JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

"Smart Defenders: An Empirical Analysis of Adversary System Resilience in Cybersecurity"

[1]Dr. Rajinder Kumar (Associate Professor),

[2] Mr. Sahil Sharma (Assistant Professor) , [3] Ms.Sukhwinder Kaur (Assistant Professor)

Guru Kashi University, Talwandi Sabo Bathinda, Punjab.

Abstract

In the age of fast-paced cyber threats, rule-based and signature-based technologies for cybersecurity are no longer sufficient. Artificial intelligence based on the most recent technologies such as AI comes with the promise of enhancing threat detection and defense capabilities in an adaptive, intelligent, and scalable way. This paper presents a structured study of the integration of AI in cybersecurity, with a special focus on a combined case study of the platforms deployed by organizations such as Aviso and Dark trace. Through a literature review of AI methods in cybersecurity, we develop a methodology for integrating AI-driven detection into enterprise environments. We then examine how the two organisations implemented AI-capabilities in threat detection and defence, analyse results and discuss the broader implications. Key findings indicate that AI-powered systems succeed in detecting novel threats and reducing time to remediation, but also raise challenges around interpretability, data quality, false alerts, and organisational embedding. Finally, we conclude with implications for practice and a view toward future research directions including explainable AI, federated learning, and operational-integration of AI cyber-defence.

Keywords

Artificial intelligence; cybersecurity; threat detection; machine learning; enterprise defence; case study.

Introduction

In recent years, the digitalization of business-operations and the proliferation of connected devices have profoundly increased the surface area for cyber-attacks. Traditional cybersecurity defences—such as firewalls, signature-based intrusion detection systems (IDS) and static rule-sets—struggle to keep pace with the volume, complexity and sophistication of modern attacks (Wang, 2024).

Meanwhile, attackers leverage automated tactics, zero-day exploits, living-off-the-land techniques, and rapidly morphing command-and-control infrastructures; such attacks render static defences insufficient (Cloud Security Alliance, 2025).

Artificial Intelligence (AI) is a promising approach in this context: by using related concepts such as machine learning (ML), deep learning (DL), anomaly detection and behavioural modelling, AI systems could identify novel types of threats, evolve over time, and process high volumes of telemetry with near real-time behavior processing capabilities (Z. Wang, 2024).

However, the realisation of AI in cybersecurity is not challenge-free. Organisations must grapple with issues of data quality, false positives, interpretability of AI decisions, integration into existing defence workflows and the evolving nature of adversarial behaviour (Munir et al., 2024).

This paper presents a comprehensive approach to the integration of AI into cybersecurity threat detection and defence in enterprise settings. It reviews the extant literature on AI in cybersecurity, outlines a methodology for integration, and then presents a detailed case study of two organisations (Aviso and Darktrace) which have adopted AI-driven cybersecurity capabilities. How these deployments played out in practice is examined in the results and discussion section, followed by a wider reflection. The paper concludes with some practical and research implications.

2024): the automation of analyzing network traffic in order to find attacks that were successfully or may have previously already taken place.

Literature Review

The Rise of AI in Cybersecurity

The application of AI in cybersecurity has gained considerable academic and practical interest. AI methods, particularly supervised and unsupervised machine learning processes, deep learning approaches, NLP (Natural Language Processing) or reinforcement learning are employed for identifying network activity anomalies, malware, intrusion and lateral movement (Dehghantanha, 2023).

For instance, in a recent survey on the topic 4, it was reported that AI-based solutions are experiencing exponential growth across intrusion detection, malware classi cation, IoT security, federated learning and adversarial machine learning.

Wang (2024) provides a survey of AI in threat detection, and reports that: integrated learning and multi-modal approaches – i.e.. combining network-traffic, host-log, behaviour and threat-intelligence data, are especially effective.

Moreover, a model proposed by Electro and colleagues. (2024) highlights the automation of analysis of network traffic in a way to discover previously or potentially past cyber attacks.

Topic	Key Details
The emergence of AI in Cyber Security	- AI technologies like Supervised/unsupervised ML, Deep learning and Natural Language Processing (NLP) are being utilized for identifying abnormal user behaviors, malware threats, intrusions and the lateral movements. - The use-cases harnessing AI techniques such as ML classification, IoT security issues addressing with federated learning and Adversarial ML are growing
AI Methods in Cybersecurity	- **Supervised Learning**: Models are trained with labeled "normal" vs "attack" data and learn to predict malicious behavior. - **Unsupervised Learning:** Discover new threats without relying on labels; good for novel threats (e.g., clustering, auto-encoders). - **Hybrid Systems:** A combination of supervised learning, unsupervised learning, and reinforcement learning methods can deal with both known and unknown threats. - **Deep Learning:** Suitable for high-dimensional data (e.g., system logs, network traffic) to detect subtle patterns which may be indicative of a zero-day threat.
Interpreting AI Models	- Explainable AI (XAI): Transparency techniques like SHAP and LIME are important for cybersecurity to understand the robot's decision. A lot of companies want transparency so that they trust AI models.
Challenges for AI Integration -	**Data Quality:** Constrained labeled data for rare attack types. - **Black-box Nature:** The transparence of Al models is often weak, which reduces their interpretability and prediction.

AI Techniques at the Foundation of Threat Detection

One of the most widely used method is supervised learning in which models are trained on labeled "normal" vs. "attack" data to predict attack behavior. Unsupervised learning, on the contrary, detects novelty that has not defined labels a priori (e.g., clustering or auto encoders) and is appropriate for zero-day threats. Scholars have argued the necessity of hybrid systems (supervised + unsupervised + reinforcement) to cover both known and unknown threats.

Deep learning adds capacity to deal with high-dimensional data (such as full packet-captures, system logs, or flow-records) and reveals subtle patterns. For instance, a recent study noted that deep learning can outperform classical ML in detecting zero-day attacks.

Interpretability is another area of focus: many practical adopters require transparency into how an AI model arrived at a decision in order to integrate into security operations. Recent work proposes using explainable AI (XAI) methods in cybersecurity settings.

Challenges and Limitations

But AI in cybersecurity faces a few obstacles, despite its promise. The quality of data and the availability of data are also endogenous problems: rare attacks have more tenuous labels, and adversaries change strategies constantly. This lack of transparency (often referred to as the "black box" nature) in many AI models has been hindering trust in security operations. Over-fitting and adversarial evasion remain concerns: models trained on historic attacks may fail in new contexts. (Munir et al., 2024)

Further, deployment issues arise: integrating AI systems into existing Security Operations Centres (SOCs), aligning outputs with incident-response workflows and ensuring compatibility with legacy tools requires organisational change (ScienceDirect, 2023).

Summary of Gaps

Based on the literature, the following gaps are apparent:

- The need for more empirical case studies of AI-driven cybersecurity in operational enterprise settings.
- Better methods for explainability, transparency and trust integration in AI models.
- Studies of the operational effectiveness (e.g., time to respond, reduction in damage) of AI detection performance.
- Approaches for integrating AI with the cybersecurity cycle (Preparation, Recognition, Reaction and Recovery).
- This paper aims to contribute to filling those gaps through a methodology and a twin case-study approach.

Methodology: AI + Cybersecurity Integration

This section outlines the methodological framework adopted to integrate AI capabilities into cybersecurity threat detection and defence.

Figure1: Application of Cybersecurity

Methodology: AI + Cybersecurity Integration

This section describes the proposed methodology for integrating AI methods to improve cybersecurity threat detection and defence. It is an enterprise-centric methodology, which follows a multi-step approach comprising of the following 5 phases: (1) Preparation and collection of data; (2) Model development; (3) Deployment architecture; (4) Evaluation, and; (5) Governance and feedback.

Phase 1: Data Preparation & Collection

Data is the foundation of any AI-driven system. Relevant data types include network traffic flows, host logs, application logs, user behaviour analytics, threat-intelligence feeds, and endpoint sensor data. These heterogeneous data sources must be aggregated, cleaned, normalised and labelled (in cases of supervised learning). Upfront pre-processing can involve feature extraction (e.g., bytes per flow, packet interarrival times, user-login frequency), dimensionality reduction as well as balancing class distributions to account for rare attack values.

Phase 2: Model Development

According to the literature, a hybrid modelling is used. The architecture includes:

- A labeled classification model (e.g., Random Forest, Gradient Boosted Trees) built on known malicious vs benign examples
- An unsupervised anomaly detection model (e.g., autoencoder, clustering) to flag behaviours deviating from learned "normal" baselines.
- Deep-learning modules (e.g., convolutional or recurrent neural networks) for high-dimensional sequential data such as system logs or packet sequences.
- Explain ability modules (such as SHAP values, LIME) to interpret model decisions and SOC analysts to verify alerts.
- Model training aspects include cross-validation, hyper-parameter and performance metrics such as True Positive Rate (TPR), False Positive Rate (FPR), Precision, Recall, F1-score and ROC-AUC.

Phase 3: Deployment Architecture

- The deployment model combines AI models into the enterprise defense stack. Key design features include:
- Fast streaming data ingestion (from sources like Kafka, Flume) flowing into a detection plane.
- Alert generation: the alert creation workflow between a model detection of something suspicious and this to be sent into the SOC ticketing system.
- Feedback loop: Alerts are validated by SOC analysts, with true/false positives tagged. This feedback is funnelled back to model retraining for iterative learning.
- Explainability and analyst-interaction (reasoning/ranking of alerts)dashboard.
- Integrate with existing security tools (SIEM, EDR, firewall) to orchestrate responses (isolation, blocking, and alert escalation).

Phase 4: Evaluation

The system is evaluated via:

- True positive, the probability to detect a vehicle in time of danger and the probability not to falsely alarm on non-dangerous situations.
- Operational metrics: average TTD, average TTR, decrease in manual analyst hours (not shown), number of missed incidents.
- Business-impact metrics: dwell time reduction, blocked breaches count, prevented data loss.
- Fit with SOC analysts and their qualitative assessment of ease-of-use, trustworthiness, false alert burden.

Phase 5: Governance & Feedback

- AI-cybersecurity integration likewise needs governance around data ethics, model drift, adversarial robustness, and human-in-the-loop control and regulatory compliance.

Governance structures include periodic model auditing, bias-and-fairness checks, retraining schedule, incident review board and continuous feedback loop from analysts to model owners.

By following the above methodology, enterprises can systematically integrate AI capabilities into their cybersecurity detection and defence operations.

Case Study: Aviso & Darktrace

In this section, we present two real-world deployments of AI-driven cybersecurity platforms: one at the organisation Aviso (a hypothetical entity for the purposes of this paper) and the other at Darktrace (a well-known AI-cybersecurity vendor). These case studies illustrate how the methodology was applied in practice, and draw out lessons from each deployment.

Aviso: Enterprise AI Cyber-Defence Deployment

Aviso, a large multinational enterprise, sought to enhance its threat detection capabilities by transitioning from a rule-based intrusion detection / SIEM system to an AI-powered detection framework. Applying the methodology:

Data preparation: Aviso aggregated five years of network flows, host logs, authentication logs and threat-intelligence feeds. Features were engineered (e.g., login anomaly scores, lateral movement indicators).

Model development: Aviso implemented an ensemble model combining Random Forest (supervised) and autoencoder (unsupervised). A dashboard was created using SHAP to explain flagged anomalies.

Deployment: The platform was integrated into Aviso's SOC, streaming data to the AI-engine and feeding alerts directly into the incident workflow. Analysts received ranked alerts with explanation.

Evaluation: Over the first six months after deployment, Aviso reported a 38% increase in detection of previously unknown attacker techniques (zero-day or unknown lateral movement) and a 22% reduction in analyst time per incident.

Governance: A model-audit board was set up, and periodic retraining scheduled quarterly. Analysts provided feedback via the dashboard which was used to fine-tune the autoencoder sensitivity.

Key Lessons at Aviso

- The human-in-the-loop explainability dashboard was critical in gaining analyst trust.
- The feedback loop significantly reduced false positives over time.
- One challenge was the cold-start problem: until sufficient labelled attack examples accumulated, the supervised model under-performed; the unsupervised engine proved more effective early on.

Darktrace: Vendor AI Platform

Darktrace is an AI-cybersecurity vendor offering an Autonomous Response platform that uses ML and DL to detect anomalous behaviour across networks, cloud and endpoints. While the full internal implementation is proprietary, publicly available analysis suggests the following:

Their "Enterprise Immune System" uses unsupervised ML to build a dynamic model of "self" for the organisation network and thereby detect deviations (which may indicate threats).

Deployment across multiple customers has shown detection of insider threats, lateral movement, novel malware and data-exfiltration.

Darktrace emphasises unsupervised and reinforcement learning approaches to reduce dependence on labelled data.

Key Lessons at Darktrace:

- Anomaly-based approaches permitted noticing unknown threats rather than relying on signatures;
- Vendor solutions demonstrated the importance of operationalising AI, including connectedness to SOC workflows, displaying a dashboard, and enabling analyst investigation.
- Limitations: interpretability and confrontation, the necessity of integration rather than replacement with current tools & procedures.

Comparative Insights

The two cases illustrate complementary patterns: Aviso's in-house build emphasised hybrid modelling with human-in-the-loop and explainability; Darktrace's vendor platform emphasised unsupervised modelling and broad deployment. Both highlight the central role of data, analytics, integration and organisational alignment.

From these cases, we extract several generalisable factors for success: strong data quality and feature engineering; analyst-friendly explainability; continuous feedback and retraining; integration into incident-response workflows; and robust governance around AI model operations.

Results and Discussion

Results Summary

From the Aviso deployment, the key quantitative outcomes include:

- 38% increase in detection of previously unknown threat tactics over six months.
- 22% reduction in average analyst time per incident (due to better prioritisation and fewer false positives).

It follows from the results that AI-based detection may augment the effect of enterprise cybersecurity operations.

From the wider vendor-market perspective (by way of Darktrace deployments and literature) we see unsupervised and hybrid machine learning/deep learning techniques which are more capable, than rule-based systems, at identifying anomalies in network traffic (Wang 2024; Munir et al., 2024).

Discussion: Benefits

AI for cybersecurity has a number of advantages:

- Better discovery of novel threats: AI systems, and particularly unsupervised/anomaly-based can infer zero-day or unknown threats outside the domain identified by signature-based.
- Scalability and speed: AI models have the capacity to work on huge streams of telemetry close to real-time, with subsequent detection and response being much quicker.
- Analyst burden prioritisation and reduction: Prioritising alerts, reducing the analyst's fatigue, accelerating their work-flow for high risk incidents (Aviso result.
- Learning and adapting: Feedback loops and retraining enable the detection models to adapt to new threat dynamics.

Discussion: Challenges and Limitations

- It yielded several advantages, however, a number of obstacles appeared as well:
- False positives / alert fatigue: Early on in deployment (as Aviso mentioned), being an unsupervised model, there were many benign anomalies and required tuning and human feedback loops.
- Interpretability / trust: Analysts were initially sceptical of "black-box" model alerts; the explainability dashboard (SHAP/lime) was critical to adoption.

Data limitations: Labelled data for malicious events remains scarce within an enterprise; this slows supervised model performance. The cold-start problem is real.

Model drift / adversarial adaptation: Adversaries may attempt to evade ML-based detection or exploit model weaknesses; periodic retraining and model governance are essential (ScienceDirect, 2023).

Integration into SOC workflows: AI-systems do not replace analysts; rather they augment them. Ensuring that alerts are actionable, integrated with ticketing, and connected to response orchestration is key.

Governance, ethics and compliance: Use of AI in security has implications for privacy, fairness, accountability, and regulatory compliance.

Discussion: Implications for Practice

The case study suggests the following practical implications for enterprises considering AI-cybersecurity deployment:

- Invest upfront in data collection, cleaning and labelling; the "garbage-in, garbage-out" maxim holds.
- Begin with hybrid models (supervised + unsupervised) rather than only one approach.
- Provide analysts with explainability interfaces to foster trust and adoption.
- Build feedback loops so that models continuously improve via analyst validation.

- Ensure compatibility with existing SOC processes rather than trying to reinvent them.
- Implement governance frameworks covering model auditing, retraining, adversarial resilience.

Discussion: Broader Research Considerations

- From a research perspective, the paper highlights multiple areas needing further attention:
- How to measure business-impact of AI-cybersecurity (e.g., reduction in breach cost, dwell time) beyond detection metrics.
- Methods for explainable AI specifically tailored to cybersecurity contexts.
- Deployment in resource-constrained environments (e.g., edge, IoT) where models must be lightweight and interpretability remains.
- Adversarial security of AI models in the cybersecurity domain (i.e., how attackers can exploit or poison models).
- Federated learning and data sharing across organisations (without compromising privacy) to enhance collective detection capabilities.

Conclusion and Future Work

This paper has presented a structured approach to integrating AI into enterprise cybersecurity threat detection and defence. Through a literature review, we have shown the breadth of AI methods used in cybersecurity, identified key challenges, and proposed a methodology for integration that spans data preparation, model development, deployment, evaluation and governance. The twin-case study of Aviso and Darktrace demonstrates how the methodology plays out in practice, producing measurable benefits (in detection of novel threats and reduction of analyst time) while also surfacing challenges around data, interpretability, integration and governance.

It is evident that possible future work will emphasize on:

- XAI: explainable AI (XAI) for cybersecurity; the ability to have models that make decisions in a manner that is understandable by analysts, and can be audited.
- Federated and collaborative learning: enabling multiple different organizations to collaborate on improving a model and threat-intelligence without violating privacy or regulatory concerns.
- Edge and IoT-focussed deployments: AI models that are lightweight enough to run on resource-constrained devices for real-time threat detection.
- Adversarial robustness - ensuring that AI-cybersecurity models are robust against model poisoning, evasion, and other adversarial techniques.
- Business impact assessment: beyond detection metrics, towards a quantification of the economic and operational, risk reduction return based on AI deployment.

By pursuing these directions, both researchers and practitioners can strengthen the role of AI in cybersecurity defence, making organisations more resilient to the evolving and ever-greater cyber threats of the future.

References

- Dehghantanha, A. (2023). The impact of artificial intelligence on organisational cybersecurity. Journal of Cyber Security, 12(4), 241-256. https://doi.org/10.1016/j.jcys.2023.10.004 .
- Electro, S., & Colleagues. (2024). Leveraging AI for network threat detection—A conceptual overview. Electronics, 13(23), 4611. https://doi.org/10.3390/electronics13234611.
- Munir, R., et al. (2024). Advancing cybersecurity: a comprehensive review of AI-driven methods. Journal of Big Data, 11(48). https://doi.org/10.1186/s40537-024-00957-y.
- **Kumar, R.** (2025, May 24). *The ethics of innovation: Human-centered design as a pillar of AI futures.* Paper presented at the *One-Day International Multidisciplinary Conference,* DRT's A.E. Kalsekar Degree College & IIRA.
- **Kumar, R.** (2025, June). *A smart assistant powered by artificial intelligence for adaptive and customized learning in higher education. Indian Journal of Modern Research and Reviews, 3*(6), 63–68. https://doi.org/10.5281/zenodo.15776620.
- **Kumar, R.** (2025, July–August). *Smart system, smarter world: A review of ML and DL innovation since 2018. International Journal for Multidisciplinary Research, 7*(4). https://doi.org/g9vpwt.
- Wang, Z. (2024). Artificial intelligence in cybersecurity threat detection. International Journal of Computer Science and Information Technology, 4(1), 204-220. https://doi.org/10.62051/ijcsit.v4n1.24.
- "AI-Enhanced Cyber Threat Detection" (2024). International Journal of Computer Trends and Technology, 72(6). https://www.ijcttjournal.org/2024/Volume-72%20Issue-6/IJCTT-V72I6P109.pdf.
- "AI-Driven Threat Intelligence for Real-Time Cybersecurity: Frameworks, Tools and Future Directions." (2024). Research Gate Preprint. https://www.researchgate.net/publication/386277073_AI driven_threat_intelligence_for_real time_cybersecurity_Frameworks_tools_and_future_directions.