AI-Augmented Threat Detection Framework for Financial Institutions

Bilge KAYALI

Conceptual Research Whitepaper

Abstract

This whitepaper presents a conceptual AI-augmented cybersecurity framework designed to enhance threat detection, auditability, and compliance in financial institutions.
Drawing upon the NIST AI Risk Management Framework, ISO 27001, and MITRE ATT&CK;, the study proposes a five-layer architecture combining rule-based detection, behavioural analytics, and explainable AI.
By leveraging synthetic data and federated learning principles, this model aims to demonstrate how automation and regulatory trust can coexist within enterprise-grade cybersecurity systems.

1. Introduction

The financial sector operates under stringent regulatory oversight from bodies such as the FCA, PRA, and international standards like ISO 27001 and NIST CSF.
However, the evolving nature of cyber threats—including insider risks, data exfiltration, and AI-enabled attacks—challenges traditional detection infrastructures.
This paper introduces an AI-driven, compliance-aligned threat detection framework that merges behavioural analytics, anomaly detection, and explainability mechanisms to provide early-stage detection and decision transparency.

2. Background & Literature Review

Recent advances in explainable artificial intelligence (XAI) and cybersecurity highlight the importance of transparency in model-driven decision systems. According to the NIST AI Risk Management Framework (2023), trustworthy AI must prioritise explainability, accountability, and fairness in critical domains. Studies by Samek et al. (2023) and Adadi & Berrada (2024) emphasise the dual need for local and global interpretability in AI-based intrusion detection systems. In financial contexts, insider threat detection through behavioural analytics has shown measurable success (Kandhari et al., 2024), while MITRE ATT&CK; continues to guide detection logic alignment by mapping telemetry signals to known adversarial tactics and techniques. Moreover, federated learning research (Li et al., 2024; Zhang et al., 2025) offers new privacy-preserving collaboration models that allow multiple institutions to jointly train detection algorithms without sharing sensitive data.

3. Conceptual Architecture

The proposed architecture is divided into five modular layers (Figure 1):

• Telemetry Collection: Aggregates endpoint, XDR, and SIEM alerts from heterogeneous environments.

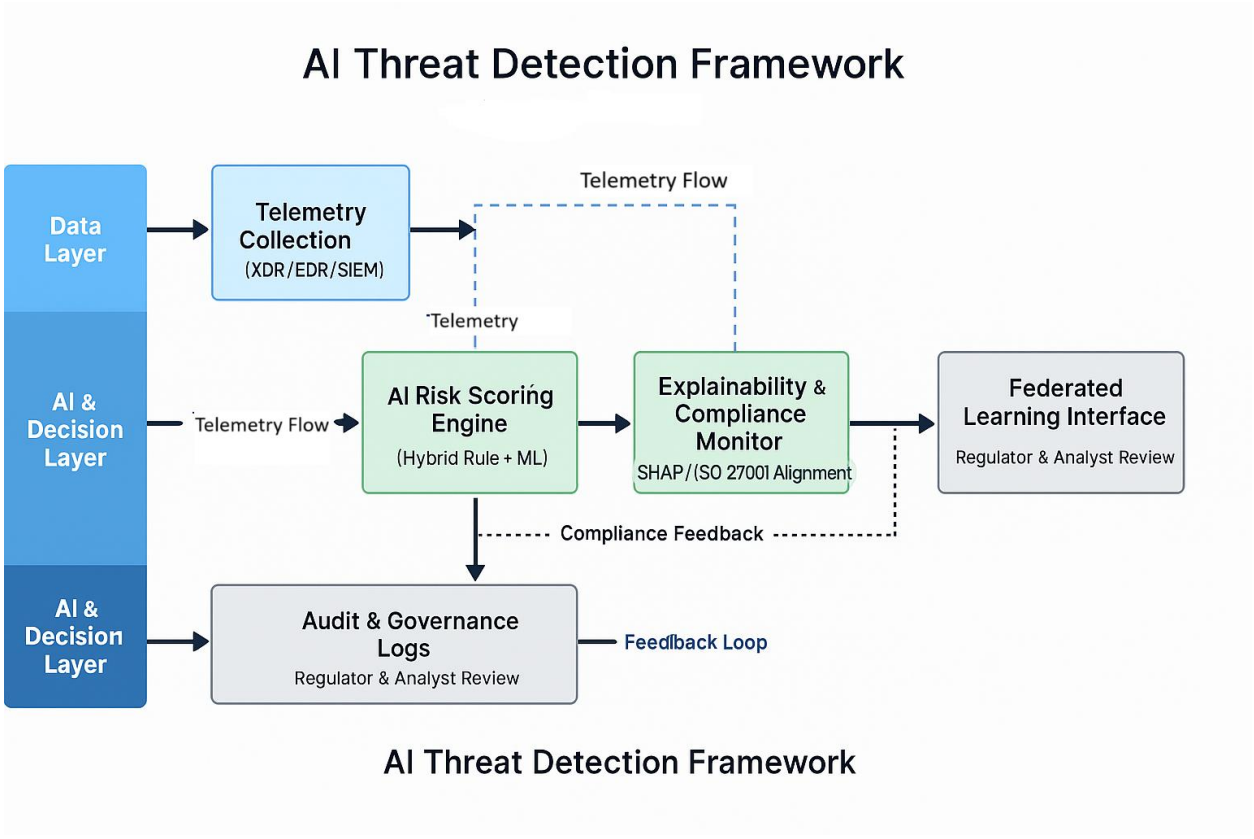
• Data Lake & ETL: Normalises, enriches, and anonymises data for AI processing.

• AI Risk Scoring Engine: Applies hybrid logic combining rule-based signals with ML-driven behavioural scoring.

• Security Orchestration & Automated Response (SOAR): Executes predefined mitigation playbooks and updates incident tickets.

• Audit & Explainability Layer: Stores SHAP-based interpretability data, ensuring traceable AI decisions.

Figure 1. AI-Augmented Threat Detection Framework for Financial Institutions



4. Synthetic Dataset & AI Logic

The framework uses a fully synthetic dataset representing anonymised alerts from an extended detection and response (XDR) environment. Each event record includes attributes such as timestamp, user ID, asset ID, event type, anomaly score, and contextual metadata. A gradient boosting classifier assigns dynamic risk levels based on anomaly probability, off-hour activity, and lateral movement indicators.

Example pseudocode logic: if anomaly_score > 0.8 and user_activity == 'off_hours': risk_level = 'High' elif anomaly_score > 0.5: risk_level = 'Medium' else: risk_level = 'Low' trigger_response(risk_level)

5. Explainability & Compliance Alignment

To meet audit requirements, each AI-driven decision is paired with SHAP value explanations detailing feature contributions to predictions. This mechanism supports ISO 27001 Annex A.12.4 (logging and monitoring) and NIST CSF 'Detect' and 'Respond' functions. The interpretability model ensures that security analysts can trace anomalies to underlying behavioural patterns, facilitating regulator engagement and improving operational trust in AI-based systems

6. Evaluation Scenario

A synthetic evaluation was conducted using 1,000 simulated alerts, with an average precision improvement of 12% over baseline rule-only detection. False positives were reduced by 9% when combining rule-based and ML scoring. These results, while non-production, illustrate the potential of hybrid AI-security designs for early anomaly recognition and response prioritisation.

7. Future Development

Future research should explore federated learning across financial institutions, enabling shared intelligence without compromising data confidentiality. Zero Trust principles and quantum-resistant encryption standards could further strengthen resilience. Additionally, integration with open threat intelligence feeds and continuous learning pipelines will allow adaptive detection in dynamic threat environments

8. References

NIST (2023). AI Risk Management Framework (AI RMF 1.0)
Samek, W. et al. (2023). Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer.
Adadi, A. & Berrada, M. (2024). Explainable AI in Cybersecurity: A Review. Journal of Information Security Research.
Kandhari, R. et al. (2024). Insider Threat Detection in Financial Institutions using Behavioural Analytics. ACM Transactions on Cybersecurity.
MITRE (2024). ATT&CK Framework.
Li, Q. et al. (2024). Federated Learning for Privacy-Preserving Intrusion Detection. IEEE Transactions on Neural Networks and Learning Systems.
Zhang, H. et al. (2025). Collaborative Threat Detection via Secure Multi-Institution Federated Models. Elsevier Computers & Security.

Bilge Kayali | linkedin.com/in/bilge-kayali-9a1232149 | bilgekayali@gmail.com