An Explainable AI-Driven Framework for Autonomous Cyber Threat Detection and Intelligent Incident Response in Enterprise Computer Networks

Trailokya Ojha

Assistant Professor, Department of Computer Science and Engineering, Kathmandu University, Nepal

Abstract

Enterprise networks face attacks that are faster, more varied and more automated than human analysts can track unaided. Machine learning has improved threat detection, but most high-performing models remain black boxes, which makes analysts reluctant to trust them and regulators reluctant to accept automated responses. This article proposes an explainable artificial intelligence (XAI)-driven framework for autonomous cyber threat detection and intelligent incident response in enterprise computer networks. The framework combines machine learning detection, explanation methods that show why an alert was raised, large language models (LLMs) grounded through retrieval-augmented generation for investigation and response, and privacy-preserving techniques that allow collaborative learning without exposing sensitive data. Drawing on nine studies, including five published between 2020 and 2022 on XAI, cybersecurity data science and explainable intrusion detection, the article sets out a five-layer architecture and a graduated incident response workflow. It argues that autonomy in security operations should expand only as far as explanations allow humans to verify, audit and correct the system's decisions.

Keywords: explainable artificial intelligence; cyber threat detection; incident response; intrusion detection; large language models; federated learning; data privacy; enterprise networks

DOI: 10.21590/ijtmh.2024100421

1. Introduction

Security operations centres (SOCs) in large enterprises receive thousands of alerts a day from firewalls, endpoints, identity systems, cloud services and business applications. Most alerts are false positives, and the real threats are often visible only when weak signals across several systems are connected. Analysts therefore spend much of their time sorting noise, while attackers exploit the delay between intrusion and response.

Machine learning has been widely applied to this problem. Classifiers and anomaly detectors can spot patterns that signature-based tools miss, and they scale to data volumes no team could review manually (Sarker et al., 2020). Yet the most accurate models, such as deep neural networks and large ensembles, rarely explain their outputs. An analyst who sees only a risk score cannot tell whether it reflects a genuine attack, a data artefact or a misconfigured sensor. Without that understanding, organisations hesitate to let the model act on its own, and much of the speed advantage is lost.

Explainable AI addresses this gap by making a model's reasoning visible, for example by showing which features drove a decision or which rules apply (Barredo Arrieta et al., 2020). Large language models add another capability. They can read logs, threat intelligence and runbooks, summarise an incident in plain language and propose response steps (Cherukuri, 2023a). Combined, XAI and LLMs make it possible to move from alert triage towards autonomous detection and response that humans can still supervise.

Three complications remain. First, security data often contain personal and regulated information, so detection systems must respect privacy and retention rules (Asuri, 2022). Second, organisations benefit from learning from each other's attacks but cannot share raw data freely, which points to federated approaches (Asuri, 2024). Third, explanations can themselves be exploited by adversaries, so XAI must be deployed with care (Capuano et al., 2022).

This article proposes a framework that addresses these issues. Section 2 reviews the literature. Section 3 presents the framework. Section 4 describes an intelligent incident response workflow. Section 5 covers privacy, governance and trust. Section 6 discusses challenges and future research, and Section 7 concludes.

2. Literature review

The literature behind this framework covers four themes: machine learning for cybersecurity, explainable AI and its application to security, LLMs and generative AI, and privacy-preserving data governance.

2.1 Machine learning for cybersecurity

Sarker et al. (2020) provide an overview of cybersecurity data science from a machine learning perspective. They describe how security data from many sources can be analysed with techniques such as feature engineering, clustering, classification and association rule learning to support data-driven security decisions. They also propose a multi-layered framework for building machine learning-based security models. Their work establishes the detection foundation on which explainability and automation can be built.

2.2 Explainable AI and its role in security

Barredo Arrieta et al. (2020) offer one of the most cited reviews of XAI. They define explainability with explicit attention to the audience who needs the explanation. They also build taxonomies of

explanation methods, including a dedicated taxonomy for deep learning. Finally, they propose the concept of responsible AI, which places fairness, explainability and accountability at the centre of real-world deployment. For security operations, the audience-centred view matters: an explanation useful to a data scientist may be useless to a SOC analyst or an auditor.

Mahbooba et al. (2021) apply XAI to trust management in intrusion detection systems (IDS). They use decision tree models, which split decisions into human-readable rules, and extract those rules from the widely used KDD benchmark dataset. They also compare the decision tree approach with other state-of-the-art algorithms. The study shows that interpretable models can make IDS behaviour understandable to experts, which strengthens trust in detection outcomes.

Zhang et al. (2022) survey XAI applications across cybersecurity, including intrusion detection, malware detection and spam filtering. They note that AI-based defences outperform signature- and rule-based approaches but are mostly deployed as black boxes. This lack of transparency reduces human confidence, particularly as attacks become more diverse. They present their survey as a roadmap for applying XAI across security domains.

Capuano et al. (2022) also review XAI in cybersecurity but highlight a double-edged effect. Explanations improve security practice, yet they may also reveal information that adversaries can exploit to evade detection. This finding is important for any framework that exposes explanations widely, and it shapes the governance controls proposed in Section 5.

2.3 Large language models and generative AI

Cherukuri (2023a) examines autonomous cyber threat detection and incident response using XAI and LLMs in enterprise networks. The study combines the interpretive power of LLMs with explainability, so that automated detection and response remain transparent to analysts. It is the closest antecedent to the framework proposed here.

Cherukuri (2023b) applies generative AI and retrieval-augmented generation (RAG) to software engineering, automated code generation and intelligent debugging. Two points transfer to security operations. Grounding generation in a relevant knowledge base improves the accuracy of technical output. In addition, GenAI can help produce and debug detection rules, response scripts and integration code, which are a constant demand in SOC engineering.

2.4 Privacy and collaborative learning

Asuri (2022) establishes a model for enterprise data privacy governance across SAP ecosystems. It relies on context-aware masking, enforcement of retention rules across the data lifecycle, and compliance with regulations in multiple jurisdictions. Security telemetry often contains user identities, business data and personal information, so the same controls are required when such data feed detection models and LLMs.

Asuri (2024) proposes a privacy-preserving federated learning framework on SAP Business Technology Platform. It supports secure, scalable and collaborative predictive analytics across distributed enterprise environments. Applied to security, federated learning allows business units,

subsidiaries or partner organisations to train shared threat detection models without pooling raw logs.

Together, these studies supply the detection methods, explanation techniques, language-model capabilities and privacy controls needed for explainable autonomous defence. The next section combines them into one framework.

3. Proposed XAI-driven framework

The framework is organised as five layers, from raw telemetry to governed action.

1. **Privacy-aware data layer.** Logs and events from network devices, endpoints, identity providers, cloud platforms and business applications such as ERP systems are collected centrally. Before storage or analysis, personal and sensitive fields are masked according to context, and retention rules are applied by data type and jurisdiction (Asuri, 2022).
2. **Detection layer.** A combination of supervised classifiers, unsupervised anomaly detectors and correlation rules identifies suspicious activity (Sarker et al., 2020). Where several organisations or business units participate, models are trained through federated learning, so that each site learns locally and shares only model updates (Asuri, 2024).
3. **Explanation layer.** Every alert carries an explanation suited to its audience (Barredo Arrieta et al., 2020). For analysts, this means the features or rules that drove the decision, such as an unusual login location combined with a privilege change. Interpretable models such as decision trees are preferred where they perform adequately (Mahbooba et al., 2021). Post-hoc explanation methods are applied to more complex models.
4. **Investigation and reasoning layer.** An LLM, grounded through RAG in the organisation's runbooks, asset inventory, past incidents and threat intelligence, interprets the explained alert (Cherukuri, 2023a; Cherukuri, 2023b). It correlates related events, maps the activity to known attack techniques, estimates impact and drafts an incident summary with recommended actions. Each statement is linked to the evidence retrieved.
5. **Response and governance layer.** Recommended actions are classed by risk and reversibility. The layer executes approved playbooks, records every decision and its explanation, and feeds outcomes back to the detection and explanation layers for improvement.

Three design principles connect the layers. First, **explanation as a precondition for autonomy**: no action runs automatically unless its triggering alert has an explanation a human could verify. Second, **evidence-grounded reasoning**: LLM outputs must cite retrievable logs and documents rather than rely on the model's general knowledge. Third, **graduated explanation exposure**: detailed explanations are restricted to authorised analysts, reflecting the risk that adversaries could use them to evade detection (Capuano et al., 2022; Zhang et al., 2022).

4. Intelligent incident response workflow

The framework supports a graduated response that ties the degree of autonomy to risk.

Tier 1: Automatic containment. Low-impact, reversible actions run without approval when the alert is well explained and confidence is high. Examples include isolating a single endpoint, disabling a suspicious session token or blocking a known malicious IP address. The system logs the action and notifies analysts with the explanation.

Tier 2: Recommended action with one-click approval. Actions with wider business impact, such as suspending a privileged account or quarantining a server, are prepared by the LLM with a summary, evidence and expected impact. An analyst reviews the explanation and approves or modifies the action.

Tier 3: Analyst-led investigation. Novel, complex or high-impact incidents, such as suspected data exfiltration from a business-critical system, are led by people. The LLM supports them by assembling timelines, querying logs and drafting reports and regulatory notifications.

A typical incident illustrates the flow. An anomaly detector flags unusual data transfers from a finance server at night. The explanation layer shows that the volume, destination and user account are all atypical. The LLM retrieves the account's recent activity, finds a login from a new country and matches the pattern to credential theft. It then recommends disabling the account and isolating the server. Because the server is business-critical, the action falls into Tier 2, and an analyst approves it within minutes rather than hours.

After each incident, analyst feedback on explanations and recommendations is used to retrain models and refine retrieval content. This loop gradually raises the share of incidents that can safely move to lower tiers. GenAI can also draft new detection rules and playbook code from incident lessons, which engineers review before deployment (Cherukuri, 2023b).

5. Privacy, governance and trust

Three controls keep the framework trustworthy. **Privacy by design** means masking personal data before it reaches models or LLM prompts and enforcing retention on logs, embeddings and incident records (Asuri, 2022). In federated settings, model updates should be securely aggregated and access to them audited (Asuri, 2024). **Accountability** requires every automated action to be traceable to its alert, explanation, evidence and approver, in line with responsible AI principles (Barredo Arrieta et al., 2020). **Adversarial resilience** requires restricting detailed explanations to authorised roles and testing models against evasion attempts (Capuano et al., 2022).

6. Challenges and future research

Open challenges include the trade-off between accuracy and interpretability, the risk of LLMs producing incorrect conclusions, the computational cost of explaining every alert in real time, and

the shortage of realistic, labelled enterprise datasets. Future research should evaluate explanation quality with working SOC analysts, measure changes in response time and error rates, and test federated detection across organisations.

7. Conclusion

Autonomous cyber defence is achievable only if humans can understand what the system is doing. The proposed framework combines machine learning detection, audience-appropriate explanations, evidence-grounded LLM reasoning and privacy-preserving collaboration. It then applies a tiered response model that links autonomy to risk. By treating explanation as a precondition for automated action, enterprises can respond to threats at machine speed while keeping control, accountability and regulatory compliance.

References

Asuri, R. (2022). Establishing enterprise data privacy governance across SAP ecosystems through context-aware masking, lifecycle retention enforcement, and cross-jurisdictional regulatory compliance. *International Journal of Humanities and Information Technology, 4*(01-03), 352–375.

Asuri, R. (2024). A privacy-preserving federated learning framework on SAP Business Technology Platform for secure, scalable, and collaborative predictive analytics in distributed enterprise environments. *The Eastasouth Journal of Information System and Computer Science, 2*(01), 173–190.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Capuano, N., Fenza, G., Loia, V., & Stanzione, C. (2022). Explainable artificial intelligence in cybersecurity: A survey. *IEEE Access, 10*, 93575–93600. https://ieeexplore.ieee.org/document/9877919

Cherukuri, M. (2023a). Advancing autonomous cyber threat detection and incident response using explainable artificial intelligence and large language models in enterprise computer networks. *International Journal of Technology, Management and Humanities, 9*(01), 175–199. https://doi.org/10.21590/ijtmh.2023090111

Cherukuri, M. (2023b). Leveraging generative artificial intelligence and retrieval-augmented generation to improve software engineering, automated code generation, and intelligent debugging systems. *International Journal of Humanities and Information Technology, 5*(04), 140–163. https://doi.org/10.21590/ijhit.05.04.11

Mahbooba, B., Timilsina, M., Sahal, R., & Serrano, M. (2021). Explainable artificial intelligence (XAI) to enhance trust management in intrusion detection systems using decision tree model. *Complexity, 2021*, 6634811. https://doi.org/10.1155/2021/6634811

Sarker, I. H., Kayes, A. S. M., Badsha, S., Alqahtani, H., Watters, P., & Ng, A. (2020). Cybersecurity data science: An overview from machine learning perspective. *Journal of Big Data, 7*, 41. https://doi.org/10.1186/s40537-020-00318-5

Zhang, Z., Al Hamadi, H., Damiani, E., Yeun, C. Y., & Taher, F. (2022). Explainable artificial intelligence applications in cyber security: State-of-the-art in research. *IEEE Access, 10*, 93104–93139. https://doi.org/10.1109/ACCESS.2022.3204051