Journal of Integrated Science, Technology and Management, 2026; 1(2):22-26

Journal of Integrated Science, Technology and Management

Research Article

AI-Driven Cyber Risk Management: A Review of Threat Detection, Prediction, and Mitigation Strategies

Vimlesh Kumar

Department of Computer Science and Engineering Teerthankar Mahaveer University, Moradabad U.P, India

ARTICLE INFO

Article history:

Received: 10 January, 2026

Revised: 20 January, 2026

Accepted: 09 March, 2026

Published: 30 March, 2026

Keywords:

cyber risk management; artificial intelligence; intrusion detection; explainable AI; advanced persistent threats; machine learning; risk prediction; mitigation strategies

ABSTRACT

As cyberattacks grow more frequent, automated, and difficult to detect using signature-based defenses, artificial intelligence (AI) has become a central pillar of cyber risk management across three interlocking functions: threat detection, risk prediction, and mitigation. This review synthesizes evidence on AI-driven approaches to each function, with particular attention to deep learning-based intrusion detection and the growing role of explainable AI (XAI) in making these systems trustworthy enough for operational deployment. Systematic evidence shows machine learning (ML) techniques — including convolutional neural networks, support vector machines, and Bayesian classifiers — substantially outperform manual, human-driven threat analysis (Alshuaibi et al., 2025), while quantitative research demonstrates that supply-chain network features measurably improve prediction of enterprise data-breach likelihood (Hu et al., 2022). A growing body of XAI research — spanning SHAP- and LIME-based interpretability methods for network intrusion detection systems and dedicated reviews of explainable deep learning for advanced persistent threat (APT) detection — addresses the interpretability gap that has historically limited analyst trust in black-box detection models (Abdul Mutalib et al., 2024; Rjoub et al., 2023). The review also examines the enterprise-architecture and infrastructure layer required to operationalize these capabilities at scale, including real-time event-driven data processing (Sannidhanam, 2021; Event Streaming Architectures for High-Volume Transaction Processing, 2022), configurable and metadata-driven enterprise platforms for embedding mitigation logic into operational workflows (Basireddy, 2022a, 2022b), audit-ready compliance architecture (Basireddy, 2023), and autonomous AI agents for continuous infrastructure remediation (Sannidhanam, 2025). The review concludes that closing the gap between AI's detection and prediction capabilities and its practical, governed deployment increasingly depends on explainability research and enterprise-architecture maturity in equal measure.

Introduction

Cyber risk management has shifted decisively from static, signature-based defense toward AI-driven approaches capable of learning and adapting to an evolving threat landscape. This shift has been driven by two converging pressures: the increasing sophistication and automation of attacker techniques, including attacker use of AI itself, and the sheer volume of network, transaction, and log data that modern organizations generate, which has exceeded the practical capacity of manual security analysis. Machine learning (ML) and, increasingly, deep learning have become the default technical foundation for detecting, predicting, and mitigating cyber risk across this expanded threat surface (Alshuaibi et al., 2025).

This review examines AI-driven cyber risk management across the same three functions addressed in its title — detection, prediction, and mitigation — but places particular emphasis on two themes that have become increasingly central to the recent literature: the specific

***Corresponding Author:** Vimlesh Kumar

Address: Department of Computer Science and Engineering Teerthankar Mahaveer University, Moradabad U.P., India

Relevant conflicts of interest/financial disclosures: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

challenge of detecting advanced, persistent, and previously uncatalogued threats using deep learning, and the parallel rise of explainable AI (XAI) as a response to the interpretability limitations of these same deep-learning models. The review draws on systematic reviews of ML in cybersecurity (Alshuaibi et al., 2025), empirical cyber-risk prediction research (Hu et al., 2022), dedicated XAI-cybersecurity literature (Abdul Mutalib et al., 2024; Kalakoti et al., 2025; Rjoub et al., 2023), and applied enterprise-architecture research addressing how detection and mitigation capabilities are operationalized within enterprise systems (Basireddy, 2022a, 2022b, 2023; Sannidhanam, 2021, 2025).

Threat Detection: From Classical ML to Deep Learning

Threat detection research has progressed from classical ML classifiers toward deep-learning architectures capable of learning more complex representations of network and system behavior directly from raw or lightly processed data. Systematic review evidence confirms that classical techniques — convolutional neural networks, support vector machines, and Bayesian classifiers — enable substantially more efficient identification of malicious activity than manual, human-driven analysis, particularly given the scale of contemporary network traffic (Alshuaibi et al., 2025). More recent comparative research finds that deep-learning-based network intrusion detection systems (NIDS) further improve detection rates and reduce false-positive rates relative to earlier ML approaches, though at the cost of increased model opacity, an issue explored in Section 3 (Tan et al., 2025).

A specific and increasingly important detection challenge is the identification of advanced persistent threats (APTs) — sophisticated, sustained intrusion campaigns designed to evade traditional detection. A recent review focused specifically on this problem finds that traditional intrusion detection systems suffer from low detection accuracy and high false-positive rates against APTs, and struggle particularly with previously unseen attack categories, motivating the integration of deep learning with explainability methods as a combined response to both the accuracy and the trust deficits of existing APT-detection approaches (Abdul Mutalib et al., 2024).

The Explainability Imperative

As deep-learning models have improved raw detection accuracy, they have simultaneously introduced an interpretability deficit that limits their practical adoption in security operations centers (SOCs), where human analysts must ultimately validate, prioritize, and act on AI-generated alerts. A substantial and rapidly growing body of research has emerged specifically to address this gap. Explainable AI (XAI) methods — including SHAP (Shapley Additive Explanations), LIME (Local Interpretable Model-agnostic Explanations), and related techniques — have been applied to network intrusion detection systems to clarify which input features drive a given alert and to what degree, converting black-box classifications into explanations that analysts can evaluate and trust (Kalakoti et al., 2025).

This XAI research serves multiple purposes beyond simple interpretability. Kalakoti et al. (2025) note that explainability helps organizations meet regulatory compliance standards for automated decision-making, improves detection-system design by identifying which features are most predictive, and provides forensic value after a security incident by helping analysts reconstruct why particular alerts were, or were not, triggered. A broader survey of explainable AI applications across cybersecurity similarly finds XAI methods increasingly integrated across detection, classification, and response functions, reflecting a field-wide recognition that detection accuracy alone is insufficient without a corresponding capacity to explain and justify AI-driven security decisions (Rjoub et al., 2023). Applied specifically to APT detection, this integration of deep learning with XAI methods such as SHAP and LIME has been proposed as a combined strategy for improving both the detection of sophisticated, evasive threats and the trustworthiness of the resulting alerts (Abdul Mutalib et al., 2024).

Risk Prediction

Beyond detecting active intrusions, predictive cyber-risk modeling seeks to forecast the likelihood of future attacks or breaches before they occur, enabling a more proactive security posture. Research applying supervised, unsupervised, and reinforcement-learning techniques to this predictive task has demonstrated the ability to forecast and pre-empt cyberattacks, while identifying data-privacy and model-maintenance challenges as ongoing constraints on predictive accuracy against a continuously evolving threat landscape (Loschilin et al., 2024).

The most methodologically rigorous predictive evidence reviewed here comes from research linking enterprise cyber risk directly to supply-chain network characteristics. Hu et al. (2022) found that incorporating supply-chain network features into a gradient-boosting prediction model improved out-of-sample predictive accuracy for enterprise data-breach likelihood by 2.3 percent relative to a model using only an organization's own

Table 1: AI/ML technique families across the cyber-risk detection, prediction, and mitigation lifecycle.

Technique Family	*Primary Function*	*Representative Evidence*
Deep neural networks (DNN/CNN)	Network intrusion detection from traffic and log data	Deep-learning NIDS approaches improve detection rates and reduce false positives relative to earlier ML methods (Tan et al., 2025)
Classical ML classifiers (SVM, Bayesian, Random Forest)	Threat detection and classification	Enable more efficient identification of malicious activity than manual analysis (Alshuaibi et al., 2025)
Gradient boosting with network-feature engineering	Cyber risk / breach prediction	Supply-chain network features improved out-of-sample breach-prediction AUC by 2.3% (Hu et al., 2022)
Supervised, unsupervised, reinforcement learning	Attack prediction and pre-emption	Applied to forecast and prevent cyberattacks before occurrence (Loschilin et al., 2024)
Explainable AI (SHAP, LIME, and related methods)	Interpretability of detection/prediction outputs	Used to clarify which features drive intrusion-detection alerts, aiding analyst trust and compliance (Kalakoti et al., 2025)
Explainable deep learning for APT detection	Detection and interpretation of advanced persistent threats	Reviews the integration of XAI methods (SHAP, LIME) with deep learning to improve APT-detection transparency (Abdul Mutalib et al., 2024)
Configurable workflow / metadata-driven architecture	Mitigation logic embedded in operations	Risk-control logic embedded directly into enterprise workflows (Basireddy, 2022a, 2022b)
Autonomous AI agents	Continuous monitoring and remediation	Self-directed detection and remediation of infrastructure anomalies (Sannidhanam, 2025)

security attributes — a finding with direct implications for predictive strategy, since it demonstrates that internally focused risk models systematically underestimate risk arising from an organization's network of suppliers and partners.

Mitigation Strategies

Effective mitigation depends on translating detection and prediction outputs into timely operational response. At the algorithmic level, ML-informed mitigation strategies tailor defensive responses to the specific threat type identified, distinguishing, for example, between active attacks seeking to disrupt systems and passive attacks seeking undetected data exfiltration. At the enterprise-systems level, this translation from detection to response depends on the architecture within which security tools operate. Basireddy (2022b) proposes a configurable enterprise workflow architecture that embeds risk-control logic, including security-relevant controls, directly into business-process workflows, enabling detected threats to trigger an operational response as part of normal process execution rather than through a separate escalation cycle. Basireddy (2022a) extends this with a metadata-centric platform architecture that allows mitigation and automated-response logic to be reconfigured dynamically as new threat types emerge, without extensive custom redevelopment.

The most advanced form of mitigation reviewed here is autonomous remediation. Sannidhanam (2025) describes autonomous AI agents deployed for cloud infrastructure operations that continuously monitor infrastructure state, detect anomalies indicative of compromise, and execute remediation steps directly, extending mitigation from human-reviewed alerting toward self-directed operational action — a capability that improves response speed but, as discussed in Section 7, introduces new governance requirements not fully addressed by existing security-operations frameworks.

Infrastructure Requirements for Real-Time Response

Because cyber threats can propagate faster than batch-oriented monitoring can process, AI-driven detection, prediction, and mitigation increasingly depend on real-time, event-driven data infrastructure. Applied research on event streaming architectures for high-volume transaction processing (2022) and real-time claims processing using event-driven architectures

(Sannidhanam, 2021) illustrates architectural patterns for processing high-volume operational data as continuous streams rather than periodic batches, directly supporting the low-latency detection and response that effective cyber risk management requires. As AI systems, including LLM-based components, become embedded in security decision-making, audit-ready compliance architecture is similarly necessary to ensure these decisions remain traceable and defensible to regulators (Basireddy, 2023).

Evidence Snapshot: AI Technique Families Across the Cyber-Risk Lifecycle

Table 1 summarizes the AI/ML technique families reviewed in this article, organized by their primary function within cyber risk management.

Challenges and Governance Gaps

Interpretability-accuracy trade-off: the deep-learning models that achieve the highest detection accuracy tend to be the least interpretable, motivating but not yet fully resolving the growing XAI literature (Kalakoti et al., 2025; Tan et al., 2025).

APT and novel-threat detection remains imperfect: even with deep learning and XAI integration, detection of advanced, previously unseen threats continues to show accuracy and false-positive limitations relative to known-threat detection (Abdul Mutalib et al., 2024).

Adversarial robustness of explanations: recent research finds that post-hoc XAI explanations themselves can be deceived or manipulated in network intrusion detection contexts, indicating that explainability is not yet a fully reliable safeguard against adversarial evasion.

Third-party and supply-chain risk visibility: predictive models built solely on internal data continue to underestimate cyber risk arising from network-connected third parties (Hu et al., 2022).

Alert fatigue: even with improved detection accuracy, large volumes of AI-generated alerts can overwhelm human analysts, motivating research into AI-assisted alert prioritization and triage alongside detection itself.

Governance of autonomous remediation: as AI systems move toward self-directed mitigation action, existing security-operations governance, designed around human-reviewed escalation, requires adaptation to oversee AI-initiated response (Sannidhanam, 2025).

Conclusion

AI-driven cyber risk management has matured substantially across detection, prediction, and mitigation, with deep-learning methods improving raw detection performance and explainable AI research increasingly addressing the interpretability gap that has historically limited analyst trust in these models. The evidence reviewed here — from systematic reviews of ML in cybersecurity, rigorous empirical work on network-informed breach prediction, and a fast-growing literature on explainable deep learning for APT detection — demonstrates real, measurable progress, but also confirms that detection accuracy, prediction accuracy, and explainability continue to advance somewhat independently rather than as a fully integrated capability. Closing this gap, together with extending enterprise-architecture and governance maturity to match the pace of AI deployment, particularly for increasingly autonomous mitigation systems, represents the central challenge for the continued advancement of AI-driven cyber risk management.

References

1. Abdul Mutalib, N. H., Md Sabri, A. Q., Abdul Wahab, A. W., Mohd Faizal Abdullah, E. R., & Aldahoul, N. (2024). Explainable deep learning approach for advanced persistent threats (APTs) detection in cybersecurity: A review. Artificial Intelligence Review, 57(11). https://doi.org/10.1007/s10462-024-10890-4
2. Alshuaibi, A., Almaayah, M., & Ali, A. (2025). Machine learning for cybersecurity issues: A systematic review. Journal of Cyber Security and Risk Auditing, 2025(1), 1-16. https://doi.org/10.63180/jcsra.thestap.2025.1.4
3. Basireddy, S. (2022a). Pioneering a future-ready metadata-centric platform architecture for dynamic configuration intelligent process automation operational agility and sustainable enterprise modernization. SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology, 14(03), 415-429. https://doi.org/10.18090/samriddhi.v14i03.27
4. Basireddy, S. R. (2022b). Digitizing risk management through configurable enterprise workflow architecture. SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology, 14(4), 231-239.
5. Basireddy, S. (2023). Audit-ready compliance platform architecture for large language model regulated enterprises. SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology, 15(01), 226-232. https://doi.org/10.18090/samriddhi.v15i01.40
6. Event Streaming Architectures for High-Volume Transaction Processing. (2022). International Journal of Advance Industrial Engineering, 10(01), 1-8.
7. Hu, K., Levi, R., Yahalom, R., & Zerhouni, E. G. (2022). Supply chain characteristics as predictors of cyber risk: A machine-learning assessment (arXiv:2210.15785). arXiv. https://arxiv.org/abs/2210.15785
8. Kalakoti, R., Vaarandi, R., Bahşi, H., & Nõmm, S. (2025). Evaluating explainable AI for deep learning-based network intrusion detection system alert classification. In Proceedings of the 11th International Conference on Information Systems Security and Privacy (ICISSP 2025) (Vol. 1, pp. 47-58). SciTePress. https://doi.org/10.5220/0013180700003899
9. Loschilin, A. V., Yarikov, V. G., & Nikishova, A. V. (2024). Machine learning methods in predicting and preventing cyber attacks. NBI Technologies, 18(2).
10. Rjoub, G., Bentahar, J., Abdel Wahab, O., Mizouni, R., Song, A., Cohen, R., Otrok, H., & Mourad, A. (2023). A survey on explainable artificial

intelligence for cybersecurity. IEEE Transactions on Network and Service Management, 20(4), 5115-5140.
11. Sannidhanam, A. H. (2021). Real-time claims processing using event-driven architectures. International Journal of Technology, Management and Humanities, 7(01), 36-50. https://doi.org/10.21590/07.01.02
12. Sannidhanam, A. H. (2025). Autonomous AI agents for cloud infrastructure operations. Journal of Contemporary Science and Technology Management, 1(01), 83-100.
13. Tan, J. K., Ong, L.-Y., & Leow, M.-C. (2025). A comparative study of deep learning-based network intrusion detection system with explainable artificial intelligence. International Journal of Electrical and Computer Engineering, 15(4), 4109-4119. https://doi.org/10.11591/ijece.v15i4.pp4109-4119

HOW TO CITE THIS ARTICLE: Kumar V. (2025). AI-Driven Cyber Risk Management: A Review of Threat Detection, Prediction, and Mitigation Strategies. *Journal of Integrated Science, Technology and Management*, 1(2), 22-26.