Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Application of AI and ML in Cybersecurity Operations: Challenges and Opportunities

Giddeon Angafor

g.angafor@greatermanchester.ac.uk

University of Greater Manchester

Research Article

Keywords: Artificial Intelligence, Machine Learning, Cybersecurity, Intrusion Detection, Threat Intelligence, Explainable AI, Adversarial Machine Learning

Posted Date: August 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10450278/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Artificial Intelligence (AI) and Machine Learning (ML) are being increasingly integrated into cybersecurity operations to enhance threat detection, automate incident responses, and improve overall resilience. Despite their advantages, some studies suggest that adding AI/ML to cybersecurity systems introduces new risks, such as adversarial attacks, data quality issues, and operational difficulties. This article examines the current landscape of AI/ML applications in cybersecurity. It highlights key challenges and opportunities and offers a structured analysis based on existing literature and empirical observations. Results indicate that while AI/ML significantly improve detection accuracy and operational efficiency, their success depends on strong data governance, model transparency, and ongoing oversight. The study concludes with recommendations for the secure and responsible deployment of AI/ML systems within cybersecurity environments.

1. Introduction

Cybersecurity operations face increasing pressure as modern digital systems become more complex and threat actors employ increasingly sophisticated tactics to deliver cyber threats. Research by [1] and [2] highlights that organisations across various sectors have adopted and continue to undergo digital transformation, which increases their reliance on interconnected systems, cloud services, and emerging technologies, including the Internet of Things (IoT). However, they argue that while these developments offer benefits such as operational efficiency and flexibility, they also expose organisations to numerous vulnerabilities. Key challenges related to digital transformation include, but are not limited to, a rise in both the frequency and sophistication of cyberattacks. [2] notes that threats like ransomware, advanced persistent threats (APTs), and insider attacks are prompting organisations to modify their security strategies. Some researchers contend that this trend has created an environment where manual analysis and static, rule-based detection methods are no longer adequate. For example, [3] emphasise that traditional signature and rule-based security detection tools struggle to identify advanced threats such as zero-day exploits, polymorphic malware, and other APTs, mainly because these attacks are designed to evade predefined detection techniques and adapt in real time to defensive measures.

AI and ML have become potent tools for future cybersecurity. Their capacity to learn from vast datasets, detect subtle anomalies, and perform predictive analytics helps security teams identify emerging threats earlier and automate complex decisions. Some studies show that ML-based systems can enhance malware detection, network traffic classification, and anomaly detection by revealing behavioural patterns that are difficult for human analysts to notice in real-time [3]. These capabilities are especially valuable in environments where data volumes and attack speeds surpass human capacity. This view is supported by [4], who state that ML and deep learning methods for malware detection and classification consistently outperform traditional signature-based and heuristic techniques, particularly against obfuscated and polymorphic malware. They achieve this by capturing complex relationships between features that are not obvious during manual inspection. They also highlight that behaviour-based models using execution logs and system calls enable near real-time detection of new malware families by

generalising from learned patterns of malicious activity rather than relying solely on known signatures [4].

Conversely, incorporating AI/ML into cybersecurity presents notable risks and operational challenges. Research by [5] emphasises that ML models can become targets of adversarial manipulation, including evasion tactics, poisoning attacks, and backdoor insertions, all of which weaken model reliability and trustworthiness. Some studies argue that adversarial machine learning has become a significant concern, as attackers increasingly exploit model vulnerabilities to bypass detection or manipulate training data. For instance, [6] agrees that although AI models enhance threat detection, they can also introduce new attack surfaces, such as adversarial manipulation of training data, model poisoning, and exploitation of model blind spots.

Beyond technical risks, some studies have identified ethical concerns such as bias, transparency, and accountability. These issues, they argue, complicate deployment decisions, especially in high-stakes security contexts where automated decisions can impact critical infrastructure or personal privacy [7]. [6] agree that AI systems often lack transparency, making it difficult for operators to understand or validate decisions in safety-critical environments. They emphasise that this lack of explainability can hinder incident response and increase operational risks.

Considering these opportunities and challenges, the main goal of this article is to explore how cybersecurity teams can strategically and responsibly incorporate AI/ML technologies into their operations. This involves analysing best practices for model governance, adversarial resilience, human-machine collaboration, and ethical risk management. By synthesising current research and practical insights, the article aims to provide a guide for utilising AI/ML to enhance cybersecurity efforts while effectively managing the associated risks.

1.1 Background

The rise of digital transformation has significantly expanded organisational attack surfaces across sectors such as finance, healthcare, and critical infrastructure. A recent survey, see [8] on safeguarding critical infrastructures, highlights that digital transformation, especially AI integrations, remote access, and data-driven automation, has widened attack surfaces in energy, transport, healthcare, and other essential systems. It emphasises that increased connectivity, particularly through legacy system integrations and third-party digital services, creates new pathways for intrusion and lateral movement, requiring more advanced security controls and continuous monitoring. Supporting this are other studies, which argue that such expansions have increased the volume, speed, and diversity of cyber assets requiring monitoring and protection, making the operational environment more complex for defenders [9]. The rapid adoption of cloud-native architectures, Internet of Things (IoT) ecosystems, and distributed computing has further heightened this complexity, as interconnected systems introduce new dependencies and potential points of failure [2]. Consequently, Security Operations Centre (SOC) teams face growing analytical workloads, including reduced visibility across environments. They also

experience higher expectations from executive leadership, who now consider cybersecurity a strategic priority at the board level. This push has driven organisations to embed AI/ML tools into IDS, EDR, and SIEM/SOC workflows, covering everything from network and host intrusion detection, endpoint behavioural analysis, to log analytics, alert triage, and threat intelligence enrichment, aimed at scaling detection and response while reducing analysts' workload.

Similarly, adversaries are utilising automation, especially AI and ML tactics, to enhance the scale, precision, and versatility of cyberattacks. Early analyses of automated threat activities demonstrated how attackers employ scripting and scanning tools to speed up reconnaissance and vulnerability detection [10]. Moreover, some research indicates that AI-powered offensive capabilities have advanced significantly beyond these initial stages, enabling adversaries to develop polymorphic malware, automate exploit creation, and execute large-scale credential stuffing or phishing campaigns with unparalleled efficiency [11].

An increasing amount of evidence shows that AI is now directly involved in social engineering and fraud schemes. A study by [12] on AI-enabled WhatsApp scams reveals how attackers use generative models to craft highly personalised messages that imitate the linguistic patterns of trusted contacts and automate the creation of convincing multimedia content, such as voice notes or deepfake videos, to extort money from unsuspecting WhatsApp users. These AI-driven scams significantly improve the success rate of social engineering attacks by exploiting human trust and reducing detectable inconsistencies that traditional awareness training relies on. Similar research suggests that generative AI is being used to produce synthetic identities, automate scam call centres, and develop realistic phishing websites that adapt dynamically to user behaviour [13, 14].

Beyond social engineering, ML-enabled offensive tools are becoming more capable of bypassing traditional rule-based defences. Adversarial ML techniques allow attackers to manipulate input data to evade detection systems, poison training datasets, or reverse-engineer model behaviour to uncover blind spots [15]. Automated vulnerability discovery systems, driven by reinforcement learning and large-scale code analysis, enable attackers to find exploitable weaknesses more quickly than human analysts, thereby reducing defenders' response times and challenging the effectiveness of static detection logic [16, 11].

Collectively, these developments highlight a significant shift, demonstrating that AI and ML are no longer merely defensive tools but also serve as powerful enablers of offensive cyber operations. As adversaries incorporate AI into reconnaissance, exploitation, and deception workflows, organisations must update their defensive strategies, recognising that traditional rule-based systems and manual analysis are no longer sufficient, particularly against AI-accelerated threats. In this context, [9] note that AI and ML have become essential tools for modern cyber defence, especially in environments with large amounts of high-dimensional data. Early studies emphasised both the potential and limitations of ML-based intrusion detection, focusing on issues such as evasion, dataset bias, and operational deployment [10]. Subsequent research has systematically linked AI techniques, such as supervised and unsupervised

learning, deep learning, and reinforcement learning, to key aspects of cybersecurity, including anomaly detection, malware classification, phishing detection, and network traffic analysis [17, 18, 11]. Recent surveys, see [11] and [19], suggest that most AI applications in SOCs focus on detection, with fewer involved in response, recovery, and resilience engineering, often relying heavily on non-explainable models and proprietary datasets.

Modern SOCs are increasingly adopting AI-driven platforms for threat intelligence integration, behavioural analytics, and endpoint protection to reduce analyst workload, improve detection accuracy, and enable near-real-time responses [9, 11]. Studies have shown that cognitive SOC concepts incorporate AI across the NIST Cybersecurity Framework functions, which include identify, protect, detect, respond, and recover, using automated alert correlation, user and entity behaviour analytics (UEBA), and adaptive playbooks to coordinate responses [19]. Similarly, the literature highlights ongoing challenges, particularly regarding explainability and trust in AI decisions, as well as integration with legacy tools. Other areas of misalignment involve data quality issues, such as labelling, and the risk of over-reliance on opaque models in critical environments [10, 17, 9]. These issues position AI-enabled SOCs not as replacements for human expertise but as socio-technical systems where automation supports human analysts, provided governance, transparency, and rigorous evaluation practices are maintained.

1.2 Research Questions

1. How are AI and machine learning presently utilised in cybersecurity operations and SOC workflows?
2. What challenges and limitations impact the successful implementation and performance of AI and ML-based cybersecurity solutions in real-world settings?
3. How can AI and ML be utilised to improve cybersecurity resilience, and what practices promote their secure and responsible deployment?

2. Literature Review

Research on AI and ML in cybersecurity has experienced a major shift over the past forty years, changing how organisations detect, analyse, and respond to cyber threats. Early foundational work established anomaly detection as a key defensive tactic, using statistical profiling and rule-based systems to spot deviations from expected behaviour [20]. Although innovative, these systems had high false-positive rates and depended on strict modelling assumptions, which limited their practical use.

By the early 1990s, improvements in computing power enabled the development of neural networks and pattern-recognition techniques, which provided greater adaptability and the capacity to learn complex relationships within network traffic [21]. These early ML-based systems signified a shift from static rule sets to data-driven detection, laying the foundation for modern intrusion detection systems (IDS).

From the mid-2000s onwards, the rise of big data, cloud computing, and distributed architectures sped up the adoption of ML in cybersecurity. Research demonstrated that supervised and unsupervised learning could enhance malware classification, spam detection, and network anomaly detection [10]. The arrival of deep learning in the 2010s further expanded capabilities, allowing models such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs) to detect subtle temporal and spatial patterns in network telemetry [22].

The rise of deep learning marked an important moment in the field. Studies indicate that deep neural networks, especially CNNs and recurrent neural networks (RNNs), outperform others in malware classification, intrusion detection, and encrypted traffic analysis [23, 24]. [25] showed that malware binaries could be converted into greyscale images, enabling CNNs to detect malicious patterns with high accuracy. This image-based approach has since been expanded to large malware datasets and incorporated into commercial endpoint protection systems. Likewise, autoencoders and LSTM networks have been used for anomaly detection in network flows, capturing temporal dependencies often missed by traditional statistical models [26].

Beyond standalone intrusion detection systems, several studies describe the integration of ML components into Security Information and Event Management (SIEM), Endpoint Detection and Response (EDR), and Security Orchestration, Automation and Response (SOAR) platforms. In these architectures, ML models assist with alert correlation across diverse data sources, risk-based prioritisation, and automated enrichment with contextual information such as threat intelligence, asset criticality, and user roles [27, 28].

Modern AI-powered SIEM platforms utilise machine learning-based correlation engines that combine logs, alerts, and threat intelligence feeds to reduce noise and highlight high-risk events [18, 17]. Behavioural analytics systems, often based on unsupervised learning techniques like clustering and isolation forests, detect deviations from normal user behaviour and have become essential for insider threat detection [29, 30]. These systems address the challenge of identifying low-and-slow attacks that evade signature-based tools. Table 1 below summarises some research illustrating how AI and ML workflows have been historically applied to improve threat detection, automate incident response, and boost overall resilience by SOCs in IDS, EDR, and SIEM applications.

Table 1: Examples of application of AI/ML in cybersecurity operations and SOC workflows.

AI/ML Model/Tool	Where / how it is applied (IDS, EDR, SIEM, SOC)	Author(s) & Date
Supervised ML for network intrusion detection (e.g. Random Forest, SVM, Gradient Boosting).	Used in network IDS to classify flows or packets as benign/malicious; commonly integrated into SOC workflows to reduce manual triage and improve detection of novel attacks compared with signature-based IDS.	[31], [9]
Deep learning IDS (CNN/LSTM, autoencoders).	Deployed as a host or network-based IDS to learn complex temporal and spatial patterns in traffic and system events; often embedded in SOC pipelines for high-volume telemetry, enabling automated anomaly detection and enrichment of SIEM alerts.	[31], [9]
Behaviour-based EDR models (supervised / unsupervised ML on process, file, registry activity).	Applied in EDR platforms to detect malware, lateral movement, and living-off-the-land techniques by modelling normal endpoint behaviour and flagging deviations; outputs feed SOC investigations and automated response playbooks.	[1], [9]
Unsupervised log clustering and anomaly detection for SIEM (k-means, DBSCAN, isolation forest).	Used across SOC/SIEM workflows to automatically score and prioritise alerts based on features such as asset criticality, historical incidents, and threat intelligence, improving mean-time-to-detect and response.	[31], [9]
ML-based alert triage and prioritisation (classification/ranking models).	Used across SOC/SIEM workflows to automatically score and prioritise alerts based on features such as asset criticality, historical incidents, and threat intelligence, improving mean-time-to-detect and response.	[31], [9]
AI-driven threat intelligence enrichment (NLP, clustering, knowledge graphs).	Applied in SOC threat intelligence and SIEM correlation to extract indicators of compromise from text sources, link them to internal telemetry, and automatically enrich IDS/EDR alerts with contextual information.	[31] [9]

Despite these advances, the literature emphasises several significant vulnerabilities and limitations. Research in adversarial machine learning shows that ML models can be manipulated via carefully crafted inputs, causing misclassification or evasion [32, 33]. This raises concerns about applying ML models in high-stakes environments where attackers actively seek to exploit weaknesses. Also, bias in training datasets, resulting from imbalanced classes, outdated attack samples, or synthetic data, can impair model performance and cause systematic errors [35].

Furthermore, the opacity of many deep learning models limits interpretability, making it difficult for analysts to understand or verify automated decisions [36]. Studies such as [37] emphasise that deep

learning models are often opaque, with complex decision boundaries that are hard to interpret. This opacity makes it difficult for analysts to understand why a model accepted an adversarial input or misclassified a malicious event, thereby restricting forensic analysis and trust in automated decisions.

Other academic reviews related to adversarial ML, such as [35], [38], and [39], explicitly list interpretability and explainability as open challenges. They argue that analysts and those charged with defending cyber threats find it difficult to examine internal representations or verify that models have learned robust, security-relevant features rather than spurious patterns that attackers can exploit. Consequently, organisations face a dual risk in security operations. This is because models can be manipulated to bypass defences, and their internal logic remains too opaque for thorough auditing, requiring additional monitoring, human oversight, and explainability tools around AI-driven security systems. Explainable AI (XAI) techniques have been suggested to bridge this gap, but adoption continues to be inconsistent across cybersecurity tools.

These studies emphasise that, while AI and ML significantly improve detection and response capabilities, they also introduce new risks. Effective governance frameworks, covering model evaluation, adversarial robustness, data quality, and human oversight. These are crucial to ensure that AI-enabled cybersecurity systems remain trustworthy, resilient, and aligned with organisational risk appetite.

3. Methods

This study uses a qualitative research design approach to explore how AI and ML are applied in contemporary security operations, and why these technologies offer both new defensive advantages and new risks, particularly adversarial manipulation, and limited interpretability. A qualitative approach is appropriate because the research aims to understand practices, challenges, and organisational settings rather than quantify isolated technical performance metrics. Qualitative methods are frequently used in cybersecurity research when analysing socio-technical systems such as SOC workflows, organisational risk practices, and analyst decision-making [37].

The study reviews peer-reviewed journals, conference papers, and industry reports published between 2010 and 2026, a period marked by the rapid adoption of AI/ML in cybersecurity tools and the rise of adversarial machine learning. Analysing this extensive body of knowledge allows the study to trace the development of AI-driven security technologies, from early anomaly-based IDS models to modern deep learning-powered EDR and SIEM platforms, while also highlighting ongoing concerns about model robustness and interpretability [39].

An analysis was carried out on AI/ML-driven cybersecurity tools, focusing on IDS, where ML models classify network traffic or identify anomalies; EDR, where behavioural ML models detect malicious processes or lateral movement; and SIEM, where ML supports log clustering, alert triage, and threat intelligence correlation. Figure 1 below presents a graphical overview of the methodology applied in the study.

These categories were selected because they represent the fundamental analytical layers of modern SOC workflows, and because adversarial ML research shows that models within these layers are particularly vulnerable to evasion attacks. For example, adversarial perturbations can cause IDS models to misclassify malicious traffic as benign or trick EDR behavioural models into ignoring malicious processes [1]. Comparing these tools helps determine where AI/ML truly provides operational value and where it introduces exploitable vulnerabilities.

To complement the literature review, the study synthesises documented case studies of SOC implementations across the financial, healthcare, and government sectors. These sectors were selected because they manage high-value systems, face sophisticated adversaries, and have adopted AI/ML at scale. Case studies provide insights into real-world deployment challenges, such as alert fatigue, model drift, adversarial evasion, and difficulties in validating opaque deep learning decisions, which are often underrepresented in purely technical research [37].

All data collected were analysed using thematic coding, a qualitative method that identifies recurring patterns across diverse sources. Thematic coding is particularly suitable for cybersecurity research because it enables the synthesis of technical, organisational, and operational insights into coherent themes. In this study, coding uncovered cross-cutting themes such as reliance on ML for high-volume detection tasks, the vulnerability of ML models to adversarial manipulation, the limited interpretability of deep learning systems, and the need for human-in-the-loop oversight in SOC environments.

By integrating a scoping literature review, comparative technical analysis, and cross-sector case studies, this methodology provides a rigorous and comprehensive basis for assessing how AI/ML enhances and complicates cybersecurity operations.

4. Results

The analysis revealed several key insights into the effectiveness, limitations, and operational impacts of AI/ML-enabled cybersecurity systems. These findings align with trends reported in recent peer-reviewed studies and industry assessments.

4.1 Enhanced Detection Capabilities

Enhanced detection capabilities in AI-driven IDS do more than just reduce false positives. Studies have shown that deep learning architectures such as CNNs, RNNs, autoencoders, and hybrid models consistently outperform traditional methods in identifying both known and zero-day attacks. This is primarily because of their ability to learn complex behavioural patterns instead of relying on static signatures. A recent study by [40] supports this. It offers a thorough review demonstrating that deep learning-based IDS achieve notably higher detection accuracy across benchmark datasets such as KDD'99, NSL-KDD, and UNSW-NB15, with CNN and autoencoder models surpassing traditional IDS in both precision and recall metrics. These architectures, according to the study, excel at modelling

high-dimensional network traffic, enabling them to detect subtle anomalies that signature-based systems often miss.

Similarly, it has been observed that AI/ML hybrid models for cybersecurity further improve detection capabilities. Research by [41] discusses a deep learning-based hybrid IDS that combines anomaly and misuse detection. Results of the study show significant improvements in recognising zero-day attacks, with the hybrid model achieving higher true-positive rates and lower false-negative rates compared to standalone ML or signature-based systems.

Furthermore, [42] contended that integrating deep learning into IDS architectures allows for real-time threat mitigation, particularly in environments with rapidly evolving attack methods. These systems remain adaptable to new threat behaviours without requiring manual signature updates, making them more resilient against polymorphic malware and adversarial evasion strategies.

4.2 Operational Efficiency

Operational efficiency in modern SOCs has been greatly enhanced by integrating AI-driven automation, especially in alert triage, event correlation, and incident response. AI and ML-enabled triage systems lessen analyst workload by automatically classifying alerts, enriching contextual information, and escalating high-risk events. These capabilities enable human analysts to concentrate on complex investigations rather than repetitive tasks. This aligns with findings from multiple peer-reviewed studies showing that automation directly improves SOC throughput and reduces operational fatigue [43].

A review of AI-enabled SOCs [44] reports that automation is one of the most widely adopted AI use cases, with organisations leveraging AI and ML to streamline detection, triage, and response workflows. These AI-enabled security operations systems reduce manual intervention by correlating disparate events, suppressing low-value alerts, and prioritising incidents based on behavioural risk scoring. Similarly, a study on the rise of cognitive SOCs by [43] emphasises that AI-driven orchestration improves analyst productivity by reducing alert fatigue and enabling faster decision-making, particularly in high-volume environments.

Further evidence from research by [9] demonstrates that AI-based automation improves efficiency within the NIST Cybersecurity Framework, particularly in detection and response. A systematic review of 189 studies found that AI substantially lowers operational overhead by automating repetitive SOC tasks, reducing mistakes, and speeding up containment processes [9]. The review also highlighted that SOCs increasingly depend on AI to handle expanding attack surfaces and growing alert volumes, with automation helping to ease staffing pressures and reduce overtime burdens.

Findings like these reinforce earlier studies suggesting that automated incident response procedures reduce containment times and enhance overall SOC responsiveness [29, 30]. This shows that, when combined with orchestration platforms, AI-enabled security solutions can improve manual alert management, allowing SOC teams to devote more time to threat hunting, forensic analysis, and strategic defence improvements.

4.3 Adversarial Vulnerabilities

Despite performance improvements, ML models remain susceptible to adversarial manipulation. Research on adversarial examples shows that attackers can subtly modify inputs, such as network packets or malware binaries, to evade detection without altering their malicious functions [33, 34]. Deep learning-based IDS systems are especially vulnerable because of their high dimensionality and reliance on learned representations. Poisoning attacks, where adversaries corrupt training data, further threaten the integrity of models and can decrease detection accuracy over time.

4.4 Data Dependency

Model performance heavily relies on the quality, diversity, and representativeness of training data. Organisations with limited telemetry or fragmented logging infrastructures find it challenging to achieve reliable detection accuracy. This is because models trained on narrow datasets often struggle to generalise to real-world environments [35]. Studies emphasise that ML models require ongoing retraining with new, high-quality data to remain effective against evolving threats [17]. Data scarcity also increases the risk of bias, leading to inconsistent or unreliable predictions.

4.5 Explainability Gaps

A recurring challenge noted in the analysis is the lack of transparency in AI/ML-based security systems. Many deep learning models function as “black boxes,” making it difficult for analysts to understand why a particular alert was generated or how a model arrived at its conclusion [36]. This lack of explainability complicates incident validation, diminishes analyst trust, and raises compliance issues for regulated industries that demand auditability and justification of security decisions. Although explainable AI (XAI) techniques are available, adoption remains limited across operational SOC environments.

5. Discussion

The findings reveal a complex and evolving landscape where AI/ML systems both enhance and complicate cybersecurity efforts. This duality is often highlighted in modern research, emphasising both the transformative potential of AI-enabled SOCs and the new types of risks they introduce.

5.1 Transformative Capabilities and SOC Enhancement

AI/ML systems significantly improve SOC capabilities by boosting detection accuracy, reducing false positives, and enabling large-scale automation. Deep learning models, such as CNNs, LSTMs, and autoencoders, have demonstrated superior performance in identifying novel threats, encrypted malicious traffic, and behavioural anomalies that traditional signature-based tools often miss [18, 24]. These advancements are especially beneficial in high-volume environments where human analysts struggle to manually manage extensive telemetry.

Automation further boosts SOC efficiency as AI-driven triage, alert correlation, and automated response processes lessen analyst fatigue and accelerate containment times. Studies indicate that automated orchestration and response can reduce manual alert handling by up to half, allowing analysts to focus on strategic investigations rather than repetitive tasks [29, 30]. This operational enhancement is essential for organisations dealing with staffing shortages and increasingly complex threat environments.

5.2 Adversarial Machine Learning as A Persistent and Growing Threat

Adversarial vulnerabilities remain one of the most significant limitations of machine-learning-based IDS. Some studies have found that even highly accurate deep learning models can be manipulated through carefully crafted adversarial inputs, enabling attackers to bypass detection while maintaining the malicious functionality of their payloads. Foundational work by [33] and [34] demonstrates that small, often imperceptible perturbations can cause ML models to misclassify malicious traffic as benign, exposing a critical weakness in systems that rely heavily on learned representations.

Some studies suggest that these risks are particularly significant in cybersecurity settings. [45] emphasised that adversarial threats targeting network traffic can exploit the high dimensionality of IDS feature spaces, causing deep learning models, especially CNNs and LSTMs, to misinterpret malicious flows. Their experiments showed that even slight feature manipulation can reduce detection rates by over 50%, highlighting the vulnerability of deep learning-based IDS against adaptive adversaries.

Building on that, studies have indicated that poisoning attacks pose an equally serious threat. In poisoning scenarios, attackers introduce malicious or misleading samples into training datasets, gradually corrupting the model's decision boundaries. [46] described poisoning as one of the most damaging forms of adversarial manipulation because it undermines the integrity of the learning process itself. [47] agree that poisoning can diminish IDS performance over time, lowering detection accuracy and increasing false negatives, especially in systems that depend on continuous retraining with live network data. Their research shows that even a small proportion of poisoned samples, such as 3–5%, can significantly distort anomaly detection models.

Some studies also highlight the problem of transferability, where adversarial examples created for one model can successfully bypass others. [48] emphasise that attackers can manipulate behavioural features used in ML-based malware detection, allowing evasion across various IDS architectures. This transferability increases the risk because attackers do not need precise knowledge of the target model to carry out successful adversarial attacks.

5.3 Data Governance as a Foundational Requirement

Data governance has increasingly been recognised as a fundamental prerequisite for effective AI/ML deployment in cybersecurity. While advanced models can detect complex threats, their reliability is chiefly limited by the quality, diversity, and consistency of the data they learn from. Consequently, poor-quality, narrow, or biased datasets lead directly to reduced generalisation, unstable performance,

and even discriminatory or misleading outputs, as highlighted by [35] in early work on ML system failures.

Several studies emphasise that organisations with fragmented telemetry, such as inconsistent logging, incomplete packet capture, or siloed monitoring tools, struggle to maintain stable model performance. For example, [47] noted that ML-based intrusion detection systems often failed when trained on datasets that do not reflect real-world traffic variability, causing models to overfit to outdated attack patterns. This challenge, they argued, is amplified in dynamic environments where threat behaviours evolve rapidly, making continuous data refresh and validation essential.

Other research, such as [17], established that data governance directly affects model resilience. They argue that high-quality, well-curated datasets are the strongest predictors of detection accuracy across ML-based IDS. Their review highlights that models trained on diverse, representative datasets consistently outperform those trained on narrow or synthetic data, especially in detecting novel or low-frequency attack types. Similarly, [49] demonstrated that dataset heterogeneity, such as covering multiple network topologies, user behaviours, and threat categories, significantly enhances anomaly detection robustness.

Beyond collection, governance includes labelling, validation, and ongoing curation, which are all vital for operational reliability. Studies on SOC automation indicate that inaccurate or inconsistent data labelling can spread errors into automated triage systems, resulting in misclassification chains that raise false positives or hide genuine threats [50]. Effective governance structures reduce these risks by implementing structured labelling procedures, regular dataset audits, and mechanisms for monitoring data drift.

Finally, some studies have emphasised the strategic importance of governance in protecting against adversarial threats. Poorly governed datasets are more vulnerable to poisoning attacks, where adversaries insert misleading samples to distort model behaviour. Research by [51] demonstrates that robust data validation pipelines greatly decrease the success rate of poisoning attempts, underlining the need for governance as a security measure, not just a data management practice.

5.4 Explainability and Trust in AI-Driven Security

Like governance, explainability, and trust have become key concerns in the operational deployment of AI-driven security systems. Although these models provide powerful detection capabilities, their often-opaque internal logic frequently prevents analysts from understanding why a particular alert was triggered. This “black-box” nature, according to some scholars, weakens trust, complicates incident validation, and poses compliance challenges, particularly in regulated sectors that require auditability and justification for security decisions [36].

Several studies have demonstrated that explainability is not just a usability feature; it is also a fundamental requirement for safe and responsible AI in cybersecurity. This clarifies why [52] contend that interpretability is vital for ensuring AI systems behave reliably during distribution shifts, adversarial

conditions, or novel attack patterns. They maintain that without transparency, analysts cannot determine whether a model's output is correct, biased, or the result of adversarial manipulation. This is especially problematic in SOC environments where false positives and false negatives have significant operational consequences.

Studies focusing specifically on AI and ML for cybersecurity highlight similar concerns. For example, [53] demonstrate that analysts are less likely to act on AI-generated alerts when the underlying reasoning is unclear, leading to slower response times and reduced SOC efficiency. Their research shows that explainability directly influences analyst trust and adoption, making it a critical factor in real-world deployment. Similarly, [54] emphasise that explainability is essential for regulatory compliance, especially in sectors such as finance, healthcare, and critical infrastructure, where automated decisions must be auditable.

Emerging explainable AI (XAI) techniques, such as SHAP, LIME, counterfactual explanations, and saliency maps, offer promising avenues for enhancing transparency. These methods assist analysts in interpreting model behaviour by emphasising feature importance, local decision boundaries, or hypothetical alternative outcomes [55]. In cybersecurity contexts, XAI can identify which packet features, behavioural indicators, or log attributes contributed to a detection event, enabling more informed triage and validation.

However, studies also emphasise that XAI adoption in cybersecurity remains limited. Some, such as [56], argue that LIME and SHAP, although powerful, introduce computational overheads that can be prohibitive in real-time detection systems. They state that integrating XAI into complex pipelines, such as deep learning-based IDS or automated SOC orchestration platforms, requires considerable engineering efforts and vendor support, which many commercial tools currently lack. Furthermore, they contend that some XAI methods struggle with high-dimensional network data, reducing their interpretability or producing explanations that analysts find difficult to act upon.

Consequently, explainability gaps remain, limiting the practical effectiveness of AI-driven detection systems in high-stakes contexts. The literature increasingly advocates for domain-specific XAI approaches tailored to cybersecurity, enhanced human-AI interaction design, and regulatory frameworks that require transparency in automated security decisions.

5.5 Ethical and Societal Considerations

Ethical and societal issues are increasingly recognised as key factors in whether AI-driven cybersecurity systems can be responsibly implemented. As these systems develop the capacity to analyse behavioural patterns, link user activities, and make semi-autonomous security choices, they pose complex ethical challenges concerning privacy, surveillance, fairness, and accountability. A wide range of academic literature, such as [57] and [17], consistently warns that without clear safeguards, AI-powered security tools may inadvertently infringe on civil liberties or perpetuate structural biases, especially in high-stakes sectors like finance, healthcare, and critical infrastructure.

Privacy concerns are among the most widely discussed issues. This is because AI-driven monitoring systems often require extensive telemetry, ranging from network flows and authentication logs to user behaviour analytics, which can inadvertently expose sensitive personal information. A study by [58] highlights that such systems may conflict with data protection frameworks like GDPR, especially when they lack clear purpose limitation, minimisation, or transparency mechanisms. Similarly, [59] argued that pervasive behavioural monitoring risks normalising surveillance practices that erode user autonomy and trust are dangerous.

In a similar vein, algorithmic bias presents another significant challenge. Bias can originate from skewed training data, flawed feature engineering, or systemic inequalities embedded in organisational telemetry. Some studies indicate that biased models may disproportionately flag certain user groups, departments, or geographic regions as anomalous, leading to unequal monitoring or higher false-positive rates. [60] demonstrate that biased AI systems can amplify existing organisational inequities, especially when deployed without fairness auditing or ongoing bias assessment. In cybersecurity contexts, this may result in unfair scrutiny of specific user populations or the misclassification of legitimate activity as malicious.

Beyond concerns of privacy and bias, scholars highlight wider societal risks. [61], for instance, warns that opaque algorithmic decision-making can generate “weapons of math destruction”, which she describes as systems that make significant decisions without accountability or recourse. In cybersecurity, this occurs when AI systems autonomously deny user access, quarantine assets, or escalate incidents without offering an understandable justification. Such actions can disrupt business operations, erode trust, and result in compliance liabilities.

To mitigate these risks, some studies emphasise the importance of strong governance frameworks. This includes ethical review procedures, model documentation, transparency reports, and compliance with legal standards such as GDPR, NIS2, and sector-specific regulations. [62] argue that responsible AI deployment involves embedding ethical principles, such as fairness, accountability, and respect for human autonomy, into the design and operation of cybersecurity systems. Without these safeguards, organisations risk deploying tools that inadvertently harm users, breach regulatory requirements, or erode societal trust in digital security practices.

6. Conclusion

AI and ML have become vital components of modern cybersecurity efforts, offering significant improvements in detection accuracy, automation, and predictive threat analysis. Their ability to process large volumes of complex data enables organisations' security teams, such as the SOC, to identify subtle anomalies, connect complex events, and respond to threats more quickly and precisely than traditional systems [18, 24]. When used effectively, these technologies can greatly enhance organisational resilience amid an increasingly complex and rapidly evolving threat landscape.

However, the results also show that adopting AI/ML introduces new risks that must be carefully managed. Adversarial machine learning poses a direct threat to model reliability, as attackers can manipulate inputs or poison training data to cause misclassification or degrade system performance [33, 34]. These risks emphasise the need for robust model evaluation, continuous monitoring, and defensive strategies such as adversarial training and secure data pipelines.

Data governance is a vital foundation for successful AI/ML deployment. Without diverse, representative, and high-quality datasets, models can develop bias, poor generalisation, or unreliable outputs [35]. Organisations must therefore dedicate resources to ongoing data collection, labelling, validation, and curation to ensure models stay accurate and resilient as threats evolve.

Explainability remains a crucial challenge, especially as Black-box models erode analyst trust, complicate incident validation, and pose compliance risks in regulated sectors [36]. Although techniques like SHAP and LIME show promise for enhancing interpretability, their integration into commercial cybersecurity tools remains limited. Developing explainable AI will be vital for ensuring accountability and transparency in AI-driven security operations.

Finally, ethical concerns such as privacy, surveillance, and algorithmic bias must be addressed to promote responsible deployment. AI systems that analyse user behaviour or network activity should include strong privacy safeguards and governance frameworks to prevent unintended harm [57].

6.1 Future Research Directions

The evolving role of AI/ML in cybersecurity underscores several key areas for future research. One such area is developing adversarial resilience in models. Creating models that can withstand evasion and poisoning attacks remains essential, and research should focus on adversarial training, defensive distillation, and real-time detection techniques. Another important area is advanced explainability methods, where there is a need for XAI techniques specifically tailored for cybersecurity, capable of providing actionable insights without overburdening analysts. Future work could aim at scalable, low-overhead interpretability tools. Additionally, standardised evaluation frameworks are lacking. The industry needs agreed-upon benchmarks, datasets, and testing protocols to enhance transparency and comparability among different vendors. An important area for future research is ethical AI and privacy preservation. This includes exploring privacy-enhancing technologies such as federated learning, differential privacy, and secure multi-party computation to reduce risks associated with behavioural monitoring. Lastly, research into human-AI collaboration models is vital. Understanding how analysts interact with AI systems and designing interfaces that promote effective collaboration will be crucial for maximising operational value.

With careful implementation, robust governance, and ongoing research, AI and ML can play a transformative role in enhancing cybersecurity resilience and supporting organisations in defending against increasingly sophisticated threats.

Declarations

Author Contribution

This article is a single-author manuscript. As sole author, I conceived the research problem, designed the study, carried out all analyses, and interpreted the results. I drafted the manuscript and completed all revisions. I am fully responsible for the accuracy and integrity of the work.

References

1. Mohamed N (2025) Artificial intelligence and machine learning in cybersecurity: a deep dive into state-of-the-art techniques and future paradigms. Knowl Inf Syst 67(8):6969–7055
2. Nagar G (2018) The Evolution of Security Operations Centers (SOCs): Shifting from Reactive to Proactive Cybersecurity Strategies. International Journal of Scientific Research and Management (IJSRM). www.ijsrm. net ISSN (e), 2321–3418
3. He Z, Davila D, Bi S, Wang T, Hou T (2025) Machine Learning for Cybersecurity: A Survey of Applications, Adversarial Challenges, and Future Research Directions. Electronics 14(23):4563
4. Redhu A, Ranganath S, Rajdeo M, Agarwal D, Subramani R, Nashine SP, Srinivasan K (2026) AI-powered malware intelligence: a review of machine and deep learning paradigms in detection and classification. Artificial Intelligence Review
5. Pal S, Yadav G, Jadidi Z, Habib A, Uddin MP, Karmakar C, Shukla S (2026) Vulnerabilities in Machine Learning for cybersecurity. Current trends and future research directions
6. Ka-Kyung K, Joon-Seok K, Shin DH, Euom IC (2026) Cybersecurity Opportunities and Risks of Artificial Intelligence in Industrial Control Systems: A Survey. Comput Model Eng Sci, 146(2)
7. Sachpelidis-Brozos N, Katsaros E, Radoglou-Grammatikis P, Kalitsios G, Sarigiannidis A, Seritan GC, Sarigiannidis P (2026) Hacking intelligence: Mapping the anatomy of adversarial threats in artificial intelligence with MITRE ATLAS. Comput Sci Rev 61:100923
8. Raval KJ, Jadav NK, Rathod T, Tanwar S, Vimal V, Yamsani N (2024) A survey on safeguarding critical infrastructures: Attacks, AI security, and future directions. Int J Crit Infrastruct Prot 44:100647
9. Giarimpampa D, Meier R, Bissyande TF, Lenders V, Klein J (2025) Exploring the Role of Artificial Intelligence in Enhancing Security Operations: A Systematic Review. ACM-CSUR 58(3):1–38
10. Sommer R, Paxson V (2010) Outside the closed world: On using machine learning for network intrusion detection. Proceedings of the 2010 IEEE Symposium on Security and Privacy, 305–316
11. Choraś M, Pawlicki M, Kozik R, Hołubowicz W (2021) Machine learning and deep learning methods for cyber security: A survey. Appl Sci 11(4):1843
12. Angafor G (2024) AI Enabled WhatsApp Scams: Emerging Threats in Social Engineering. J Cybersecur Res 12(2):45–62

13. Borgaonkar R, Hirschi L, Park S, Shaik A (2023) Synthetic Social Engineering: AI Driven Fraud and Identity Manipulation, vol 128. Computers & Security, p 103123
14. Huang Y, Zhang L, Chen X (2024) Generative AI in Phishing and Online Fraud: A Systematic Analysis. IEEE Transactions on Dependable and Secure Computing
15. Papernot N, McDaniel P, Goodfellow I (2018) Adversarial Machine Learning: Attacks and Defenses. IEEE Commun Mag 56(5):80–87
16. Bilge L, Dumitraş T (2012), October 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 (pp. 833–844)
17. Buczak AL, Guven E (2016) A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Commun Surv Tutorials 18(2):1153–1176
18. Sarker IH, Jyothi A, S., Abawajy J (2020) AI driven cybersecurity: Machine learning and deep learning for cyber threat intelligence. Knowl Inf Syst 62(6):2405–2437
19. Alazab M, Khan S, Krishnan S, Venkatraman S (2020) AI based security operations and cognitive SOCs: A survey. IEEE Access 8:110205–110223
20. Denning DE (1987) An intrusion detection model. IEEE Trans Software Eng 13(2):222–232
21. Debar H, Becker M, Siboni D (1992), May A neural network component for an intrusion detection system. In S&P (pp. 240–250)
22. Alarm S, Raza M, Nawaz F (2023) Deep learning-based intrusion detection system for IoT networks. IEEE Trans Netw Serv Manage 10(3):431–442
23. Kim J, Park J, Lee H (2020) Deep learning-based intrusion detection systems: A survey. Electronics 9(5):1–28
24. Shone R, Ngoc TN, Phai VD, Shi Q (2018) A deep learning approach to network intrusion detection. IEEE Trans Emerg Top Comput Intell 2(1):41–50
25. Nataraj L, Karthikeyan S, Jacob G, Manjunath BS (2011) Malware images: Visualization and automatic classification. Proceedings of the 8th International Symposium on Visualization for Cyber Security, 1–7
26. Javaid A, Niyaz Q, Sun W, Alam M (2016) A deep learning approach for network intrusion detection system. Eai Endorsed Trans Secur Saf 3(9):21
27. Salo J, Hämäläinen T, Hiltunen T (2019) Machine Learning in Security Information and Event Management: Alert Classification and Correlation Approaches. Proceedings of the 2019 IEEE Conference on Cyber Security and Resilience
28. Yali R (2026) Data Science and Machine Learning for Cyber Intrusion Detection: A Systematic Review. Mach Learn Res 11(1):8–21
29. Chio C, Freeman D (2018) Machine learning and security: Protecting systems with data and algorithms. O'Reilly Media, Inc.

30. Eberle W, Holder L, Cook D (2013) Finding Outlier Behaviors in Multi–Source System Logs. IEEE Symposium on Computational Intelligence in Cyber Security
31. Waschke M, Vaarandi R, Guerra–Manzanares A (2026) A Survey of Machine Learning Techniques for Security Event Correlation and Automated SOC Analytics. J Cybersecur Digit Forensics 14(2):55–78
32. Li JH (2018) Cyber security meets artificial intelligence: a survey. Front Inform Technol Electron Eng 19(12):1462–1474
33. Goodfellow IJ, Shlens J, Szegedy C (2014) Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572
34. Biggio B, Roli F (2018) Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recogn 84:317–331
35. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., … Dennison,D. (2015). Hidden technical debt in machine learning systems. Advances in neural information processing systems, 28
36. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D (2018) A survey of methods for explaining black box models. ACM-CSUR 51(5):1–42
37. Sharma P, Gupta R, Al–Kuwari S (2026) Interpreting Deep Neural Networks: Challenges, Methods, and Applications in Security–Critical Systems. IEEE Trans Dependable Secur Comput 23(1):112–130
38. Askhatuly A, Berdysheva D, Berdyshev A, Adamova A, Yedilkhan D (2025) Adversarial Attacks and Defense Mechanisms in Machine Learning: A Structured Review of Methods, Domains, and Open Challenges. IEEE Access
39. Qiao Y, Sathyanarayana NB, Shi C, He Z, Wang T, Hou T (2026) A survey on adversarial machine learning: Attacks, defenses, real-world applications, and future research directions. Neurocomputing, 132670
40. Garg A (2026) A Review of Advancements in Deep Learning Approaches for Intrusion Detection Systems. J Artif Intell 8(1):273–298
41. Azeez SD, Ahmed SR, Ilyas M, Miah ASM, Farid FA, Karim MHA (2026) IntrusionNet: Deep Learning-Based Hybrid Model for Detection of Known and Zero-Day Attacks. Computers Mater Continua 88(1):1
42. Gangrade M (2025), August Integrating Deep Learning with Intrusion Detection Systems: A New Paradigm in Cyber Threat Mitigation. In 2025 World Skills Conference on Universal Data Analytics and Sciences (WorldSUAS) (pp. 1–13). IEEE
43. Khayat M, Barka E, Serhani MA, Sallabi F, Shuaib K, Khater HM (2025) Empowering security operation center with artificial intelligence and machine learning—A systematic literature review. IEEE Access 13:19162–19197
44. Binbeshr F, Imam M, Ghaleb M, Hamdan M, Rahim MA, Hammoudeh M (2025) The rise of cognitive SOCs: A systematic literature review on AI approaches. IEEE Open Journal of the Computer Society
45. Xu W, Evans D, Qi Y (2017) Feature squeezing: Detecting adversarial examples in deep neural networks. arXiv preprint arXiv :170401155

46. Barreno M, Nelson B, Sears R, Joseph AD, Tygar JD (2010) The Security of Machine Learning. Mach Learn 81:121–148
47. Apruzzese G, Colajanni M, Ferretti L, Guido A, Marchetti M (2019) On the Effectiveness of Machine Learning for Botnet Detection in the Presence of Adversaries. IEEE Trans Neural Networks Learn Syst 30(1):126–140
48. Rigaki M, Garcia S (2020) A Survey of Privacy Attacks in Machine Learning. ACM-CSUR 53(6):1–33
49. Ring M, Wunderlich S, Scheuring D, Landes D, Hotho A (2019) A Survey of Network–based Intrusion Detection Data Sets. Computers Secur 86:147–167
50. Ferrag MA, Maglaras L, Moschoyiannis S, Janicke H (2020) Deep Learning for Cyber Security Intrusion Detection: Approaches, Datasets, and Comparative Study. J Inform Secur Appl 50:102419
51. Biggio B, Nelson B, Laskov P (2012) Poisoning attacks against support vector machines. arXiv preprint arXiv :12066389
52. Doshi-Velez F, Kim B (2017) Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608
53. Anderson HS, Woodbridge J, Filar B (2021) Deep Learning for Cybersecurity: Trust, Explainability, and Analyst Adoption. IEEE Secur Priv 19(3):72–80
54. Batarseh FA, Freeman L (2021) Explainable AI in Regulated Industries: Challenges and Opportunities. AI Magazine 42(3):35–48
55. Molnar C (2019) Interpretable machine learning: a guide for making black box models explainable. Leanpub. Creative Commons Attribution
56. Carvalho DV, Pereira EM, Cardoso JS (2019) Machine Learning Interpretability: A Survey on Methods and Metrics. Electronics 8(8):832
57. Brundage M, Avin S, Clark J, Toner H, Eckersley P, Garfinkel B, Dafoe A, Scharre P, Zeitzoff T, Filar B (2018) & others. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv preprint arXiv:1802.07228
58. Wachter S, Mittelstadt B (2019) A Right to Reasonable Inference: Re–thinking Data Protection Law in the Age of Big Data and AI. Columbia Business Law Review, 2019(2), 494–620
59. Zuboff S (2023) The age of surveillance capitalism. Social theory re-wired. Routledge, pp 203–213
60. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., …Barnes, P. (2020, January). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 conference on fairness, accountability, and transparency (pp. 33–44)
61. O'Neil C (2017) Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy. Crown
62. Floridi L, Taddeo M (2016) What is data ethics? Philosophical transactions. Series A, Mathematical, physical, and engineering sciences, 374(2083), 20160360

Figures

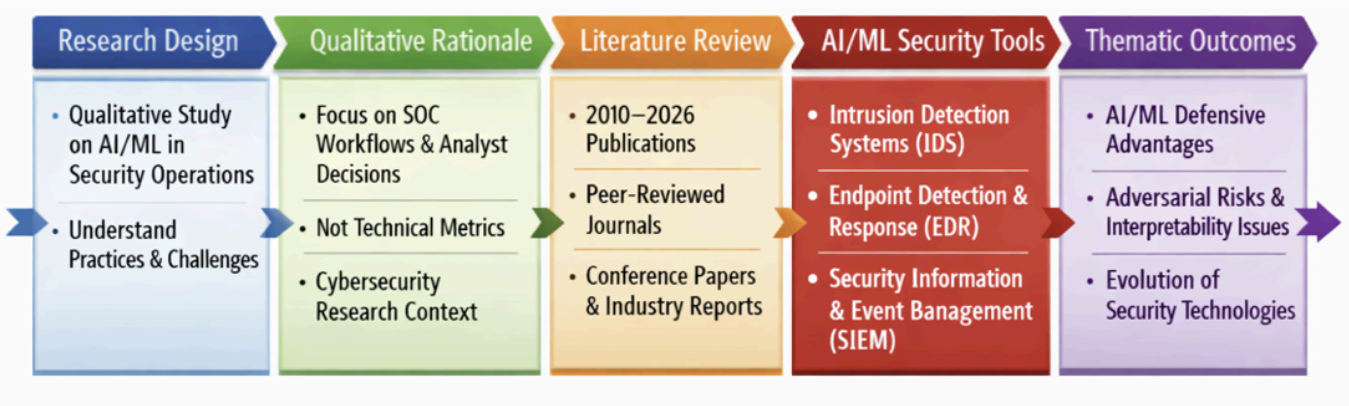


Figure 1

Graphical swimlane representation of the research methodology.