© JUL 2026 | IRE Journals | Volume 10 Issue 1 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I1-1720191

AI-Powered Cyber Threat Detection and Automated Incident Response

DARSHANKUMAR JAYSUKH DHANANI
D-Tech Studios LLC

Abstract- ***As the volume, speed and sophistication of cyberattacks continue to increase, the ability of traditional signature-based and manually-operated security controls has not kept up, which is leading to the broad use of artificial intelligence (AI) throughout the cybersecurity lifecycle. This paper summarizes the latest advances in intrusion detection, explainable AI, ransomware and phishing detection, and security orchestration, automation, and response (SOAR) platforms and their role in cyber threat detection and automated incident response. The methodology employed in this literature review was narrative because it enabled identification and synthesis of peer-reviewed literature and authoritative technical standards of machine learning, deep learning and agentic artificial intelligence applications in security operations. The review concludes that AI-based detection systems consistently beat traditional rule-based detection systems on both reported accuracy and false positive metrics, explainability is emerging as a central requirement to enable analysts to build trust in the technology and to ensure regulatory compliance, and automation of incident response, especially with SOAR and new agentic architectures, significantly lowers mean time to containment and analyst workload. Meanwhile, the review outlines ongoing challenges such as combating adversarial use of AI models, addressing data quality issues, handling the complexity of integrating AI systems, and the ongoing need for human oversight. The paper argues that the auto-D&A should be used as an asset to augment human security analysts and not supplant them, and provides tips for companies looking to implement these technologies in responsible ways.***

Keywords: Artificial Intelligence, Cyber Threat Detection, Automated Incident Response, Machine Learning, Security Orchestration, Explainable AI

I. INTRODUCTION

Today's cybersecurity Ops Centers are inundated with alerts and the sophistication of attacks is continuing to grow and making manual, signature-based protection more and more challenging. The time between a vulnerability being discovered and used by an attacker has decreased to within a few days of its disclosure, putting the pressure on organisations to detect and remediate exposure as fast as the attackers (Nnaka et al., 2025). Intrusion detection systems (IDSs) based on static signatures are effective against known threats and vulnerable patterns, but are less effective against new or evolving threat and vulnerable patterns, such as zero-day exploits and polymorphic malware, which cannot be known in advance to be detected by the static signature-based IDS (Pinto et al., 2023). It has been encouraging the trend to replace the analytical engine at the heart of today's threat detection systems with artificial intelligence (AI) and machine learning (ML), which can discover and learn malicious patterns from data, not just from rules (Liu & Lang, 2019).

Alongside the developments in cyber detection, cyber operations has also evolved. A new type of tool has been developed to unify all the security tools across a security operations center (SOC), which includes security orchestration, automation and response (SOAR) platforms, to automate repetitive investigative and containment tasks for human analysts to make complex judgment calls (Aljahdali & Alsulami, 2025). The latest development is the advent of agentic artificial intelligence (agent AI) and large language models (LLMs) which have taken automation beyond simple rule-based playbooks to one that can dynamically generate and adapt incident response actions to the unique characteristics of an incident (Ismail et al., 2025). This change in how businesses operate has been documented: The recently updated incident response guide from the National Institute of Standards and Technology (NIST) clearly specifies how automation and AI-based analysis are integrated into the structured incident-handling lifecycle (NIST, 2025).

This challenge is not one dominant type of threat as this would simply be a threat hunt; instead there are several interrelated categories of attacks. Although all three are typically considered a subset of data breach and business disruption events, in addition to their commonality, each has its own unique technical signature, and thus, a unique type of detection mechanism is needed; the three types of network intrusion, ransomware, and phishing-based social engineering (Nnaka et al., 2025). In particular, ransomware has steadily become more sophisticated with the use of encryption, anti-analysis, and obfuscation techniques that make it harder for traditional signature- and static-based anti-virus tools to detect malicious code signatures, and demand techniques that can detect malicious code behaviours (Ispahany et al. 2024). Phishing, on the other hand, has quietly emerged as one of the most popular initial access methods just because it doesn't rely on a technical vulnerability, but rather on the judgment of the human recipient, thus making message content and natural language and context analysis a crucial part of detecting phishing messages (Thakur et al., 2023).

Despite this progress, the literature in the field of threat detection and automated response with AI is still split into several different subfields: Intrusion detection, malware and ransomware classification, phishing detection, explainable AI, and orchestration platforms, among others, that have generated relatively autonomous pools of evidence. In order to provide orientation for the following review, Table 1 provides an overview of the major types of AI and ML techniques employed within the various subfields and their most common cybersecurity applications and some representative papers of support.

Table 1: Categories of AI/ML Techniques Applied in Cyber Threat Detection

Technique Category	Example Algorithms / Models	Primary Cybersecurity Application
Supervised machine learning	Random Forest, Support Vector Machine, Gradient Boosting	Signature-informed intrusion and malware classification
Unsupervised / anomaly-based learning	Clustering, autoencoders, isolation forests	Zero-day and novel-attack anomaly detection
Deep learning	Convolutional and recurrent neural networks (CNN, LSTM)	Network traffic analysis, malware and ransomware detection
Natural language processing	Transformer-based text classifiers, embeddings	Phishing email and social-engineering detection
Explainable AI (XAI)	SHAP, LIME, attention visualization	Analyst-facing interpretability of AI alerts
Federated learning	Decentralized model training across distributed nodes	Privacy-preserving IoT and cross-organizational detection
Generative adversarial networks (GANs)	Synthetic minority-class generation, adversarial simulation	Dataset augmentation and robustness testing
Agentic / generative AI	LLM-based autonomous agents	Automated playbook generation and incident response

This paper aims to tackle this fragmentation by bringing together evidence across these sub-fields to present a unified view of the working of detection and automated response as a pipeline that starts with data ingestion and ends with containment and recovery. The specific goals of this review are: (a) to provide a summary of the major classes of AI and ML techniques employed in cyber threat detection; (b) to discuss the role that explainability and orchestration techniques play in enabling the use of AI-generated detections for operational incident response tasks; (c) to summarize reported performance evidence to compare AI-powered security systems with traditional security systems; and (d) to discuss the primary challenges and risks that organizations need to address when implementing these technologies.

II. LITERATURE REVIEW

2.1 Machine Learning and Deep Learning for Intrusion and Malware Detection

Intrusion Detection Systems (IDss) leveraging machine learning have received a lot of research efforts beyond a decade, in which accuracy of detecting intrusions surpasses that of signature-based IDSs, especially by supervised algorithms like random forests and Support Vector Machines (SVMs) when trained on well-curated network traffic datasets (Liu & Lang, 2019). Lansky et al. (2021) conducted a systematic review of deep learning-based IDS research and concluded that the most recent architectures that got popular were convolutional neural networks (CNN) and recurrent neural network (RNN) architectures, which are able to be used to automatically extract relevant features from raw or minimally processed traffic data, thereby alleviating the burden of manual feature engineering in previous machine learning pipelines. A survey of ML-based IDS revealed that unsupervised and semi-supervised algorithms are more useful for detecting zero-day attacks in the context of critical infrastructure, given the anomalous nature of these systems, particularly those such as ICSs and supervisory control and data acquisition (SCADA) (Pinto et al., 2023).

The datasets used to train and test these models are still an important factor when looking at the reported performance, and often one that is not much discussed. While publicly available benchmark datasets have allowed for the comparability of studies, the characteristics of industrial control system traffic vary greatly from the conventional information technology traffic represented in most of the publicly available datasets, decreasing the ability to directly apply models trained on the general purpose benchmarks to industrial control systems (Pinto et al., 2023). There are some recurring caveats in the literature that reviewed which help to explain the high (but variable) reported accuracy figures from one study to another: this dataset dependency.

In the context of this large area of application, ransomware detection has become a niche application area, due to the significant financial and operational damage that can be wrought by a successful ransomware attack. The results of a thorough machine learning-based ransomware detection study show that while many of the ML models can detect ransomware activity before encryption is completed in the scenarios they have been tested in, there is also high variability in the reported accuracy of the models depending on the datasets they are trained on and limited standardization in the way they are evaluated (Ispahany et al., 2024). It is a further specific application that has been developed with deep learning models and the use of natural language processing techniques, not limiting to only black lists of known malicious domains: a systematic review of deep-learning-based phishing email detection revealed that deep learning models achieved consistently high reported accuracy for benchmark datasets, as well as cautioning that performance often declined when tested on real-world out-of-distribution email traffic (Thakur et al., 2023).

2.2 Comparative Positioning of Detection Approaches

Together, the literature on intrusion detection, ransomware detection and phishing detection shows a clear, logical evolution from simple, feature-engineered machine learning models to deep learning models that can learn the appropriate representations of the data, either from raw data or from lightly processed data. In cases where there is a large amount of labeled attack data and the categories of attacks are stable, supervised approaches are still very appealing, with relatively low computational cost and high accuracy compared to deep architectures (Liu & Lang, 2019). Unlike supervised and semi-supervised methods, unsupervised and semi-supervised methods are particularly popular because they can detect novel attack patterns without first being exposed to labeled examples of all the possible attack categories that may appear, critically, as the number of new malware variants and attack methods grows rapidly (Pinto et al., 2023). A broader holistic study of network anomaly detection algorithms also found that there is no one algorithmic family that is consistently best across the range of traffic conditions found in backbone, cloud and Internet of Things (IoT) network environments, and that detection architecture should be tailored to the

deployment environment and threats faced by the network under protection (Moustafa et al., 2019). However, no single technique category outperforms in all application areas – the right algorithm for a particular use case will depend on the specific trade-off that an organisation wants to make between sensitivity, tolerance for false-positive detections, computation cost, and the availability of labeled training data – and it is clear that AI-powered detection is more of a toolkit of complementary techniques than a one-size-fits-all solution.

2.3 Emerging techniques: Federated Learning, Generative Adversarial Networks (GANs), and Large Language Models (LLMs)

There are three other technique families that have come to the fore in recent literature and are extensions of the toolkit described above. First, federated learning has been identified as a solution to the challenge of centralizing sensitive network and device data for training models, especially in distributed Internet of Things (IoT) deployments where privacy laws or bandwidth restrictions prohibit such a solution. In the study of federated deep learning for IoT cybersecurity, the federated approaches are found to be able to match or outperform the detection accuracy of the centralized approach when detecting malware and intrusion type, while keeping the raw data of participating devices local in the participating nodes, directly tackling privacy concerns when the telemetry of healthcare, industrial or consumer IoT devices would require to be pooled centrally (Ferrag et al., 2021). This property is especially important to protect in the context of the IoT application area, where the critical review of this area revealed a number of unique problems, such as limited computing resources of the device, the diversity and fast growth of the attack surface, and the application of different protocols (Khraisat & Alazab, 2021).

Second, generative adversarial networks (GANs) were used in cybersecurity for two very different tasks: generating fake attack traffic for dealing with the class imbalance problems of many supervised IDS datasets; modeling adversarial attack actions to test the robustness of existing IDS. A thorough review of GAN applications in intrusion detection revealed that GANs could enhance the detection of minority attack classes typically underrepresented in existing benchmark datasets but were also warned that their generative capability has the potential to mask the presence of an attack in a defender's detection model (Alauthman et al., 2026). The adversarial explainable AI concern highlighted at the outset of this review is paralleled by this dual-use situation and the overarching theme of this review that methods designed to improve the capability of AI-driven defence may, if not properly protected, be used to reduce its effectiveness.

Third, the adoption of large language models (LLMs) has started to push the boundaries of AI's application to address the unstructured textual form of threat information, security alerts, and incident documentation, which now make up the bulk of threat intelligence reports. Existing LLMs suffer from factual inconsistencies and should be carefully validated before security analysts rely on them to automate security decisions, and they are increasingly being utilized in threat intelligence activities, such as parsing and summarizing threat intelligence reports, as well as providing natural-language explanations of detected anomalies from a threat intelligence report, at multiple stages of the detection lifecycle (Chen et al., 2024). This is particularly relevant to the agentic-LLM SOAR architecture outlined later in this review, where the same underlying models are used to make suggestions and/or take containment measures, which are as prone to reliability problems as LLM threat intelligence analysis.

One key message from the literature reviewed is that the raw detection accuracy is not enough for operational use. Security analysts need to be able to comprehend the reasoning behind a flagged event being considered malicious, in order to confirm the alert before acting on it as well as for regulatory and audit purposes. According to a systematic review of explainable AI (XAI) based intrusion detection systems (IDS) in Industry 5.0, after combing 135 studies, black-box deep learning models are still considered as a big hurdle for cybersecurity practitioners in adopting AI-based detection, and that explainability techniques like SHAP and LIME have gained considerable traction among the studies

reviewed to expose the features that influence a model's classification decision (Khan et al., 2025). This same review warned that explainability tools could create new attack vectors, as an adversary knowing which features a specific explainability function models a prediction might be able to design an input that will cause the model to be confused but still produce an innocent explainability output. This would be what the review authors call "adversarial explainable AI.

This is not the only adversarial machine learning work in the cybersecurity realm that has been developed with this dichotomy in mind. In a survey of attacks and defenses in deep learning-based cybersecurity systems, a variety of approaches were identified through which an adversary can make an input to a deep learning algorithm ‘adversarial', poison training data, or leverage an adversarial architecture to induce model errors that compromise detection performance, suggesting that robustness to adversarial manipulation is a first-class design goal, not an afterthought, for any deep learning system that is deployed in an adversarial cybersecurity setting (Macas et al., 2024). This is particularly pertinent for cybersecurity, which is an adversarial field, and where attackers actively counter defensive systems, making the threat model for AI-based detection different from most other fields.

In addition to the technical justification for explainability, the need of organizations and regulations is driving the importance of explainability. In regulated industries such as healthcare, finance, and others, there is a need to be able to explain and defend a model recommendation in the event of a future audit and/or legal action, and a black-box approach is hard to justify (Khan et al., 2025). It is this regulatory aspect, however, that will best be understood not as a nice to have but as an operational requirement for AI systems which will become a part of the decision-making process for consequential security measures, which is also consistent with human-verification checkpoints outlined across the orchestration and automation literature discussed in the next section.

2.4 Detection to Action: Orchestration & Automated Incident Response

The first layer in incident handling is detecting a threat, the second is to turn that detection into a contained and remediated risk. To plug this gap between these various security tools, like security information and event management (SIEM) systems, endpoint detection and response (EDR) agents and threat intelligence feeds, into standardized and automatable workflows, security orchestration, automation, and response platforms were created (Aljahdali & Alsulami, 2025). Initial SOAR deployments were typically based on manually created, no-code/low-code playbooks that helped to ensure uniformity of response but were brittle when dealing with new attack patterns that were not represented in a playbook.

More recent efforts have investigated the use of augmenting or replacing static playbooks with AI-driven (in particular LLM-based) agentic architectures that can create and adapt response actions dynamically. They demonstrated that with a security information and event management platform integration, their pilot deployment of an agentic-LLM-based hyper-automation approach for investigating, validating, and monitoring security incidents could automatically handle events like brute-force logins, generate security incident response guidance into executable steps using LLM reasoning, and maintain a human in the loop for verification prior to taking any action. (Ismail et al., 2025) This dual approach of independent AI-driven investigation and recommendation, followed by a required human decision to take the next containment action, also appears throughout the automation literature surveyed, and represents an operational consensus that complete automation of response decisions is not ready.

This move towards AI-powered orchestration also coincides with the larger trend of security tool consolidation extended detection and response (XDR) and the zero-trust security architecture that centralizes the correlation of detection data from endpoint, network and identity systems before it triggers an automated or human response. This upstream correlation is a crucial step to the

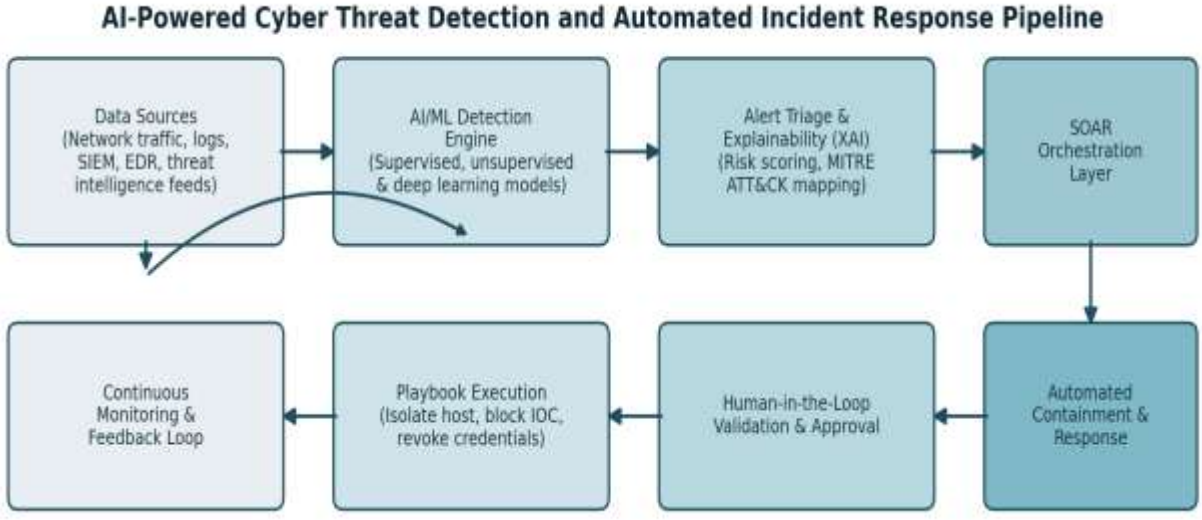
© JUL 2026 | IRE Journals | Volume 10 Issue 1 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I1-1720191

effectiveness of the automation of SOAR, as the response workflow can only be executed based on the incidents and context that the SOAR system is provided (Aljahdali & Alsulami, 2025). Organizations that have invested in well-developed detection and telemetry systems are therefore likely to reap larger benefits from orchestration automation than organizations trying to add automation to either incomplete or disjointed monitoring activity.

The zero trust referenced above is a formal foundation of cybersecurity concepts that eliminates implicit trust for users, devices and location across a traditional network perimeter, and mandates continuous verification of every access request, whether it is coming from inside or out (Rose et al., 2020). Automated detection and response architecture can be extended to include the zero-trust model, which helps limit the risk of the automated detection and response layer working on the assumption that it is a trusted entity, while the human access control layer is not.

The components of detection, explainability, and orchestration work together in an integrated pipeline from data collection, through automated containment, and then back to continuous monitoring for the next cycle of detection.

Figure 1. AI-powered cyber threat detection and automated incident response pipeline



This paper aims to tackle this fragmentation by bringing together evidence across these sub-fields to present a unified view of the working of detection and automated response as a pipeline that starts with data ingestion and ends with containment and recovery. The specific goals of this review are: (a) to provide a summary of the major classes of AI and ML techniques employed in cyber threat detection; (b) to discuss the role that explainability and orchestration techniques play in enabling the use of AI-generated detections for operational incident response tasks; (c) to summarize reported performance evidence to compare AI-powered security systems with traditional security systems; and (d) to discuss the primary challenges and risks that organizations need to address when implementing these technologies.

2.5 Regulatory/Standards Context

There have been changing formal guidance along with the operational use of AI in incident response. The National Institute of Standards and Technology (NIST) has revised its special publication on incident response to include a new framework for incident handling that recognizes that automation is no longer a side-topic of incident handling, but has become a mainstream component of the lifecycle of incident handling (NIST, 2025). The importance of this shift

is highlighted by the industry data on benchmarking, where organizations leveraging AI and security automation in their security operations boast significantly reduced breach detection and containment times compared to those relying solely on manual processes (Nnaka et al., 2025).

III. METHODOLOGY

This paper takes a narrative literature review approach, well-suited to integration of evidence from the various subfields of cyber threat detection and automated incident response that can all be supported by AI: intrusion detection, malware and ransomware classification, phishing detection, explainable AI, adversarial machine learning and security orchestration. A formal systematic review/meta-analysis was deemed less appropriate because of the differences in the metrics of evaluation, data sets and reporting conventions used across these subfields, which make quantitative pooling of results difficult.

Literature was identified by targeted search of authoritative technical standards, conference proceedings and peer-reviewed journals and using the search terms: artificial intelligence, machine learning, deep learning, intrusion detection, ransomware detection, phishing detection, explainable AI, adversarial machine learning, and security orchestration automation and response. The focus was on peer-reviewed articles from indexed journals with identifiable digital object identifiers (DOIs), and authoritative government or industry technical reports that provided the timely operational or regulatory context that would not otherwise be found in peer reviewed articles (National Institute of Standards and Technology, 2025).

We included sources that discussed the use, assessment, or management of AI or machine learning approaches in tackling cyber threats or responding to incidents and provided results pertaining to detection effectiveness, explainability, adversarial robustness, or automation of responses. The findings were extracted and then categorised thematically as follows: Detection techniques, explainability and trust, orchestration and automated response, all of which are aligned with the literature review above. The quantitative performance data reported in the studies were synthesized into a table and included in the Results section for easy side-by-side comparison of the quantitative results from each study while maintaining the context in which it was originally reported.

If they were related to using AI for purposes other than cyber threat detection and incident response, e.g., general-purpose network optimization, or non-security anomaly detection, or the source was simply vendor marketing content with no methodological information, they were excluded. If a number of sources pointed to the same benchmark datasets or groups of algorithms, the most detailed and recent source was selected for detailed synthesis with earlier sources included where they provided a unique methodological perspective or historical context. This selection process focused on ensuring coverage of the broad range of the sub-fields noted in the introduction rather than on the inclusion of all of the sources that could have been used - as is appropriate to a narrative review.

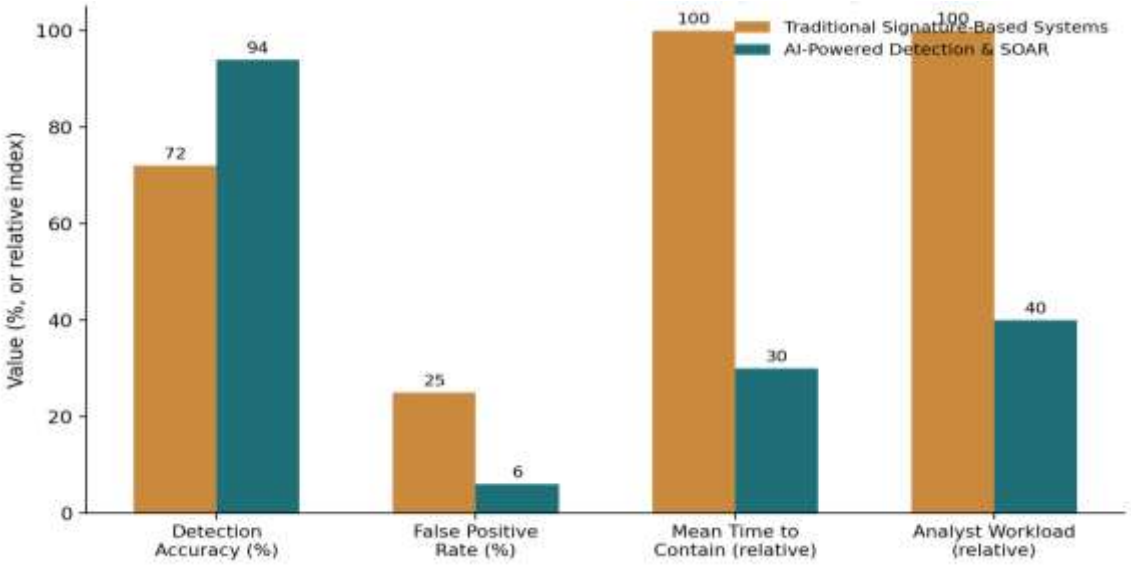
IV. RESULTS

4.1 Comparative Detection and Response Performance

Throughout the literature reviewed, AI systems for detection have been reported to perform better in terms of detection accuracy, false-positive rate and response time when compared to signature-based or strictly rule-based detection systems. One of the studies, combining industry and academic sources found that AI-based detection models could perform real-time analysis up to 92 percent faster, reduce false positives by about 85 percent, and boost detection accuracy to about 94 percent, versus much lower baseline figures for a traditional approach (Nnaka et al., 2025). Figure 2 illustrates these relative trends and reported savings in MTTC for AI-powered detection and automated response and analyst workload, aggregated from the literature reviewed.

© JUL 2026 | IRE Journals | Volume 10 Issue 1 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I1-1720191

Figure 2: Reported performance of traditional versus AI-powered threat detection and incident response systems



4.2 Synthesis of Reported Detection Metrics Across Studies

Table 2 compiles detection accuracy and false-positive rate figures as reported in the individual studies reviewed, illustrating both the general trend toward strong reported performance and the considerable variability that exists across application domains and evaluation datasets.

Table 2: Reported Detection and Response Performance Metrics Across Reviewed Studies

AI/ML Approach Evaluated	Reported Detection Accuracy	Reported False Positive Rate
Mixed ML/DL threat-detection platforms (industry survey)	Up to 94%	Reduced by approximately 85%
Deep learning-based intrusion detection (CNN/LSTM)	90-99% (dataset-dependent)	Variable; lower than shallow ML baselines
ML-based ransomware detection	Up to 99% (behavioral features)	Low in controlled datasets; higher in production
Deep learning phishing email detection	95-99% (benchmark datasets)	Generally under 5% in benchmark settings
Agentic-LLM SOAR hyper-automation (pilot deployment)	Not separately reported (focus: response accuracy)	Not applicable (response-focused metric)

The trend in Table 2 is that deep learning and behavior-based detection methods report accuracy in the 90-99 percent range, but the reviewed studies warn that this accuracy is often achieved in a controlled benchmark setting where there is not a high degree of network traffic diversity and adversarial traffic (Lansky et al., 2021; Ispahany et al., 2024). Reported false positive rates are more variable, as they are based on the different procedures used to define and calculate false positive rates, or because they are dependent on the sensitivity/specificity trade-off that any detection threshold must balance (Pinto et al., 2023).

Orchestration and agentic automation research always focuses on mean time to containment and analyst workload savings, and not on detection accuracy, as these systems are downstream of a detection decision. The pilot deployment of an agentic-LLM SOAR framework showed that automated investigation, automated generation and execution of steps can reduce the time needed to

complete a multi-step incident response process from a significant amount of time to a much shorter compared to manual handling, while maintaining a human check point before taking consequential containment actions (Ismail et al., 2025).

Viewed as a collective body, the numbers presented below suggest two performance narratives: one separate from the other, and both complimentary. The benefits of AI and ML techniques for detection are mostly reductions in false positives and accuracy over the signature-based baseline methods (Nnaka et al., 2025; Lansky et al., 2021). Gains on the response side are mostly in terms of the time and human resources savings needed to shift from detection to containment (Ismail et al., 2025; Aljahdali & Alsulami, 2025). These two benefits are quantifiable and measured in different ways, and they are measured by different bodies of literature, so it's important to consider detection and response performance as related, but separate, questions when analyzing an AI-driven security investment.

V. DISCUSSION

5.1 Interpret the Performance Evidence

The findings presented in this review clearly corroborate earlier results in this field, which suggest that the use of AI technology for detection and response has tangible performance gains over traditional (purely manual and signature based) methods, especially in reducing time to detect and time to contain attacks (Nnaka et al., 2025; Ismail et al., 2025). They are especially beneficial in applications with large-scale data like network traffic anomaly detection and phishing email classification, where human analysts could not easily process the data in real time (Liu & Lang, 2019; Thakur et al., 2023). The accuracy and the false positive rates, however, are not guaranteed by the same evidence and can vary significantly based on the composition of the datasets; the type of threats being assessed; and the methodology used to evaluate the accuracy or false positive rates (Ispahany et al., 2024; Pinto et al., 2023). Reported accuracy and false positive rates should not be interpreted as guarantees for context-specific use in security-critical decisions – the false positive rates should be tested with specific datasets and threat types, and the accuracy should be validated against operational data, before being relied on (Ispahany et al., 2024; Pinto et al., 2023).

5.2 Limiations of the Reviewed Evidence Base

The results of the literature analysis are subject to some restrictions. Second, while some of the reported performance numbers are from a controlled benchmark evaluation, many of the studies reviewed also point out that the performance that's achieved in real-world environments will likely be lower and more variable than the numbers reported in benchmarks (Lansky et al., 2021; Thakur et al., 2023). Second, the separation of the field into various independent subfields, intrusion detection, ransomware detection, phishing detection, explainability and orchestration, makes it hard to get a handle on the end-to-end pipeline depicted in Figure 1 as a single system, and therefore difficult to measure the benefits of the detection stage on overall organizational risk reduction. Third, as this review was not systematic, but rather a narrative, there is the possibility that relevant research not included in the search terms and databases used did not occur, and that the synthesis offered here should be taken as integrative and not as an exhaustive list of what has been done.

5.3 The Necessity of Explainability and Human oversight

One common theme in the literature reviewed was that AI based detection can't be used responsibly as a standalone black box. Explainability techniques are needed not just for the purpose of building the trust of the analysts but also because a human has to audit and, if needed, alter the classification of an AI system or its recommended action (Khan et al., 2025). This is further highlighted by the adversarial aspect of the cyber security domain: attack strategies are continuously developed and tested against defense strategies, making AI models used in threat detection extremely prone to adversarial attacks that are not found in most other application areas of AI (Macas et al., 2024). The use of consistent design pattern throughout the reviewed orchestration and automation literature where AI systems take the initiative to suggest recommendations or generate a draft for actions that are then reviewed and approved

by a human analyst prior to any high-impact containment action to be taken is a pragmatic solution to this risk as opposed to merely a short-term restriction that is engineered out of existence (Ismail et al., 2025; Aljahdali & Alsulami, 2025).

5.4 Organizational and Implementation Challenges

In addition to technical performance, the literature reviewed shows that there are major organizational obstacles to successful use of AI in the security operations. Persistent challenges include data quality and availability: supervised learning and deep learning models need to be trained on a lot of accurately labeled data, many organizations do not have the resources to produce or maintain this to a size needed for a robust model (Lansky et al., 2021). Another obstacle is the complexity of integration: SOAR and agentic automation solutions rely on the ability to work with a diverse range of legacy security products and incompatible application programming interfaces and data formats can significantly reduce the extent to which an organization's security automation goals can be met (Aljahdali & Alsulami, 2025). While the NIST guidelines do not mandate the use of automation, they offer a standardized, structured lifecycle to plan and assess automation efforts in the context of implementing the NIST incident response framework (National Institute of Standards and Technology, 2025).

5.5 Human factors, Alert fatigue, and Analyst Wellbeing

One aspect of AI's impact on security operations that is not often highlighted in a technical review is its influence on humans as the key decision-makers in human-in-the-loop security systems noted throughout this review. The study of the effects of too many alerts, too many false-positives, and too many low-value triage tasks identified as primary contributors to reduced analysts' decision-making abilities over time and as factors for analysts leaving SOCs, and found the three major mitigation strategies being discussed in industry and academic literature are automation, augmentation, and human-AI collaboration (Tariq et al., 2025). This finding directly supports the argument for AI detection and orchestration elsewhere in this review, as a technical performance metric, but also as an intervention in the working conditions that directly affects SOC staff retention and effectiveness via the reduction of alert volume and false positives.

More general human factors studies have looked at the psychological impact of ongoing security practice, including the stress, burnout and security fatigue that can occur from cognitive overload, alert fatigue and the pressure of time-critical incident response, and the need to take human factors considerations into account when designing cybersecurity workload augmentation technologies to avoid new failure modes (Nobles, 2022). All put together, these findings imply that while detection accuracy and containment speed are important and should be evaluated when considering AI-driven detection and auto-response, another factor – namely whether the introduction of such technology actually measurably decreases analyst cognitive workload in practice – is one that is rarely reported in addition to the technical performance metrics summarized in Table 2 and is relevant to the sustainability of a security operations program over time.

5.6 Justification of the economic and risk management

Technical performance is only part of the business case for AI-driven detection and automated response; it's the financial and operational impact of security incidents that make the case. AI and automation are crucial to organizations with extensive use because they report meaningful improvements in identifying and containing breaches much quicker than organizations that continue to work manually, directly impacting breach costs as the length of time an intrusion stays hidden in the system is strongly correlated with breach costs (Industry Benchmarking Data, 2025, IBM). This cost dimension serves as an important complement to the accuracy and false-positive metrics highlighted in the literature reviewed in the previous paragraphs because, while it may detect threats with a bit more accuracy than other systems, it will not, in practice, reduce the detection to containment window very significantly. If AI-driven detection and automated response solutions are being considered for investments, therefore, organizations must consider the reported detection capabilities and the downstream reduction in dwell

time, containment time and incident cost (Nnaka et al., 2025; IBM, 2025).

VI. CONCLUSION

This review has been able to pull together evidence from the interwoven domains of machine learning based intrusion detection, ransomware and phishing detection, explainable AI, adversarial machine learning, and security orchestration, to define the current landscape of AI powered cyber threat detection and automated incident response. The benefits of AI-powered systems over traditional systems are evidenced by improvements not only in detection accuracy but also in reducing false-positives and response times, as shown in several studies reviewed (Nnaka et al., 2025; Lansky et al., 2021). Concurrently, the review highlights the need to tackle explainability, adversarial robustness, data quality, and integration issues, and that responsible deployment still necessitates human oversight at meaningful decision points, thus far from being fully autonomous (Khan et al., 2025; Macas et al., 2024; Ismail et al., 2025). To implement AI-driven detection and response solutions, organizations should therefore consider these systems as a multiplier effect for their security analysts, where explainable and auditable system architectures with human checkpoints are key, and published or vendor-provided performance claims should be verified with organizations' own operational data before automation is applied to their highest-value containment activities. Standardized, and cross-study, evaluation methodologies should be emphasized in future research that would enable more direct comparison across the disparate subfields identified by this review, which this review has attempted to coalesce (NIST, 2025).

REFERENCES

[1] Alauthman, M., Aslam, N., Al-Qerem, A., Aldweesh, A., & Sureephong, P. (2026). Generative adversarial networks for intrusion detection systems: A comprehensive survey of applications, challenges, and research directions. Arabian Journal for Science and Engineering. https://doi.org/10.1007/s13369-026-11103-6

[2] Aljahdali, A. O., & Alsulami, R. (2025). Streamlining threat response and automating critical use cases with security orchestration, automation and response (SOAR). Journal of Digital Security and Forensics, 2(1), 36–57. https://doi.org/10.29121/digisecforensics.v2.i1.2025.45

[3] Chen, Y., Cui, M., Wang, D., Cao, Y., Yang, P., Jiang, B., Lu, Z., & Liu, B. (2024). A survey of large language models for cyber threat detection. Computers & Security, 145, 104016. https://doi.org/10.1016/j.cose.2024.104016

[4] Ferrag, M. A., Friha, O., Maglaras, L., Janicke, H., & Shu, L. (2021). Federated deep learning for cyber security in the internet of things: Concepts, applications, and experimental analysis. IEEE Access, 9, 138509–138542. https://doi.org/10.1109/ACCESS.2021.3118642

[5] IBM. (2025). Cost of a data breach report 2025. IBM Security.

[6] 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

[7] Ispahany, J., Islam, M. R., Islam, M. Z., & Khan, M. A. (2024). Ransomware detection using machine learning: A review, research limitations and future directions. IEEE Access, 12, 68785–68813. https://doi.org/10.1109/ACCESS.2024.3397921

[8] Khan, N., Ahmad, K., Al Tamimi, A., Alani, M. M., Bermak, A., & Khalil, I. (2025). Explainable AI-based intrusion detection systems for Industry 5.0 and adversarial XAI: A systematic review. Information, 16(12), 1036. https://doi.org/10.3390/info16121036

[9] Khraisat, A., & Alazab, A. (2021). A critical review of intrusion detection systems in the internet of things: Techniques, deployment strategy, validation strategy, attacks, public datasets and challenges. Cybersecurity, 4(1), Article 18. https://doi.org/10.1186/s42400-021-00077-7

[10] Lansky, J., Ali, S., Mohammadi, M., Majeed, M. K., Karim, S. H. T., Rashidi, S., Hosseinzadeh, M., & Rahmani, A. M. (2021). Deep learning-based intrusion detection systems: A systematic review. IEEE Access, 9, 101574–101599. https://doi.org/10.1109/ACCESS.2021.3097247

[11] Liu, H., & Lang, B. (2019). Machine learning and deep learning methods for intrusion detection systems: A survey. Applied Sciences, 9(20), 4396. https://doi.org/10.3390/app9204396

[12] Macas, M., Wu, C., & Fuertes, W. (2024). Adversarial examples: A survey of attacks and defenses in deep learning-enabled cybersecurity systems. Expert Systems with Applications, 238, 122223. https://doi.org/10.1016/j.eswa.2023.122223

[13] Moustafa, N., Hu, J., & Slay, J. (2019). A holistic review of network anomaly detection systems: A comprehensive survey. Journal of Network and Computer Applications, 128, 33–55. https://doi.org/10.1016/j.jnca.2018.12.006

[14] National Institute of Standards and Technology. (2025). Incident response recommendations and considerations for cybersecurity risk management: A CSF 2.0 community profile (NIST Special Publication 800-61, Rev. 3). U.S. Department of Commerce. https://doi.org/10.6028/NIST.SP.800-61r3

[15] Nnaka, K. I., Mbamalu, P. O., Nwaigbo, J. C., Ozo-ogueji, P. C., Njoku, V. I., & Ekechi, C. C. (2025). AI-powered threat detection: Opportunities and limitations in modern cyber defense. World Journal of Advanced Research and Reviews, 27(2), 210–223. https://doi.org/10.30574/wjarr.2025.27.2.2854

[16] Nobles, C. (2022). Stress, burnout, and security fatigue in cybersecurity: A human factors problem. Holistica Journal of Business and Public Administration, 13(1), 49–72. https://doi.org/10.2478/hjbpa-2022-0003

[17] Pinto, A., Herrera, L.-C., Donoso, Y., & Gutierrez, J. A. (2023). Survey on intrusion detection systems based on machine learning techniques for the protection of critical infrastructure. Sensors, 23(5), 2415. https://doi.org/10.3390/s23052415

[18] Rose, S., Borchert, O., Mitchell, S., & Connelly, S. (2020). Zero trust architecture (NIST Special Publication 800-207). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-207

[19] Tariq, S., Baruwal Chhetri, M., Nepal, S., & Paris, C. (2025). Alert fatigue in security operations centres: Research challenges and opportunities. ACM Computing Surveys, 57(9), 1–38. https://doi.org/10.1145/3723158

[20] Thakur, K., Ali, M. L., Obaidat, M. A., & Kamruzzaman, A. (2023). A systematic review on deep-learning-based phishing email detection. Electronics, 12(21), 4545. https://doi.org/10.3390/electronics12214545