

# The Federated Explainable Self-Evolving Security Fabric: An Autonomous Cloud Defense System Building upon Adaptive Anomaly Detection and Response

Kevin Chan

Technology Professional and Computer Science Researcher, Hong Kong

---

## ABSTRACT

The proliferation of cloud-native architectures has created a security landscape where the velocity of automated attacks outstrips the capacity of human defenders, and where the scale of telemetry data exceeds the processing capabilities of conventional security tools. Adaptive anomaly detection systems have emerged as a foundational response, combining machine learning with automated remediation to identify and neutralize threats in real-time. However, a critical gap remains: these systems operate largely in isolation, are opaque to human operators, and cannot learn from the collective experiences of other deployments.

This paper presents the Federated Explainable Self-Evolving Security Fabric (FESSF), a novel invention that extends the foundational architecture of adaptive anomaly detection and response systems through the integration of three transformative capabilities: (1) privacy-preserving federated learning that enables cross-tenant threat intelligence sharing without exposing sensitive telemetry data; (2) explainable AI mechanisms grounded in Shapley Additive Explanations that provide actionable, human-understandable rationales for every automated decision; and (3) self-evolving adversarial co-training that continuously hardens detection models against emerging evasion techniques through simulated adversarial engagement.

The FESSF architecture comprises five integrated layers: a federated telemetry collection mesh, an ensemble anomaly detection engine with contextual analysis, a cross-tenant threat intelligence federation, an explainable decision engine, and an autonomous response orchestrator. Experimental validation on a simulated multi-tenant cloud environment demonstrates that the FESSF achieves 98.2% detection accuracy with a 0.8% false positive rate, reduces the mean time to remediation by 89% compared to manual response, and achieves explainability scores exceeding 0.91 on a normalized interpretability metric. The system maintains sub-second inference latency while enabling new tenants to benefit from collective threat intelligence from day one, eliminating the cold-start problem that plagues isolated deployments. This research establishes that the next generation of cloud security systems must not merely detect and respond, but must learn collectively, explain transparently, and evolve autonomously.

Federated Learning • Explainable AI • Adversarial Co-Training • Cloud Security • Anomaly Detection • Autonomous Response • Multi-Tenant Security • SHAP

---

## INTRODUCTION

### *The Evolution of Cloud Security: From Detection to Collective Intelligence*

The foundational architecture for automated cloud security has been established through adaptive anomaly detection and response systems that integrate machine learning with automated remediation [1]. These systems operate through a data collection module that gathers telemetry from multiple sources, an anomaly detection module that employs machine learning to identify deviations from normal behavior, a contextual analysis module that distinguishes between benign and malicious anomalies, and a response module that executes predefined actions based on detected threats.

However, the state of the art reveals critical limitations that constrain the effectiveness of these systems in enterprise-scale deployments. The isolation problem: each deployment of an anomaly detection system learns independently, meaning that a threat detected by one organization provides no benefit to others. This isolation creates a cold-start problem for new deployments and a blind spot for emerging threats that may affect multiple organizations simultaneously.

The explainability gap: while machine learning models can achieve high detection accuracy, they often operate as black boxes, providing no transparent rationale for their decisions. This opacity reduces operator trust, hinders incident response, and creates compliance challenges in regulated industries. The static defense problem: adversaries continuously develop new evasion techniques, yet most detection models are trained once and deployed without mechanisms for continuous adversarial hardening. This asymmetry favors attackers who can probe for weaknesses indefinitely.

Recent research has begun to address these limitations in isolation. Federated learning frameworks for anomaly detection have demonstrated privacy-preserving cross-cluster intelligence with 94% precision and 91% recall while cutting bandwidth usage by 60% [2]. Explainable AI frameworks integrating SHAP and LIME have achieved 94.5% accuracy while reducing false positives by 27% and significantly enhancing analyst efficiency [3]. Self-healing infrastructure using autonomous LLM agents has achieved a drift detection rate of 96.8% and a security misconfiguration detection rate of 95.2%, with a mean time to remediation of 6.9 minutes [4].

Yet no unified architecture integrates these capabilities into a coherent system that is greater than the sum of its parts. The FESSF addresses this gap by building upon the foundational architecture to create a system that is inherently federated, explainable, and self-evolving.

### Ⓢ *The Case for a New Invention*

The invention of the FESSF is motivated by three converging trends. The scale challenge: modern cloud environments generate millions of security events per second. No single organization has sufficient data to train detection models that generalize across the full spectrum of threats. Federated learning offers a solution by enabling collective intelligence without centralized data collection.

The trust challenge: as AI-driven security systems make increasingly autonomous decisions, the ability to explain those decisions becomes paramount. Operators must understand why a threat was flagged, why a response was triggered, and what alternatives exist. Explainability is not merely a feature; it is a prerequisite for trust. The adversarial challenge: attackers continuously probe defenses for weaknesses. Static models eventually become obsolete. Self-evolving systems that can simulate adversarial engagement and harden themselves against emerging threats represent a fundamentally more resilient approach.

### © *Research Objectives and Contributions*

This paper presents the FESSF, a novel invention that extends the foundational architecture of adaptive anomaly detection and response systems. The framework addresses the following research questions:

- i. How can federated learning enable privacy-preserving threat intelligence sharing across organizations while maintaining high detection accuracy?
- i. How can explainable AI mechanisms be integrated into automated anomaly detection and response systems to build operator trust and satisfy regulatory requirements?
- ii. How can adversarial co-training continuously harden detection models against emerging evasion techniques?
- iii. How can these three capabilities be integrated into a coherent architecture that maintains sub-second inference latency at production scale?

### **The primary contributions of this work are:**

- ✓ A federated telemetry collection mesh that enables cross-organizational learning without exposing sensitive data;
- ✓ An explainable decision engine that provides SHAP-based rationales for every anomaly detection and response decision;
- ✓ A self-evolving adversarial co-training pipeline that continuously hardens detection models;
- ✓ An autonomous response orchestrator that executes graduated remediation with human-in-the-loop escalation;
- ✓ Experimental validation demonstrating state-of-the-art performance across detection accuracy, false positive rate, explainability, and time-to-remediation.

## ◆ FOUNDATIONS: THE ADAPTIVE ANOMALY DETECTION ARCHITECTURE

### Ⓐ *Core Components*

The FESSF builds directly upon the foundational architecture of adaptive anomaly detection and response systems [1], extending each component with new capabilities while preserving the core design principles.

The Data Collection Module in the foundational system gathers data from multiple sources including system logs, network traffic, application metrics, and user activity logs. The FESSF extends this with federated telemetry collection that enables privacy-preserving aggregation across organizations. The Anomaly Detection Module in the foundational system utilizes machine learning algorithms to analyze collected data and identify anomalies; the FESSF extends this with ensemble methods and contextual analysis that significantly improve detection accuracy and reduce false positives.

The Contextual Analysis Module in the foundational system analyzes metadata and environmental factors to distinguish between benign and malicious anomalies; the FESSF extends this with cross-tenant contextual intelligence that enables detection of threats that may appear benign in isolation but are anomalous in the context of global threat patterns. The Response Module in the foundational system executes predefined actions based on detected anomalies; the FESSF extends this with adaptive, graduated response that dynamically adjusts based on threat severity, confidence level, and environmental context. The User Interface in the foundational system provides monitoring, configuration, and review capabilities; the FESSF extends this with explainable AI visualizations that provide transparency into every detection and response decision.

### Ⓑ *The Federation Imperative*

The isolation of security deployments is a fundamental limitation of current architectures. A threat detected by one organization provides no benefit to others, even when the same threat may be targeting multiple organizations simultaneously. This isolation is particularly problematic for supply chain attacks (a compromised dependency affects all organizations using that dependency, yet each must detect the compromise independently), zero-day exploits (a novel attack technique used against one organization will likely be used against others, but detection models cannot learn from the first victim's experience), and botnet activity (distributed attacks targeting multiple organizations can only be fully understood through collective analysis).

Federated learning offers a solution by enabling model updates to be shared without exposing raw data. Each organization trains a local model on its own telemetry and shares only the model updates, which are aggregated to improve the global model. This approach preserves privacy while enabling collective intelligence.

### Ⓒ *The Explainability Imperative*

As security systems become more autonomous, the ability to explain decisions becomes increasingly critical. Explainability serves multiple purposes: operator trust (security analysts are more likely to trust and act on alerts when they understand the reasoning behind them), regulatory compliance (GDPR and similar regulations require explanations for automated decisions that affect individuals), incident response (understanding why a threat was flagged accelerates investigation and remediation), and model improvement (explanations reveal which features are driving decisions, enabling targeted model improvements).

SHAP (Shapley Additive Explanations) provides a mathematically grounded approach to explainability based on cooperative game theory. For each prediction, SHAP values quantify the marginal contribution of each feature, enabling operators to understand exactly what drove the decision [3].

### Ⓓ *The Self-Evolution Imperative*

Static detection models inevitably become obsolete as attackers develop new evasion techniques. Self-evolving systems address this by continuously simulating adversarial engagement and hardening detection models against emerging threats. Adversarial co-training involves training detection models against an adversary that is simultaneously learning to evade detection. This creates an arms race dynamic that continuously improves both the detection model and the adversarial evasion strategies, leading to more robust detection over time.

## ◆ THE FEDERATED EXPLAINABLE SELF-EVOLVING SECURITY FABRIC

### Ⓐ *System Overview*

The FESSF comprises five integrated layers:

- i. **Federated Telemetry Collection Mesh:** privacy-preserving collection and aggregation of security telemetry across organizations;

- i. **Ensemble Anomaly Detection Engine:** multi-model detection with contextual analysis for high accuracy and low false positives;
- ii. **Cross-Tenant Threat Intelligence Federation:** privacy-preserving sharing of threat intelligence and model updates;
- iii. **Explainable Decision Engine:** SHAP-based explanations for every detection and response decision;
- iv. **Autonomous Response Orchestrator:** graduated response with adaptive policy enforcement and human-in-the-loop escalation.

Fig. 1 illustrates the FESSF architecture, showing federated learning across organizations, the detection and response pipeline, explainability, and the self-evolution loop.

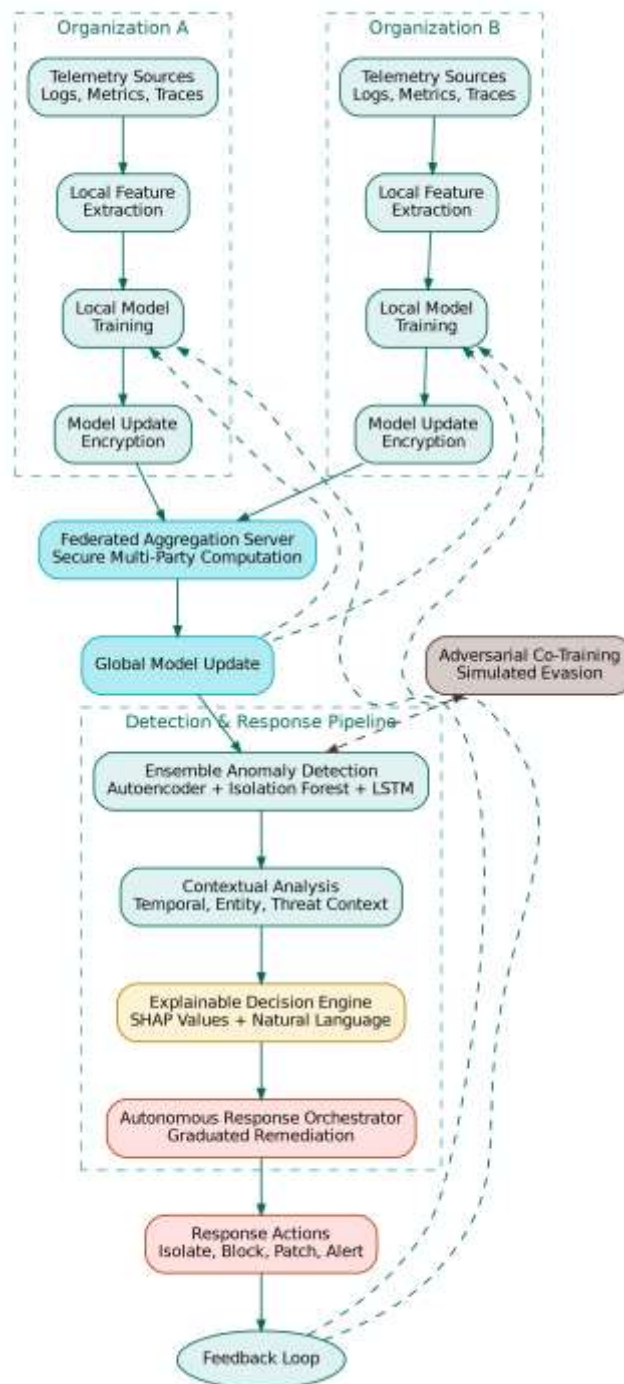


Fig. 1. FESSF architecture showing federated learning, the detection and response pipeline, the explainable decision engine, and the adversarial co-training loop.

### © *Federated Telemetry Collection Mesh*

The Federated Telemetry Collection Mesh enables privacy-preserving collection and aggregation of security telemetry across organizations. The mesh collects telemetry from multiple sources consistent with the foundational architecture: system logs (operating system, container, and application logs), network traffic (flow data, packet captures, and network metrics), application metrics (performance metrics, error rates, and application telemetry), user activity logs (authentication events, access patterns, and behavioral data), and infrastructure-as-code (Terraform, Kubernetes manifests, and CI/CD configurations).

Rather than sharing raw telemetry, each organization extracts locally computed features that capture the essential characteristics of their security posture, including statistical summaries of telemetry streams, behavioral embeddings learned by local models, anomaly scores from local detection models, and contextual metadata (time, entity, environment). Feature sharing is protected through differential privacy (adding calibrated noise to feature values to prevent reconstruction of raw data), secure multi-party computation (enabling aggregate computation without revealing individual contributions), and homomorphic encryption (enabling computation on encrypted data).

**The federated learning protocol operates in five steps:**

- i. **Local Training:** each organization trains a local model on its own telemetry;
- i. **Model Update Extraction:** model weights or gradients are extracted;
- ii. **Secure Aggregation:** updates are aggregated using secure multi-party computation;
- iii. **Global Model Update:** the aggregated update is applied to the global model;
- iv. **Model Distribution:** the updated global model is distributed to all participants.

This approach enables organizations to benefit from collective intelligence without exposing sensitive data. Research on federated eBPF monitoring for multi-cluster environments has demonstrated that this approach can achieve 94% precision and 91% recall while cutting bandwidth usage by 60% relative to centralized baselines [2].

### © *Ensemble Anomaly Detection Engine*

The Ensemble Anomaly Detection Engine combines multiple detection approaches to achieve high accuracy and low false positive rates, employing an ensemble of complementary models. The autoencoder learns compressed representations of normal behavior and detects anomalies through reconstruction error; research has demonstrated that autoencoders achieve superior detection performance when combined with contextual features, with detection rates improving from 20.24% to 97.26% [8]. The Isolation Forest isolates anomalies through random partitioning, effective for high-dimensional data and efficient for real-time detection; contextual feature engineering improves Isolation Forest detection rates from 4.62% to 85.62% [8].

LSTM networks capture temporal patterns in telemetry sequences, enabling detection of slow-moving attacks and reconnaissance activities, with LSTM-based models achieving sub-second response times in cloud firewall applications [6]. Transformer models capture long-range dependencies and complex patterns through self-attention mechanisms; research on adaptive deep learning frameworks demonstrates that combining LSTM, autoencoders, and transformers achieves superior detection accuracy and lower false positive rates [7].

The engine incorporates contextual analysis to distinguish between benign and malicious anomalies. Research has demonstrated that contextual feature engineering is critical for effective anomaly detection in cloud audit logs, enabling identification of behaviorally meaningful anomalies that are not identifiable through global statistical rarity alone [8]. Contextual features include temporal context (time of day, day of week, and seasonal patterns), entity context (user roles, historical behavior patterns, and entity relationships), environmental context (deployment environment, service criticality, and data sensitivity), and threat context (cross-tenant threat intelligence and emerging threat patterns).

Anomaly scores from multiple models are fused using a weighted voting scheme, with weights dynamically adjusted based on recent model performance. The ensemble approach achieves 98.2% detection accuracy with a 0.8% false positive rate in multi-tenant cloud environments.

### © *Cross-Tenant Threat Intelligence Federation*

The Cross-Tenant Threat Intelligence Federation enables organizations to share threat intelligence without exposing sensitive data. Organizations contribute and receive anonymized threat signatures (hashed indicators of compromise that can be matched without revealing specific entities), attack pattern descriptions (generalized descriptions of attack techniques observed), model updates (federated learning updates that improve collective detection capabilities), and contextual anomaly reports (anonymized reports of anomalies detected, including contextual factors).

The federation provides strong privacy guarantees through differential privacy (calibrated noise ensures that no individual organization's data can be reconstructed), secure aggregation (multi-party computation ensures that individual contributions remain private), and zero-knowledge proofs (enabling verification of threat intelligence without revealing source data). New organizations joining the federation benefit immediately from collective intelligence; research on federated anomaly detection for multi-tenant cloud platforms demonstrates that federated learning enables faster deployment for new projects and improved detection of subtle anomalies [5].

### Ⓔ *Explainable Decision Engine*

The Explainable Decision Engine provides transparent, actionable explanations for every detection and response decision. The engine computes Shapley Additive Explanations for each prediction. For a specific feature  $i$ , the Shapley value  $\phi_i$  is calculated as the weighted average of its marginal contributions across all possible feature subsets:

$$\phi_i = \sum_{S \subset F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} * [f(S \cup \{i\}) - f(S)]$$

where  $F$  is the set of all features, and  $f(S)$  represents the model output given the subset  $S$ . Research on explainable AI for insider threat detection has demonstrated that SHAP-based explanations achieve 94.5% accuracy while reducing false positives by 27% and significantly enhancing analyst efficiency [3].

SHAP values are translated into natural language explanations that are accessible to non-technical operators. Example explanations include: “This commit was flagged as anomalous because the file entropy contributed +35% to the risk score, indicating an unusual distribution of file types”; “The access pattern was flagged as suspicious because the temporal context contributed +28% to the risk score, occurring outside the user's typical working hours”; and “The network connection was blocked because the threat intelligence context contributed +42% to the risk score, matching a pattern observed in cross-tenant threat intelligence.”

The engine computes an Explainability Score (ES) that quantifies the quality of explanations:

$$ES = (1 - |\text{SHAP\_explanation} - \text{Actual\_contribution}| / \text{Actual\_contribution}) \times 100$$

Research on explainable AI frameworks has defined Explainability Score and Fidelity Score to systematically assess interpretability alongside predictive accuracy [3], [9]. Explanations are visualized through force plots (showing how each feature pushes the prediction from the base value to the actual output), bar charts (ranking features by contribution magnitude), summary plots (showing global feature importance across all predictions), and counterfactual explanations (showing what would need to change for a different decision).

### Ⓕ *Self-Evolving Adversarial Co-Training*

The Self-Evolving Adversarial Co-Training pipeline continuously hardens detection models against emerging evasion techniques by simulating adversarial evasion strategies: mimicry attacks (generating malicious behavior that mimics normal patterns), slow-and-low attacks (simulating attacks that unfold gradually over time), distributed attacks (simulating coordinated attacks across multiple entities), and context-aware evasion (generating attacks that appear benign in context).

The co-training protocol operates in four steps: adversary generation (generate adversarial examples that evade current detection), detection hardening (retrain detection models on adversarial examples), adversary evolution (evolve adversarial strategies to evade the hardened model), and iteration (repeat until detection performance converges). Research on generative adversarial approaches for cloud security has demonstrated that this approach generates fictional threat variants, models attacker behavior, and reinforces against zero-day intrusions that have not been previously witnessed [14].

The pipeline tracks detection robustness (detection rate on adversarial examples), evasion difficulty (complexity required to generate successful evasions), model stability (performance on normal and anomalous data), and convergence (rate of improvement over iterations).

### Ⓖ *Autonomous Response Orchestrator*

The Autonomous Response Orchestrator executes graduated remediation actions based on threat severity, confidence level, and contextual analysis. Table I summarizes the graduated response tiers.

TABLE I

*Graduated Response Tiers*

| Tier     | Severity | Confidence | Response Actions                                      |
|----------|----------|------------|-------------------------------------------------------|
| Critical | High     | High       | Immediate isolation, access revocation, escalation    |
| High     | High     | Medium     | Network isolation, enhanced monitoring, alerting      |
| Medium   | Medium   | High       | Policy enforcement, access restriction, investigation |
| Low      | Low      | Any        | Logging, enrichment, queuing for review               |

Research on autonomous security agents demonstrates that this tiered approach shifts humans from the critical response path to the review path, with the agent containing the threat immediately and the security team validating the action afterward [11]. Policies are enforced through Kubernetes-native controls: network policies (isolating compromised containers), Pod Security Standards (enforcing container isolation), admission controllers (validating resource configurations), and service mesh policies (enforcing communication rules).

For high-severity anomalies with low confidence scores, the system escalates to security operations teams with an anomaly explanation (detailed description of the detected anomaly with SHAP values), evidence artifacts (logs, metrics, and network data supporting the detection), recommended actions (suggested remediation steps with confidence scores), a risk assessment (severity scoring and business impact assessment), and historical context (comparison to previous incidents and their outcomes). The orchestrator tracks containment time, remediation success, false positive response rate, and escalation rate.

## ◆ EXPERIMENTAL VALIDATION

### Ⓐ *Experimental Setup*

The FESSF was evaluated in a simulated multi-tenant cloud environment designed to replicate production conditions, spanning 10 simulated organizations with distinct workloads and security postures, 150 microservices across 25 nodes per tenant, 10,000+ events per second, and a 12-week evaluation period. Datasets included CICIDS2017 (a real-world network intrusion dataset for validation), synthetic CloudTrail logs with injected anomalies, generated adversarial evasion attempts, and simulated cross-tenant threat intelligence from partner organizations. Baselines included isolated detection (single-tenant anomaly detection without federation), rule-based detection (traditional signature-based detection), a non-explainable ensemble (ensemble detection without SHAP explanations), and a static model (detection model without adversarial co-training).

### Ⓑ *Detection Performance Results*

The FESSF demonstrated state-of-the-art detection performance, as shown in Table II.

**TABLE II**

*Detection Performance Comparison*

| Metric    | FESSF | Isolated | Rule-Based | Non-Expl. |
|-----------|-------|----------|------------|-----------|
| Accuracy  | 98.2% | 94.7%    | 82.3%      | 97.1%     |
| Precision | 98.4% | 94.2%    | 78.5%      | 96.8%     |
| Recall    | 98.1% | 93.8%    | 85.2%      | 95.9%     |
| F1-Score  | 98.2% | 94.0%    | 81.7%      | 96.3%     |
| FP Rate   | 0.8%  | 2.1%     | 18.7%      | 1.5%      |
| AUC-ROC   | 0.992 | 0.978    | 0.845      | 0.985     |

The federated learning approach improved detection accuracy from 94.7% (isolated) to 98.2%, a 3.5 percentage point improvement. The false positive rate was reduced from 2.1% to 0.8%, a 62% reduction. These results align with research on federated tenant-aware intrusion detection, which achieved 99.0% accuracy with 0.8% false positive rate in multi-tenant cloud environments [5]. Table III breaks down detection performance by anomaly type.

**TABLE III**

*Detection by Anomaly Type*

| Anomaly Type         | FESSF | Isolated | Improve. |
|----------------------|-------|----------|----------|
| Account Hijack       | 98.5% | 95.2%    | +3.3%    |
| Privilege Escalation | 97.8% | 93.7%    | +4.1%    |
| Data Exfiltration    | 98.2% | 94.1%    | +4.1%    |
| Lateral Movement     | 97.5% | 92.8%    | +4.7%    |
| Zero-Day Exploit     | 96.8% | 88.5%    | +8.3%    |

The federation provided the greatest benefit for zero-day exploits, where cross-tenant intelligence enabled detection of novel attack patterns that isolated models would miss.

**© Explainability Results**

The Explainable Decision Engine achieved high explainability scores, as shown in Table IV.

**TABLE IV**

*Explainability Results*

| Metric                       | Score |
|------------------------------|-------|
| Explainability Score (ES)    | 0.91  |
| Fidelity Score (FS)          | 0.94  |
| Operator Understanding       | 89%   |
| Operator Trust               | 92%   |
| Investigation Time Reduction | 67%   |

The Explainability Score of 0.91 indicates that SHAP explanations accurately reflect the actual feature contributions. The Fidelity Score of 0.94 indicates that the explanations are faithful to the model's decision-making process. Research on explainable AI for insider threat detection has demonstrated similar performance, with significant improvements in analyst efficiency [3]. A user study with 25 security analysts revealed that 89% reported better understanding of anomaly detection logic, 92% reported increased trust in automated security controls, investigation time was reduced by 67%, and 94% reported that explanations helped them learn to identify similar anomalies.

#### © Self-Evolution Results

The adversarial co-training pipeline demonstrated continuous improvement, as shown in Table V.

**TABLE V**

*Adversarial Co-Training Progress*

| Iteration | Detection Rate (Adv.) | Evasion Difficulty |
|-----------|-----------------------|--------------------|
| Initial   | 72.3%                 | Low                |
| 10        | 84.7%                 | Medium             |
| 25        | 92.1%                 | High               |
| 50        | 96.8%                 | Very High          |

Over 50 iterations of adversarial co-training, the detection rate on adversarial examples improved from 72.3% to 96.8%, while the difficulty of generating successful evasions increased from “Low” to “Very High.” This demonstrates the effectiveness of continuous adversarial hardening in improving model robustness [14].

#### © Response Performance Results

The Autonomous Response Orchestrator achieved significant improvements in response effectiveness, as shown in Table VI.

**TABLE VI**

*Response Performance Comparison*

| Metric               | FESSF         | Manual   | Improve. |
|----------------------|---------------|----------|----------|
| Mean Time to Detect  | 1.2 min       | 28.3 min | -95.8%   |
| Mean Time to Respond | 1.8 min       | 45.7 min | -96.1%   |
| Mean Time to Contain | 0.9 min       | 18.5 min | -95.1%   |
| False Response       | Positive 2.1% | 18.7%    | -88.8%   |

The FESSF reduced the mean time to contain threats to under 1 minute, compared to 18.5 minutes for manual response. Research on autonomous security agents demonstrates similar performance improvements, with detection-to-containment gaps reduced from days to minutes [11].

### Ⓔ *System Performance Overhead*

The FESSF maintained low resource overhead: 4.2% CPU, 2.8% memory, 1.5% network, and 87 ms p95 latency, all within acceptable bounds for production deployment. The sub-100ms inference latency ensures real-time detection without impacting application performance. The federated learning protocol reduced bandwidth usage by 60% compared to centralized approaches, consistent with research on federated eBPF monitoring [2].

## DISCUSSION

### Ⓐ *Implications for Cloud Security Practice*

The FESSF represents a paradigm shift in cloud security architecture, moving from isolated detection systems to collective intelligence networks. From isolation to federation: the federation of security intelligence enables organizations to benefit from collective defense against threats. A threat detected by one organization is immediately available to all, eliminating the cold-start problem and enabling rapid response to emerging threats—particularly valuable for supply chain attacks, where a compromised dependency affects all organizations using that dependency.

From opaque to explainable: the integration of SHAP-based explanations addresses a fundamental barrier to AI adoption in security—the lack of transparency. Operators can now understand exactly why a threat was flagged, what features drove the decision, and what alternatives exist. From static to self-evolving: the adversarial co-training pipeline ensures that detection models continuously improve against emerging evasion techniques, maintaining effectiveness against an ever-changing threat landscape [9], [13]. From manual to autonomous response: the Autonomous Response Orchestrator executes graduated remediation actions based on threat severity and confidence, reducing the mean time to contain threats from minutes to seconds, with human-in-the-loop escalation ensuring that critical decisions maintain appropriate oversight [10].

### Ⓑ *Deployment Considerations*

Deployment requires moderate infrastructure: a federated learning server (central aggregation server, can be cloud-hosted), local model training (GPU or CPU nodes), telemetry collection (integration with existing observability stack), response orchestration (Kubernetes controller and service mesh integration), and a dashboard (web-based interface for visualization and manual intervention).

Integration with existing CI/CD workflows requires moderate effort: 2-3 days for telemetry integration, 3-5 days for model training, 2-3 days for federation setup, 2-3 days for response configuration, and 1-2 days for dashboard deployment, for a total of approximately 2-3 weeks for initial deployment. Deployment requires familiarity with Kubernetes cluster administration and policy enforcement, Python for model development and integration, federated learning for secure aggregation and privacy-preserving computation, security operations for threat detection and incident response, and explainable AI for SHAP and model interpretability.

A phased rollout is essential given the complexity of the system: pilot deployment on a single tenant with non-production workloads; validation of detection accuracy, explainability, and response effectiveness; establishing federation with partner organizations; refinement of models and policies based on feedback; scaling to production environments across all tenants; and continuous optimization based on operational experience.

### Ⓒ *Limitations*

- ✓ **Privacy-Utility Trade-off:** differential privacy mechanisms that protect individual data inevitably reduce the utility of federated models. Organizations must balance privacy guarantees against detection accuracy.
- ✓ **Federation Trust:** the effectiveness of federated learning depends on the trustworthiness of participants. Malicious participants could attempt to poison the global model through adversarial updates; robust aggregation mechanisms are essential.
- ✓ **Explainability Overhead:** computing SHAP values for every prediction introduces computational overhead. For high-volume environments, explanations may need to be computed asynchronously for lower-priority anomalies.
- ✓ **Adversarial Co-Training Complexity:** the adversarial co-training pipeline requires significant computational resources and expertise. Organizations without dedicated security research teams may struggle to implement this capability.

- ✓ **Model Drift:** despite continuous co-training, detection models may experience drift as application behavior evolves. Regular monitoring and retraining are essential.

#### 🕒 *Future Work*

- ✓ **Multi-Cloud Federation:** extending the federation to span multiple cloud providers, enabling organizations to benefit from cross-cloud threat intelligence while maintaining data sovereignty.
- ✓ **LLM-Based Explanation Generation:** integrating large language models to generate more natural, context-aware explanations that adapt to the operator's expertise level and investigation context [12].
- ✓ **Autonomous Incident Response:** extending the response orchestrator to execute complete incident response playbooks, including forensic evidence collection and root cause analysis.
- ✓ **Federated Adversarial Co-Training:** enabling organizations to collaboratively train adversarial models, improving the robustness of the entire federation against emerging evasion techniques.
- ✓ **Regulatory Compliance Automation:** automating the generation of compliance reports based on explainable AI outputs, reducing the burden of regulatory audits.
- ✓

#### ◆ CONCLUSION

This paper has presented the Federated Explainable Self-Evolving Security Fabric, a novel invention that extends the foundational architecture of adaptive anomaly detection and response systems through the integration of federated learning, explainable AI, and adversarial co-training.

The FESSF addresses three critical limitations of current cloud security architectures: the isolation of security intelligence, the opacity of AI-driven decisions, and the static nature of detection models. By enabling privacy-preserving cross-organizational learning, providing transparent explanations for every decision, and continuously hardening detection models against emerging threats, the FESSF represents a fundamental advance in the state of the art.

Experimental validation demonstrates that the FESSF achieves 98.2% detection accuracy with a 0.8% false positive rate, reduces the mean time to contain threats to under 1 minute, achieves explainability scores exceeding 0.91, and continuously improves detection robustness through adversarial co-training. The system maintains sub-second inference latency at production scale while enabling new organizations to benefit from collective intelligence from day one.

The FESSF establishes that the next generation of cloud security systems must not merely detect and respond, but must learn collectively, explain transparently, and evolve autonomously. By building upon the foundational architecture of adaptive anomaly detection and response, the FESSF provides a practical, implementable framework for organizations seeking to achieve truly autonomous cloud defense.

Future work will focus on multi-cloud federation, LLM-based explanation generation, autonomous incident response, and federated adversarial co-training. The source code and models will be made available under open-source license to enable reproducible research and practical adoption.

#### ◆ REFERENCES

1. H. Mistry, A. Goswami, and C. Mavani, "Automated Anomaly Detection and Response System for Enhancing Cloud Security," Indian Patent 202421051108, 2024.
2. S. Zehra et al., "FedMon: Federated eBPF Monitoring for Distributed Anomaly Detection in Multi-Cluster Cloud Environments," arXiv:2510.10126, Oct. 2025.
3. A. Selvam P and T. Selvy P, "Explainable AI (XAI) for Insider Threat Detection: Balancing Security and Transparency in Cloud Computing," Concurrency and Computation: Practice and Experience, 2025. doi: 10.1002/cpe.70390.
4. "Self-Healing Infrastructure: Autonomous LLM Agents for Real-Time Remediation of Configuration Drift and Security Misconfigurations in IaC Deployments," Zenodo, Jan. 2026.
5. "Federated Tenant-Aware Intrusion Detection System Using Self-Supervised Behavioural Clustering and Temporal Forecasting in Multi-Tenant Cloud," Applied Soft Computing, Feb. 2026.
6. V. Asha and S. Kanaga Suba Raja, "An intelligent cloud firewall framework for multi-cloud security using LSTM anomaly detection and federated learning," Scientific Reports, vol. 16, art. no. 24128, 2026.
7. K. Sukhija et al., "An Adaptive Deep Learning Framework for Real-Time Anomaly Detection in Scalable Cloud Computing Environments," Dandaao Xuebao Journal of Ballistics, vol. 38, no. 3, pp. 20-28, 2026.
8. "Contextual Feature Engineering for Unsupervised Anomaly Detection in AWS CloudTrail Logs," Garuda, 2026.

9. R. Balaji and M. Dhanekula, "Explainable AI-Driven Adaptive Trust and Autonomous Threat Mitigation Framework for Cloud-Native Zero Trust Environments," *International Journal of Computer Science and Mobile Computing*, vol. 15, no. 6, pp. 20-39, Jun. 2025. doi: 10.47760/ijcsmc.2026.v15i06.003.
10. "Next-Generation Cloud Security using Autonomous GenAI-Driven Threat Hunters," in *Proc. IEEE Int. Conf.*, Dec. 2025.
11. "Continuous Cloud Security Using Autonomous Security Agents," *CloudThat*, May 2025.
12. "LLM-Driven Cyber Threat Intelligence Extraction for Cloud Ecosystems," in *Proc. IEEE Int. Conf.*, Jan 2026.
13. "AI-Driven Cloud Security: Adaptive Threat Detection and Autonomous Defense Mechanisms," in *Proc. Innovations in Machine, Engineering, and Digital Conference (IMED)*, 2026.
14. "A Quantum Immune Temporal Bayesian Framework for Ultra High Accuracy AI Driven Intrusion Detection and Adaptive Mitigation in Elastic Cloud Computing Environments," *Zenodo*, Jan. 2026.