use cases were also excluded from the final synthesis.

2.4. Screening and Data Extraction

An initial pool of 96 records was identified through structured searches across academic databases and industry sources. After removing duplicates and screening titles and abstracts for relevance, 66 records remained. Full texts of 34 of these were assessed based on inclusion and exclusion criteria. In addition, 9 more sources were included through manual reference checks, expert recommendations, and grey literature tracking, bringing the total to 43 sources for the final qualitative synthesis. These comprised 25 peer-reviewed journal articles, 2 peer-reviewed conference papers, 7 industry whitepapers and technical documentation, 7 authoritative blog posts and vendor-authored web articles, and 2 documents reflecting government or standards-related cybersecurity frameworks. The entire screening and inclusion workflow is summarized in Figure 1, a PRISMA-style flow diagram adapted for this narrative review.

Data were manually extracted using a standardized form capturing AI/ML model type (e.g., supervised, unsupervised, deep learning), application domain (e.g., SIEM, SOAR, anomaly detection), key metrics (e.g., mean time to detect, false positive rate), integration notes, and reported advantages or challenges. Thematic categorization was applied to support structured comparison across tools and approaches.

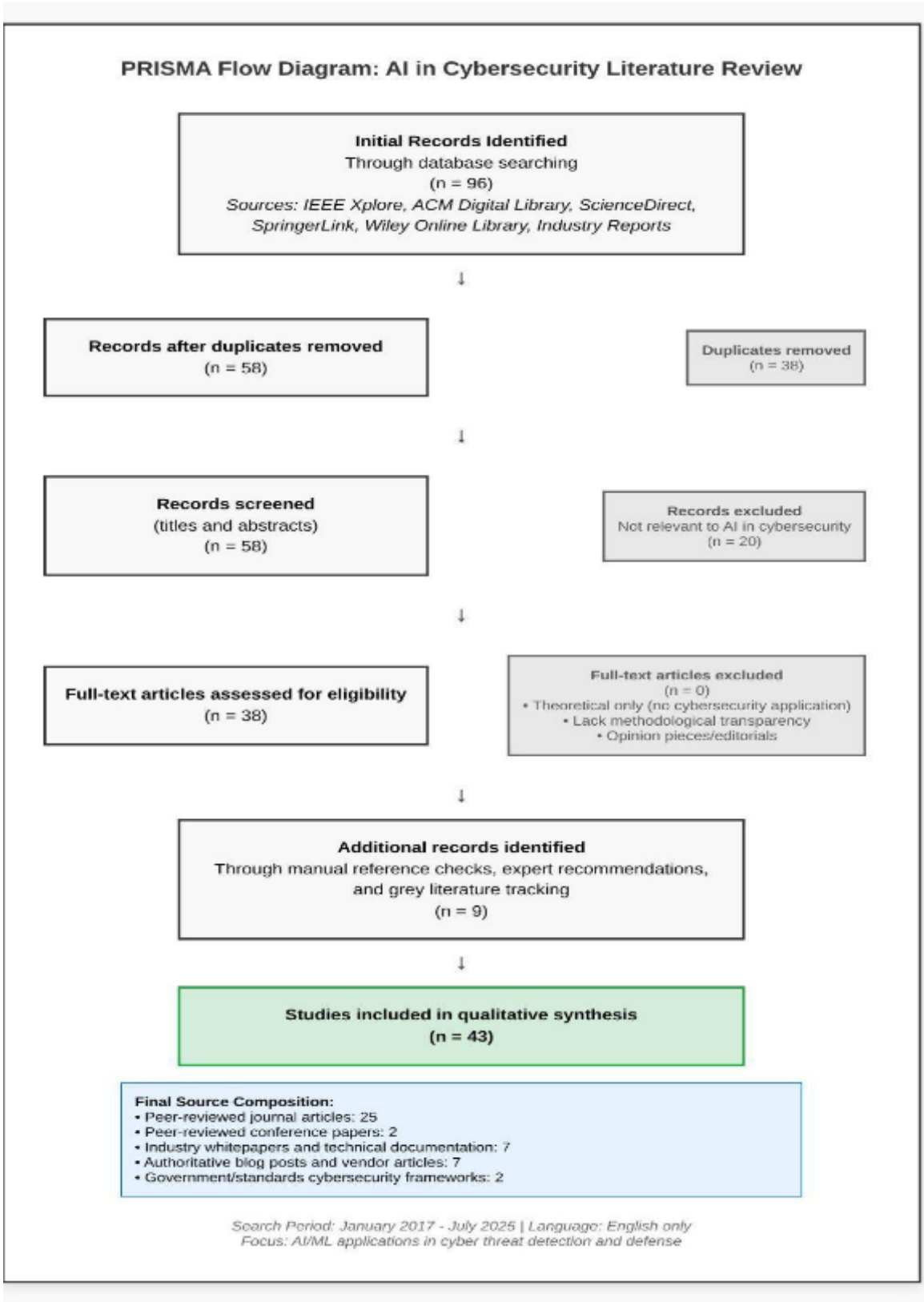

Figure 1 PRISMA flow diagram illustrating the literature screening and selection process for this structured narrative review on AI in cybersecurity [44]

3. Discussion

3.1. AI Models Used in Threat Detection

Artificial intelligence (AI) and machine learning (ML) have transformed the field of cyber threat detection by enabling systems to identify malicious behaviors, adapt to evolving threats, and operate at scale [2], [8]. These capabilities are driven by a range of algorithmic models that can be broadly categorized into supervised, unsupervised, and reinforcement learning approaches [16]. A comparative summary of these AI models, their cybersecurity use cases, and associated trade-offs is presented in Table 1

Supervised learning models are trained on labeled datasets that distinguish between benign and malicious activities. Common algorithms include support vector machines (SVMs), decision trees, and random forests, which are frequently applied in malware classification, spam filtering, and phishing detection [18]. For example, Fatima et al. showed that the optimized ensemble and linear classifiers like SGD, Extra Trees, Random Forest, and MLP had very high accuracy and F1-scores when used on spam email detection on three benchmarked datasets [17]. These models do very well in situations where the threat is known but may perform poorly when a new or obfuscated attack is detected because of the situations where models rely on historical data [18].

In contrast, unsupervised learning algorithms do not require labeled information and are particularly effective in identifying anomalies or deviations from expected behavior that may signal emerging or novel threats [19]. Common techniques used for detecting lateral movements, insider threats, and zero-day attacks include clustering algorithms such as K-means and DBSCAN, as well as autoencoders [1], [20]. Although these models offer greater flexibility, they often suffer from high false positive rates when not properly calibrated [21]. The distinct workflows of supervised and unsupervised learning models for cybersecurity threat detection are illustrated in Figure 2, highlighting differences in data requirements, processing stages, and detection outputs. The figure offers valuable insight into their respective implementation logic and operational distinctions [45]

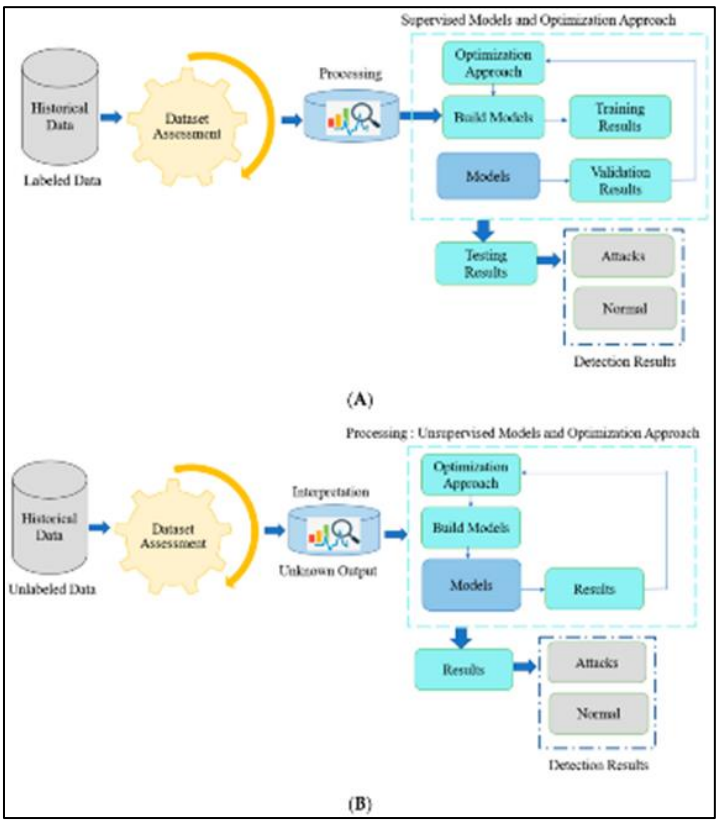

Figure 2 Comparison of supervised and unsupervised learning pipelines for cybersecurity threat detection [45]. (A) Supervised models rely on labeled historical data and involve model training, validation, and testing to classify inputs as attacks or normal. (B) Unsupervised models operate on unlabeled data, identifying patterns or anomalies through interpretation without prior labeling. Both approaches include optimization processes and yield detection results for threat identification

Deep learning techniques, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have been adopted for both supervised and unsupervised tasks due to their ability to learn complex patterns in high-dimensional data [22]. Recent studies have also explored graph neural networks (GNNs) for modeling relational data in network topologies and transformer-based architectures for log analysis and sequential prediction [23], [24].

Reinforcement learning (RL) is an emerging paradigm that allows detection systems to learn optimal responses to threats through interaction with the environment [25]. While RL has shown promise in dynamic defense scenarios such as honeypot adaptation and deception-based strategies, its application remains limited due to challenges in environment modeling, reward specification, and computational cost [26].

A critical distinction in AI-powered detection lies between anomaly-based and signature-based models. The former identifies deviations from normal patterns, enabling the discovery of novel threats but often at the expense of increased false positives [28]. The latter, though effective in recognizing known threats, fails to generalize across emerging attack vectors [1]. Contemporary detection frameworks increasingly adopt hybrid models that combine the strengths of both approaches to enhance detection precision and reduce alert fatigue [29].

As shown in Figure 3, the majority of AI/ML applications in cybersecurity are concentrated in intrusion detection and threat intelligence, each accounting for 25% of documented use cases [1]. This distribution reflects the prioritization of real-time attack detection and contextual analysis in modern security strategies. Notable real-world deployments include IBM Watson for Cybersecurity, which leverages natural language processing to correlate threat data [30], and Darktrace's Enterprise Immune System, which uses unsupervised learning to detect behavioral anomalies in corporate networks [31].

These examples underscore the diversity of AI models currently used in threat detection and highlight ongoing trade-offs between accuracy, interpretability, and adaptability in dynamic threat landscapes.

Table 1 Comparison of AI Models and Their Applications in Cybersecurity Threat

AI Model Type	Key Algorithms	Primary Applications	Strengths	Limitations
Supervised Learning	SVM, Decision Trees, Random Forest, MLP	Malware classification, spam/phishing detection	High accuracy on known threats; fast classification	Requires large labeled datasets; poor at novel attack detection
Unsupervised Learning	K-Means, DBSCAN, Autoencoders	Anomaly detection, insider threats, and zero-day attacks	Can detect unknown threats; no need for labeled data	High false positives; requires tuning of anomaly thresholds
Deep Learning	CNN, RNN, GNN, Transformer Models	Log analysis, behavioral modeling, image-based IDS	Learns complex patterns; adaptable to high-dimensional data	Computationally intensive; lacks interpretability ("black box")
Reinforcement Learning	Q-learning, DQN, Policy Gradient Methods	Dynamic defense, deception systems, and honeypot control	Learns optimal responses; suitable for adaptive defense scenarios	Sparse real-world deployment; reward modeling is complex
Hybrid Models	Combined supervised + unsupervised or DL models	Behavioral analysis, threat correlation, alert tuning	Improved generalization; balances precision and recall	Integration complexity requires constant re-training

Note: Adapted from sources including [1], [17]–[18], [20]–[26], [29].

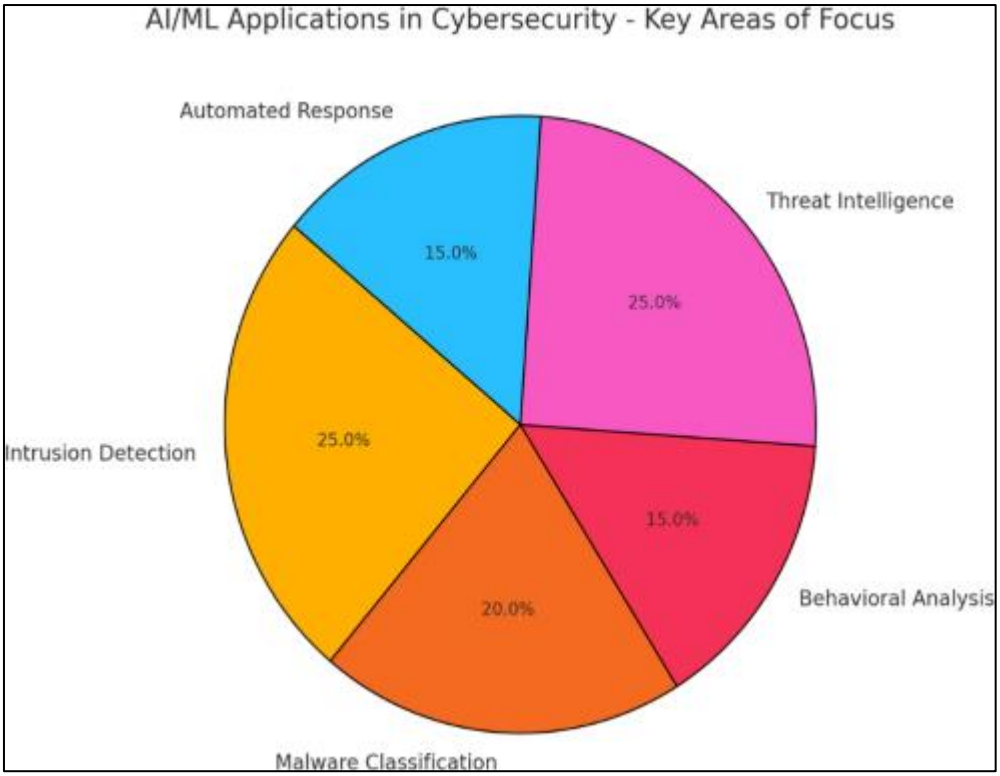

Figure 3 Key functional areas of AI/ML application in cybersecurity[1]. Intrusion detection and threat intelligence represent the largest focus areas (25% each), followed by malware classification (20%), behavioral analysis (15%), and automated response (15%)

3.2. Integrated Security Platforms

The integration of artificial intelligence into cybersecurity operations has extended beyond standalone detection models to comprehensive platforms that unify data collection, analysis, and automated response. Some of the latter include Security Information and Event Management (SIEM), Security Orchestration, Automation, and Response (SOAR), and Extended Detection and Response (XDR) systems, which are becoming the central architectural frameworks by which AI-guided threat detection and incident management is being actualized [32].

SIEM platforms, such as Splunk Enterprise Security and IBM QRadar act as a centralized location where logging data can be ingested, aggregated and correlated, all among various sources, including firewalls, intrusion detection systems, endpoint agents and cloud workloads [33]. Historically rule-driven, new SIEMs integrate machine learning and behavioral analytics to detect the odd exceptions, raise red flags' suspicious mode and rank alerts according to contextual risk score. For example, the Adaptive Response Framework by Splunk uses AI to streamline the process of enriching the threats, as well as auto-initiating dynamic response actions among the integrated tools. These improvements suppress analyst fatigue and false alarms and allow the detection of the stealthy multi-stage attacks that would increasingly elude fixed correlation rules.

SOAR platforms including Cortex XSOAR by Palo Alto Networks and Microsoft Sentinel are aimed at automating and orchestrating the incident response processes in a heterogeneous security setting [34]. AI plays a crucial role by enabling context-aware playbook selection, intelligent alert prioritization, and adaptive remediation strategies. For example, Cortex XSOAR includes supervised learning model integration to correlate the alerts with the previous case data, and Microsoft Sentinel employs natural language processing (NLP) to break down threat intelligence feeds and prescribe the necessary actions. This automation helps not only promote faster response times and consistency in incident processing procedure but can also help minimize human error and relieve the pressure on manual input of analysts.

Extended Detection and Response (XDR) solutions that are the next evolution in detection and response platforms and will enable the elimination of the operational silos between endpoint, network, cloud and identity telemetry that have traditionally existed. XDR platforms provide a more contextual and correlated threat detection capability by joining visibility across these areas and subjecting high-volume security information to AI-driven analytics in an effort to

normalize it and contextualize and interpret boxes of data beyond the capacity of security teams in real-time. Premier vendors, including CrowdStrike Falcon XDR, Palo Alto Networks Cortex XDR and Trend Micro Vision One, use tricks like escort learning, behavior analytics and automation correlation engines to follow advanced attack chains that can slip detection in silos [35]. Such convergence minimizes alert and investigation cycle times collaborative proactive containment of advanced persistent threats (APTs) and lateral movement along the hybrid environment.

The integration of modern security platforms to external threat intelligence frameworks and APIs, including MITRE ATT&CK, STIX/TAXII, as well as commercial threat feeds, is one of the major strengths of such platforms. AI models are employed to continuously ingest, analyze, and learn from these external sources, enabling dynamic updates to risk scoring algorithms, alert enrichment, and adaptive tuning of detection thresholds. This continuous learning pipeline allows platforms to shift from static, signature-based defense models toward proactive, intelligence-driven threat detection. As threat landscapes evolve rapidly, such integrations ensure that detection systems remain current, context-aware, and resilient against both known and emerging attack techniques [36] and are supported by comprehensive cross-telemetry correlation engines typical of XDR solutions [5]

Overall, the synergy between AI techniques and integrated security platforms has fundamentally reshaped modern cyber defense strategies. These systems now operate not merely as repositories of security telemetry, but as intelligent orchestration engines capable of autonomous decision-making, context-aware alerting, and rapid incident remediation. However, to fully realize their potential, successful deployment demands rigorous calibration, robust data governance, and seamless interoperability across heterogeneous environments.

Table 2 Comparison of SIEM, SOAR, and XDR Platforms in AI-Powered Cyber Defense

Platform	**Primary Function**	**AI Integration**	**Strengths**	**Limitations**	**Examples**
SIEM (Security Information and Event Management)	Centralized log aggregation, event correlation, and alerting	Machine learning for anomaly detection, adaptive correlation rules, threat scoring	Broad visibility across infrastructure; supports compliance and forensic analysis	High setup cost; noisy alerts; static rules require frequent tuning	Splunk Enterprise Security, IBM QRadar
SOAR (Security Orchestration, Automation, and Response)	Automated incident response and workflow orchestration	Context-aware playbooks, NLP for threat intel parsing, supervised models for alert triage	Speeds up response time; standardizes remediation; reduces analyst workload	Requires high-quality integrations and rule engineering	Cortex XSOAR, Microsoft Sentinel
XDR (Extended Detection and Response)	Unified detection across endpoint, network, cloud, and identity layers	Ensemble learning, behavioral analytics, and real-time signal correlation	Full-stack visibility; reduced alert fatigue; better detection of APTs	Still evolving; vendor lock-in; integration complexity	CrowdStrike Falcon XDR, Cortex XDR, Trend Micro Vision One

Note: Table compiled by the author using information from [5], [32]–[36].

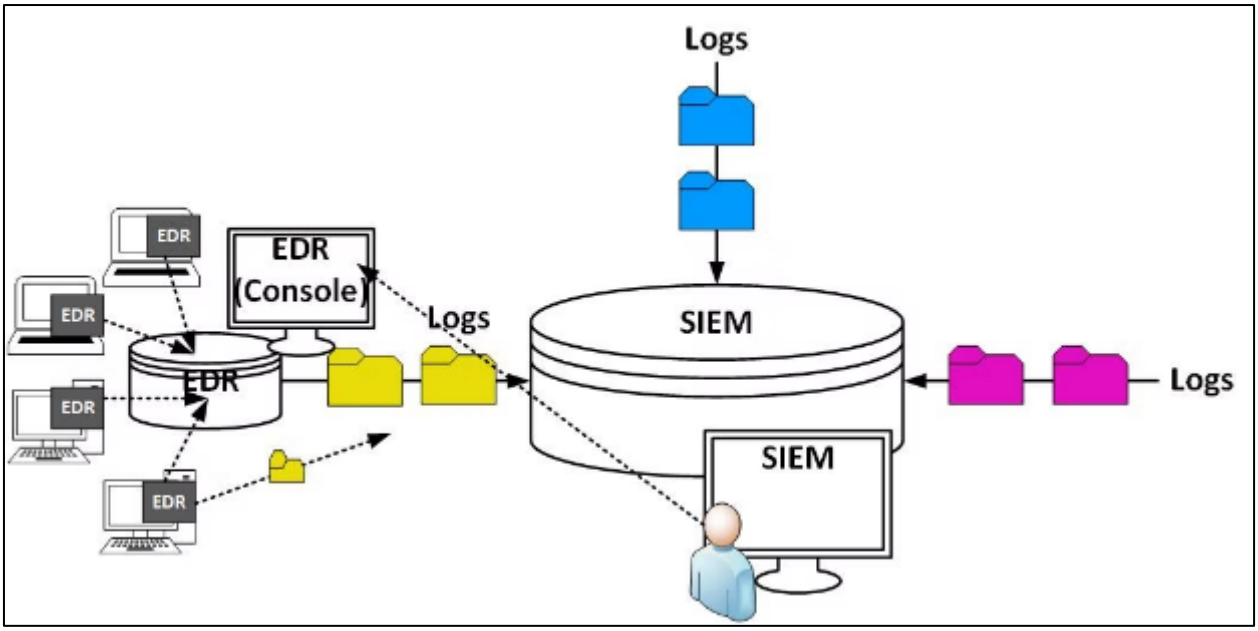

Figure 4 Integration of SIEM, SOAR, and EDR components in a unified cybersecurity architecture [32]. Logs from endpoint detection and response (EDR) tools are centralized through the SIEM system, which feeds into SOAR for automated incident response

3.3. Advantages of AI in Cyber Defense

With the use of artificial intelligence (AI) in cybersecurity activities, the possibilities of improving the identification, prioritization, and mitigation of threats is very promising. AI enhances both the speed and accuracy of cyber defense systems in several different ways by supplementing the effectiveness of more conventional detection mechanisms with learning-based systems and contextual scrutiny.

The potential of AI, in terms of the actual-time analysis of huge amounts of log information that are created in endpoints, servers, cloud-based systems, and network gadgets, is one of the most critical advantages of AI. This is in contrast to rule-based systems, which may be based on known threat signatures and thus unable to identify anomalies or possible breaches with a minimal delay, since the AI models, especially when driven by streaming analytics and pattern recognition, can process the data generated by the events dynamically [37]. This is particularly useful to contemporary organizations that face a multi-vectoring attack and time oriented exploits.

Another significant advantage is the lessened number of false positives, which is one of the difficulties of standard intrusion detection systems. By training on the typical behavior of their users, applications and systems, AI-based platforms can stop malicious deviations by using the baseline behavior as a filter. The effectiveness of such a behavior-based method is that it drastically reduces the number of irrelevant alerts that the security analysts are shown which enhances operational capabilities and allows them to triage incidents much faster [8].

AI also enables the use of predictive analytics, allowing cybersecurity systems to proactively identify potential threats, and intervene against it before it happens. With the help of correlation between past threat patterns and the present level of activity in the system, predictive models are further able to point to indications of compromise (IoCs) as well as suspicious behavioural chains, even when no full attack signature is available. Such a proactive approach contributes to early containment and hinders the increase of damage [38].

In addition, AI has the potential to be adaptively learned entailing that its systems constantly improve with respect to new and unfolding patterns of threat. With attackers changing tactics in order to avoid detection, AI algorithms have the ability to retrain on newly updated datasets or integrate new threat intelligence providing them an opportunity to optimize their detection logic. Such flexibility minimizes the chances of model obsolescence and improves ability to withstand dynamism [37].

The advantages are also supported by the performance indicators that compare the AI-based detection system with the classical approaches to cybersecurity. In terms of the operational metrics, as illustrated in Figure 5, AI-based strategies

show significant results in terms of major improvement. As an example, AI models have enhanced acceleration of real-time analysis to 92% up by 45%, and the number of false positives was reduced by 85%, as compared to 25 percent. There was an increase in the accuracy of detection of threats by 26 percent to 94 percent as predictive capability and adaptive learning efficiency increased to 78 percent and 89 percent respectively. The response to incidents also greatly increased-it increased to 88 percent as compared to 40 percent. These metrics support the concept of AI being a potentially powerful tool in advancing threat detection as well as making the least amount of work by the security analysts thus allowing it to take appropriate steps to mitigate the breach in time, hence the strategic value it may have in the contemporary cyber-security environment.

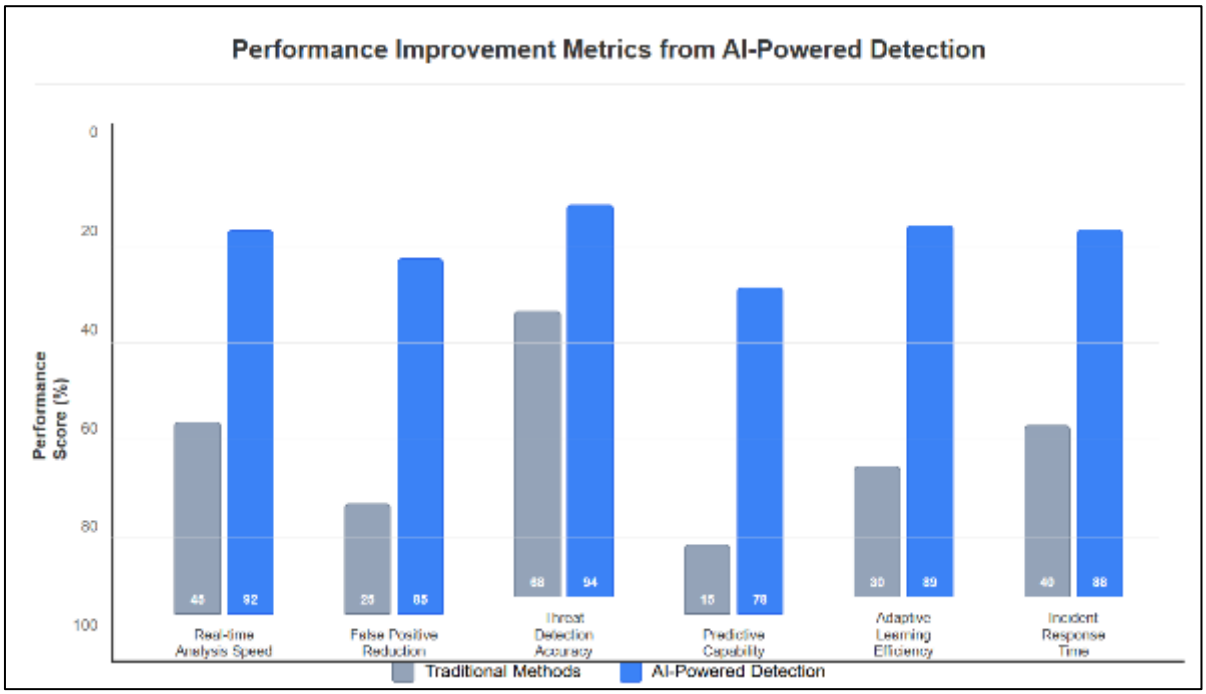

Figure 5 Performance improvement metrics from AI-powered detection systems compared to traditional methods. AI-driven approaches show strong gains in real-time speed of analysis, false positives, accuracy of detection of threats, predictive performance, and adaptive learning and response

3.4. Limitations and Ongoing Challenges

While artificial intelligence (AI) has proven to have significant potential in terms of improving cyber defense, there are still a number of limitations that remain in the way of its full-scale implementation and improving its effectiveness. These issues exist both in technical, organizational, and regulatory spheres, and display important gaps that are to be closed to achieve secure, scalable, and ethically aligned AI usage. The main strengths and weaknesses of AI in the field of cyber defense regarding such areas as detection, adaptation, and governance are provided in Table 3 below.

One of the most time-sensitive issues of AI-based cybersecurity is the data imbalance and scarcity of labeling, mostly in supervised learning environments. Security data are usually biased probably because the number of malicious events is surpassing the quantity of benign events and gives the model a low probability of generalizing well to unseen or infrequent classes of attack patterns [39]. In addition, the process of labeling cybersecurity data, particularly for more complex behavior such as lateral movement or polymorphic malware requires expert knowledge and a lot of time to label which can cause delays and inconsistencies when training a model.

The fact that AI models are prone to adversarial attacks and inputs purposefully altered to avoid being detected or used to trick classification is another major issue. In cyber security scenarios, the attacker may generate minor aberrations in the characteristics of the traffic or in the structure of the payload, which evades the defenses without affecting the malicious purpose. In particular, recent studies have shown how adversarial inputs may be used to attack IoT-enabled systems identifying weaknesses even in time-series models that might be used for predictive maintenance applications [40]. The presence of these risks establishes the importance of strong adversarial defense mechanisms to AI-powered cyberspace.

Lack of transparency, with most AI models being represented as an opaque or black box, especially within deep learning models, is a significant drawback to their actual use in security challenged settings. The security analysts also face an issue in validating decision-making logic, such as the decision to flag an alert or the decision to miss that alert, as well as incident response and reporting regulatory compliance. Even though explainable AI (XAI) methods such as SHAP and LIME are being used more frequently, they tend to be local in nature and their analysis is conducted using anecdotal rather than precise measures. This has cast doubt on their reliability and usefulness in operational settings, including cybersecurity [41].

Other than the technical limitations, organizational constraints also present a barrier. The major expense when implementing AI in cybersecurity is the cost of infrastructure, the cost of software, and experts who are skilled in the field. Most organizations do not have the internal knowledge to work on the development, fine tuning, and in-life management of any AI-based solutions and it creates a talent shortage, further restricting adoption. Also, it can be cumbersome and resource-demanding to integrate AI into the old systems and workflow, particularly where there is no interoperability standard [42].

Finally, regulatory, legal, and ethical issues also persist. AI as a tool to observe network traffic and user activities attracts the risks of violating data privacy and harboring prejudices or not adhering to regulatory frameworks like GDPR, HIPAA, or NIST guidelines. Increasingly, there is a question of whether algorithms should be accountable when AI systems make decisions that affect the security posture of an organization or privacy rights of an individual. The use of AI in cyber defense is bound to introduce new risks despite reducing pre-existing risk levels unless there are policy guidelines and governance structures to manage all interactions between AI and the cyber defense apparatus [43].

Ultimately, while AI significantly augments cyber defense capabilities, its implementation should be done with caution, ensuring that models are transparent, transparent, robust and in line with general operation and regulatory bodies.

Table 3 Summary of Key Advantages and Limitations of AI in Cyber Defense

Category	Advantages	Limitations/Challenges
Detection Efficiency	Real-time analysis of large-scale log and telemetry data [36]	Imbalanced datasets and rare attack labels hinder generalization [40]
Accuracy	Reduced false positives through behavioral baselining [37]	Adversarial inputs can fool AI models [41]
Proactivity	Predictive analytics enable preemptive threat identification [38]	Black-box models reduce transparency and interpretability [42]
Adaptability	Continuous learning from new threat data [39]	Model drift and retraining requirements create maintenance overhead
Operational Value	Scalable response automation via SOAR/XDR platforms [32–34]	High implementation costs and skill gaps [43]
Governance and Ethics	Enhanced compliance through policy-aware AI agents (emerging)	Privacy, regulatory, and accountability issues under GDPR, NIST, HIPAA, etc. [44]

4. Conclusion

This review examined the evolving role of artificial intelligence (AI) in modern cyber defense, with a particular focus on its applications across detection, prediction, response automation, and threat mitigation. The integration of AI technologies, ranging from supervised learning and unsupervised clustering to deep neural networks and reinforcement learning, has significantly enhanced cybersecurity operations' speed, scale, and precision. By leveraging large-scale threat intelligence, behavioral baselines, and contextual inference, AI systems are increasingly capable of detecting both known and unknown attacks with reduced false positives and greater operational efficiency.

However, it is necessary to note that AI is not a silver bullet. It is highly effective in the use of data-driven detection of anomalies and triage automation, but more so if good data, variable modeling, and overseeing structures are utilized. Trusting the AI models blindly (black-box AI models) can create potential bugs especially when manipulating the

opacity of the system or adding adversarial inputs are used maliciously. Therefore, it is important to establish a sustainable balance between intelligent automation and human-led governance.

Moreover, the broad use of AI in the domain of cybersecurity requires scalable, explainable, and ethically safe deployments. Explanable AI (XAI) is especially important to generate trust of analysts, satisfy the regulatory requirements, and make informed decisions in critical incidents. To counterbalance the increasingly sophisticated nature of cyber threats, the interpretability, generalizability and security of AI models should be made the primary focus of future research and development in order to secure sustainable and reliable cyber defense systems.

Recommendations

To ensure the responsible and effective use of AI in cybersecurity, future work must aim at developing explainable AI (XAI) to promote more transparency and confidence in responding to the automation decision. Benchmarking datasets with standardized and realistic diversity of attacks are also required to introduce and enhance intrusion detection systems based on AI. Also, interdisciplinary cooperation between AI-developing, security, and policymaking is essential to ensure the adjustments of technological advancement to the ethical norm and control. The ability to withstand manipulations and drift in AI models by investing in secure AI architecture will further strengthen defense capabilities against adversarial manipulation and model drift.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Mohamed, N. (2025a). Artificial intelligence and machine learning in cybersecurity: A deep dive into state-of-the-art techniques and future paradigms. Knowledge and Information Systems, 67, 6969–7055. https://doi.org/10.1007/s10115-025-02429-y

[2] Mohamed, N. (2025b). Cutting-edge advances in AI and ML for cybersecurity: A comprehensive review of emerging trends and future directions. Cogent Business & Management, 12(1). https://doi.org/10.1080/23311975.2025.2518496

[3] Talukder, Md. A., Islam, Md. M., Uddin, M. A., Hasan, K. F., Sharmin, S., Alyami, S. A., & Moni, M. A. (2024). Machine learning-based network intrusion detection for big and imbalanced data using oversampling, stacking feature embedding and feature extraction. Journal of Big Data, 11(1). https://doi.org/10.1186/s40537-024-00886-w

[4] Armonk, N. Y. (2023, April 24). IBM launches new QRadar security suite to speed threat detection and response. IBM Newsroom. https://newsroom.ibm.com/2023-04-24-IBM-Launches-New-QRadar-Security-Suite-to-Speed-Threat-Detection-and-Response

[5] Aarness, A. (2025, January 7). XDR vs. SIEM vs. SOAR: What's the difference? Crowdstrike.com. https://www.crowdstrike.com/en-us/cybersecurity-101/next-gen-siem/xdr-vs-siem-vs-soar/

[6] Mannem, K. R. (2025). Human-AI collaboration in DevOps: Enhancing operational efficiency with smart monitoring. European Journal of Computer Science and Information Technology, 13(18), 113–125. https://doi.org/10.37745/ejcsit.2013/vol13n18113125

[7] Nadella, S., Gonaygunta, N. H., Kumar, N. D., & Pawar, P. (2024). Exploring the impact of AI-driven solutions on cybersecurity adoption in small and medium enterprises. World Journal of Advanced Research and Reviews, 22(1), 1199–1197. https://doi.org/10.30574/wjarr.2024.22.1.1185

[8] Kaur, R., Gabrijelčič, D., & Klobučar, T. (2023). Artificial intelligence for cybersecurity: Literature review and future research directions. Information Fusion, 97(101804), 1–29. ScienceDirect. https://doi.org/10.1016/j.inffus.2023.101804

[9] Safitra, M. F., Lubis, M., & Fakhrurroja, H. (2023). Counterattacking cyber threats: A framework for the future of cybersecurity. Sustainability, 15(18), 13369. MDPI. https://www.mdpi.com/2071-1050/15/18/13369

[10] Brandão, P., & Silva, C. (2025). Unveiling the shadows—a framework for Apt's defense AI and game theory strategy. Algorithms, 18(7), 404–404. https://doi.org/10.3390/a18070404

[11] Gan, C., Lin, J., Huang, D.-W., Zhu, Q., & Tian, L. (2023). Advanced persistent threats and their defense methods in industrial internet of things: A survey. Mathematics, 11(14), 3115. https://doi.org/10.3390/math11143115

[12] Diana, L., Dini, P., & Paolini, D. (2025). Overview of intrusion detection systems for computer network security. Computers, 14(3), 87. https://doi.org/10.3390/computers14030087

[13] Le, T. D., Le-Dinh, T., & Uwizeyemungu, S. (2025). Cybersecurity analytics for the enterprise environment: A systematic literature review. Electronics, 14(11). https://doi.org/10.3390/electronics14112252

[14] Ismail, Kurnia, R., Brata, Z. A., Nelistiani, G. A., Heo, S., Kim, H., & Kim, H. (2025). Toward robust security orchestration and automated response in security operations centers with a hyper-automation approach using agentic artificial intelligence. Information, 16(5), 365. https://doi.org/10.3390/info16050365

[15] Berrios, S., Leiva, D., Olivares, B., Allende-Cid, H., & Hermosilla, P. (2025). Systematic review: Malware detection and classification in cybersecurity. Applied Sciences, 15(14), 7747–7747. https://doi.org/10.3390/app15147747

[16] Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. SN Computer Science, 2(3), 1–21. Springer. https://link.springer.com/article/10.1007/s42979-021-00592-x

[17] Fatima, R., Fareed, M. M. S., Ullah, S., Ahmad, G., & Mahmood, S. (2024). An optimized approach for the detection and classification of spam emails using ensemble methods. Wireless Personal Communications, 139(1), 347–373. https://doi.org/10.1007/s11277-024-11628-9

[18] Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. 2010 IEEE Symposium on Security and Privacy, 305–316. https://doi.org/10.1109/sp.2010.25

[19] Pinto, A., Herrera, L.-C., Donoso, Y., & Gutierrez, J. A. (2024). Enhancing critical infrastructure security: Unsupervised learning approaches for anomaly detection. International Journal of Computational Intelligence Systems, 17(1). https://doi.org/10.1007/s44196-024-00644-z

[20] Bagui, S. S., De, S., Mishra, A., Mink, D., Bagui, S. C., & Eager, S. (2025). Detecting cyber threats in UWF-ZeekDataFall22 using K-Means clustering in the big data environment. Future Internet, 17(6), 267–267. https://doi.org/10.3390/fi17060267

[21] Alabadi, M., & Çelik, Y. (2020). Anomaly detection for cybersecurity based on convolutional neural network: A survey. 2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA). https://doi.org/10.1109/hora49412.2020.9152899

[22] Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. Journal of Big Data, 8(1). https://doi.org/10.1186/s40537-021-00444-8

[23] Kuntur, S., Krzywda, M., Wróblewska, A., Paprzycki, M., & Ganzha, M. (2024). Comparative analysis of graph neural networks and transformers for robust fake news detection: A verification and reimplementation study. Electronics, 13(23), 4784–4784. https://doi.org/10.3390/electronics13234784

[24] Khemani, B., Patil, S., Kotecha, K., & Tanwar, S. (2024). A review of graph neural networks: concepts, architectures, techniques, challenges, datasets, applications, and future directions. Journal of Big Data, 11(1). https://doi.org/10.1186/s40537-023-00876-4

[25] Xie, J. (2024). Application study on the reinforcement learning strategies in the network awareness risk perception and prevention. The International Journal of Computational Intelligence Systems/International Journal of Computational Intelligence Systems, 17(1). https://doi.org/10.1007/s44196-024-00492-x

[26] Moric, Z., Dakic, V., & Regvart, D. (2025). Advancing cybersecurity with honeypots and deception strategies. Informatics, 12(1), 14–14. https://doi.org/10.3390/informatics12010014

[27] Alnasser, O., Al Muhtadi, J., Saleem, K., & Shrestha, S. (2025). Signature and anomaly-based intrusion detection system for secure IoTs and V2G communication. Alexandria Engineering Journal, 125, 424–440. https://doi.org/10.1016/j.aej.2025.03.068

[28] Malhotra, P., Singh, Y., Anand, P., Bangotra, D. K., Singh, P. K., & Hong, W.-C. (2021). Internet of things: Evolution, concerns and security challenges. Sensors, 21(5), 1809. https://doi.org/10.3390/s21051809

[29] Shaik, I. A. (2025). Hybrid threat detection systems: A synergistic approach to modern cybersecurity. European Journal of Computer Science and Information Technology, 13(43), 62–69. https://doi.org/10.37745/ejcsit.2013/vol13n436269

[30] IBM. (2017, February 13). IBM Delivers Watson for Cyber Security to Power Cognitive Security Operations Centers. IBM UK Newsroom. https://uk.newsroom.ibm.com/2017-02-13-IBM-Delivers-Watson-for-Cyber-Security-to-Power-Cognitive-Security-Operations-Centers

[31] Darktrace. (2023). Darktrace: "The clear leader in anomaly detection." Darktrace.com. https://www.darktrace.com/news/451-research-calls-darktrace-the-clear-leader-in-anomaly-detection

[32] Brenconin. (2023, April 14). The log and pony show - security orchestration automation response (SOAR). Croninity. https://www.croninity.com/post/the-log-and-pony-show-security-orchestration-automation-response-soar

[33] Shelke, P., & Frantti, T. (2025). Exploring the possibilities of Splunk Enterprise Security in advanced cyber threat detection. International Conference on Cyber Warfare and Security, 20(1), 605–613. https://doi.org/10.34190/iccws.20.1.3326

[34] Networks, P. A. (2025). Brute force investigation—Generic. Cortex XSOAR Documentation. https://xsoar.pan.dev/docs/reference/playbooks/brute-force-investigation---generic

[35] Varshini. (2025, July 18). 10 best XDR solutions in 2025. GBHackers Security | #1 Globally Trusted Cyber Security News Platform. https://gbhackers.com/best-xdr-solutions

[36] Deepwatch. (2025, May 15). Breach intelligence. Deepwatch. https://www.deepwatch.com/glossary/breach-intelligence

[37] IBM. (2024). Artificial intelligence (AI) cybersecurity | IBM. Www.ibm.com. https://www.ibm.com/ai-cybersecurity

[38] Orekha, C. D. I. P. O. (2024). Predictive cyber defense: Harnessing AI and ML for anticipatory threat mitigation. International Journal of Research Publication and Reviews, 5(9), 3122–3132. https://doi.org/10.55248/gengpi.5.0924.2669

[39] Ring, M., Wunderlich, S., Scheuring, D., Landes, D., & Hotho, A. (2019). A survey of network-based intrusion detection data sets. *Computers & Security*, *86*, 147–167. https://doi.org/10.1016/j.cose.2019.06.005

[40] Amato, F., Cirillo, E., Fonisto, M., & Moccardi, A. (2024). Detecting adversarial attacks in IoT-enabled predictive maintenance with time-series data augmentation. *Information*, *15*(11), 740. https://doi.org/10.3390/info15110740

[41] Saarela, M., & Podgorelec, V. (2024). Recent applications of explainable AI (XAI): A systematic literature review. *Applied Sciences*, *14*(19), 8884–8884. https://doi.org/10.3390/app14198884

[42] Naseer, I., Akram, S., Masood, T., Rashid, M., & Jaffar, A. (2023). Lung cancer classification using modified U-Net-based lobe segmentation and nodule detection. *IEEE Access*, *11*, 60279–60291. https://doi.org/10.1109/access.2023.3285821

[43] Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. *Science*, *361*(6404), 751–752. https://doi.org/10.1126/science.aat5991

[44] Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., & McGuinness, L. A. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *British Medical Journal*, *372*(71). https://doi.org/10.1136/bmj.n71

[45] Talaei Khoei, T., & Kaabouch, N. (2023). A comparative analysis of supervised and unsupervised models for detecting attacks on the intrusion detection systems. *Information*, *14*(2), 103. https://doi.org/10.3390/info14020103