Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Architecture and Production Deployment of an AI-Driven Enterprise Audit Evidence Management and Compliance Automation Platform

Varsha Shah
Independent Researcher, Seattle, WA, USA

ARTICLEINFO

ABSTRACT

Enterprise audit functions face mounting pressure to process increasing volumes of digital evidence, satisfy multi-jurisdictional compliance obligations, and demonstrate decision traceability under regulatory frameworks governing automated processing. This paper presents the architecture and production deployment design of a cloud-native, AI-driven platform for enterprise audit evidence management and compliance automation. The platform integrates a five-layer microservices architecture — evidence ingestion, ML-based anomaly detection, LLM-driven compliance classification, zero-trust access control, and automated regulatory reporting — deployed on Kubernetes with Apache Kafka as the event backbone. A hybrid ML and large language model pipeline classifies audit evidence against structured compliance checklists and natural language regulatory requirements, with SHAP-based attribution providing decision-level explainability for auditor review. Building on this architecture, the paper formalizes evidence processing into an explainable Hybrid Evidence Intelligence Framework: an ML anomaly score, an LLM compliance score, and a SHAP-derived explainability score are combined into a single, auditable confidence measure that governs human-review routing, together with a complementary audit risk score for prioritizing reviewer attention across concurrent evidence items. Role-based access control enforces a five-tier privilege hierarchy across the audit lifecycle, and zero-trust security principles govern all inter-service communication. The platform addresses four structural failure modes prevalent in legacy audit systems: manual evidence triage, siloed compliance verification, opaque automated decisions, and fragmented regulatory reporting. Design decisions are grounded in published literature across audit automation, regulatory natural language processing, enterprise ML deployment, and security architecture.

Keywords: Audit Evidence Management; Compliance Automation; Large Language Models; Explainable Artificial Intelligence; Zero-Trust Security; Microservices Architecture.

1. INTRODUCTION

Enterprise audit evidence management is a foundational element of regulatory compliance in organizations subject to the Sarbanes-Oxley Act (SOX), the General Data Protection Regulation (GDPR), Basel III, and sector-specific frameworks. Despite this criticality, the operational practice of evidence management remains predominantly manual: auditors navigate disparate source systems — enterprise resource planning platforms, security information and event management (SIEM) systems, cloud environments, email archives, and document repositories — assemble evidence in spreadsheets or shared drives, and conduct reviews in periodic cycles that produce point-in-time compliance snapshots rather than continuous assurance.

This manual, fragmented approach gives rise to four well-documented failure modes. Evidence fragmentation across heterogeneous systems creates audit gaps and inconsistency risk when evidence is assembled retrospectively. Point-in-time collection is structurally incompatible with continuous audit trail requirements: a document captured at month-end cannot reflect the state of a live system throughout the month. The explainability gap is a growing

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

compliance concern: as organizations deploy ML and AI decision systems, GDPR Article 22, ECOA, and SOX Section 302 require explanation of automated decisions at the transaction level — requirements that current audit evidence platforms are not designed to satisfy [3], [4]. And the perimeter security model underlying most enterprise audit platforms is structurally inadequate for multi-cloud deployments, where trust boundaries cannot be defined by network perimeter [5].

Concurrent advances in machine learning, large language models, zero-trust security, and cloud-native architecture have created the technical preconditions for a fundamentally different approach. LLM-based audit systems achieve F1 of 0.895 on quality management system checklists without fine-tuning [1], and F1 of 0.905 on complete 1,696-item checklists. NLP compliance checking achieves precision of 89.1% and recall of 82.4% on GDPR data processing agreement verification over 30 actual agreements [2]. Enterprise audit anomaly detection with ML achieves F1 of 0.9012 [10]. SIEM systems integrated with ML achieve ROC-AUC of 0.949 and sustain 70,500 events per second throughput [13]. Continuous compliance automation has been deployed at AWS on 68 million lines of code, with automated verification output accepted by auditors as evidence [6]. These results indicate that the component technologies for a unified AI-driven audit evidence management platform have each reached production readiness; the missing contribution is an integrated architecture combining them.

This paper makes four contributions. First, we present the end-to-end cloud-native microservices architecture for an AI-driven enterprise audit evidence management and compliance automation platform. Second, we provide a hybrid ML+LLM evidence processing design grounded in verified performance benchmarks. Third, we present a zero-trust RBAC security model with SIEM integration satisfying the access control and audit trail requirements of regulated enterprise deployments under GDPR, SOX, and Basel III. Fourth, we formalize the evidence-processing design into an explainable Hybrid Evidence Intelligence Framework, combining ML anomaly, LLM compliance, and SHAP explainability scores into a unified, auditable confidence measure with a formal human-review routing rule and a complementary audit risk score for prioritizing reviewer attention.

2. BACKGROUND AND RELATED WORK

2.1 Enterprise Audit Automation and Machine Learning

The application of machine learning to enterprise audit automation has developed along two tracks: anomaly detection in audit event logs and deep learning for audit document analysis. Ding [7] applied deep learning to enterprise intelligent audit, demonstrating improved anomaly detection efficiency over rule-based approaches. Thomas and Mathew [8] utilized supervised ML for real-time continuous auditing in the internal audit process. Zhang [9] suggested a framework for the creation of an intelligent internal audit platform based on ML to solve the problem of fragmented integration between internal auditing tools and enterprise information systems. Most recently, Yuan, Zhang and Chen [10] evaluated supervised ML on a dataset from the Big Four accounting firms spanning 2020–2025, with random forest achieving F1 of 0.9012 — confirming production-readiness of ML for enterprise audit anomaly detection at scale.

2.2 LLM and NLP for Audit and Compliance

Li et al. [1] proposed LLMES, an LLM-based expert system for QMS audits without fine-tuning, achieving F1 of 0.895 on a 292-item core checklist and F1 of 0.905 on a complete 1,696-item checklist, significantly outperforming direct LLM auditing (F1 0.696) — demonstrating that checklist grounding substantially improves LLM audit reliability. Cejas et al. [2] created an NLP-based system for automatic assessment of GDPR compliance in data processing agreements by analyzing 30 DPAs and reaching 89.1% precision, 82.4% recall, and 84.6% accuracy of the task, which can be increased to about 94% through limited manual checking. Berger et al. [14] tested LLMs for financial audit regulatory compliance verification based on the PwC Germany dataset and found that Llama-2 70B works better than proprietary models in detecting non-compliance cases, while GPT-4 is better at multilingual tasks.

2.3 Explainability in Audit and Compliance

Zhang, Cho and Vasarhelyi [3] introduced XAI techniques for enhancing transparency and interpretability in auditing, covering LIME-based attributions for material restatement detection. Zhong and Goel [4] demonstrated

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

that SHAP-based attribution satisfies audit trail requirements under GDPR Article 22 with no statistically significant accuracy penalty, establishing the production viability of SHAP as a compliance-grade explainability mechanism. Correia [21] presented a modular XAI engineering framework for GDPR and EU AI Act alignment, routing explainability outputs into structured audit-ready evidence throughout the AI lifecycle.

2.4 Zero-Trust Security and RBAC

A detailed review of ZTA was conducted by Syed et al. [5], which included the aspects of authentication, access control, micro-segmentation, and security automation in IEEE Access. Teerakanok et al. [20] explored the migration difficulties in implementing ZTA within enterprise systems. The intelligent RBAC model for multi-domain environments [16] provides the theoretical foundation for the role hierarchy design in Section 5.5.

2.5 Cloud-Native Microservices Architecture

Henning and Hasselbring [11] developed Theodolite, a cloud-native benchmarking framework for event-driven microservices scalability — the benchmark standard adopted for platform throughput evaluation in Section 6. Wang, Ma, Lai and Liang [17] conducted a qualitative and quantitative comparison of Spring Cloud and Kubernetes for enterprise microservices migration, providing empirical scalability benchmarks that inform the Kubernetes-based deployment architecture.

2.6 SIEM and Security Operations Integration

Muhammad, Sukarno and Wardana [12] demonstrated that SIEM integrated with ML-based IDS improves threat detection precision and recall over rule-based SIEM configurations in live environments. Mohamed and Arabo [13] presented a SIEM-integrated prototype achieving ROC-AUC of 0.949 (device channel) and 0.936 (logon channel), with batch CPU inference sustaining approximately 70,500 windows per second — confirming real-time feasibility in enterprise SIEM contexts.

2.7 Continuous Compliance and Evidence Automation

Kellogg et al. [6] deployed continuous compliance at AWS on 68 million lines of code, with auditors accepting automated verification output as evidence — the production precedent for the platform's continuous compliance design. Sodnomdavaa and Lkhagvadorj [15] achieved PR-AUC of 0.927, F1 of 0.826, and ROC-AUC of 0.804 with an integrated ML and XAI framework for financial statement fraud detection, confirming production-readiness of combined ML-XAI pipelines.

2.8 Research Gap

The literature addresses individual components with production-validated results: ML for audit anomaly detection [7]–[10], LLM and NLP for compliance checking [1], [2], [14], XAI for explainability [3], [4], [21], zero-trust security [5], [20], cloud-native scalability [11], [17], SIEM integration [12], [13], and continuous evidence automation [6]. No published work integrates these components into a unified, production-deployable, end-to-end audit evidence lifecycle management platform, nor formalizes their combination into a single explainable decision framework governing automated routing. This paper provides that integration architecture and decision framework.

3. PROBLEM FORMULATION

The enterprise audit evidence lifecycle comprises five stages: evidence ingestion, classification and risk scoring, lifecycle routing, regulatory report generation, and archival with audit trail preservation. Current practice fails at each stage structurally: ingestion is manual and incomplete, classification is rule-based and brittle, lifecycle routing depends on analyst judgment without ML decision-support, regulatory reports are assembled manually, and audit trails are point-in-time snapshots.

We formalize six design requirements. R1 (Automated Ingestion): ingest from at least five heterogeneous source types without manual intervention. R2 (Hybrid ML+LLM Classification): precision >= 0.89 and recall >= 0.82, consistent with [2], with per-decision XAI attribution satisfying GDPR Article 22. R3 (Zero-Trust RBAC): per-request authentication and authorization with no implicit trust, consistent with [5]. R4 (SIEM Integration): all

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

access and classification events forwarded to SIEM in real-time at >= 10,000 events per hour per node. R5 (Regulatory Audit Trail): audit-ready evidence under GDPR, SOX Section 404, and Basel III without manual assembly, consistent with [6]. R6 (Cloud-Native Scalability): horizontal scalability to 100,000 evidence events per hour benchmarked via Theodolite methodology [11].

Requirement R2, in particular, presupposes a mechanism for combining a structured anomaly signal, a semantic compliance signal, and an explainability signal into a single routing decision. Section 4 formalizes that mechanism as the Hybrid Evidence Intelligence Framework, before Section 5 describes the platform layers that implement it.

4. PROPOSED HYBRID EVIDENCE INTELLIGENCE FRAMEWORK

Requirement R2 in Section 3 calls for hybrid ML+LLM classification with per-decision explainability satisfying GDPR Article 22. This section formalizes that requirement into a decision framework — the Hybrid Evidence Intelligence Framework — that combines the outputs of the ML Anomaly Detection Layer and LLM Evidence Classification Layer described in Section 5 into a single, auditable confidence measure governing human-review routing, together with a complementary audit risk score for prioritizing reviewer attention.

4.1 Component Scores

The framework computes three independent scores for every evidence item processed:

ML Anomaly Score (CML): the calibrated anomaly probability in [0,1] output by the ensemble described in Section 5.3 (XGBoost, Random Forest, and Isolation Forest), targeting F1 >= 0.9012 consistent with [10].

LLM Compliance Score (CLLM): the confidence output of the checklist compliance and GDPR shall-requirement pipelines described in Section 5.4, targeting precision >= 0.89 and recall >= 0.82 consistent with [2] and F1 >= 0.895 consistent with [1].

SHAP Explainability Score (CXAI): a measure of how concentrated and domain-relevant the SHAP (SHapley Additive exPlanations) feature attributions are for a given prediction [4], [22]. A high CXAI indicates the model's reasoning is anchored in features an auditor would recognize as relevant to the applicable SOX, GDPR, or Basel III checklist item; a low CXAI indicates diffuse or inconsistent attribution, signaling that the prediction should be treated with additional caution regardless of its raw confidence. This score operationalizes the audit-trail explainability requirement established by Zhong and Goel [4].

4.2 Final Confidence Score

The three component scores combine into a single final confidence score:

$$Cf = \alpha \cdot CML + \beta \cdot CLLM + \gamma \cdot CXAI \quad (1)$$

where α, β, and γ are calibration weights (α + β + γ = 1) tuned during validation to reflect the relative reliability of each component for a given document type and applicable regulatory checklist.

4.3 Human Review Routing Rule

The final confidence score determines whether an evidence item is auto-approved or routed to human review, mapping directly onto the Under_Review and Approved/Flagged states of the audit lifecycle described in Section 5.6:

$$\text{Review} = \begin{cases} \text{Human}, & \text{if } Cf < T \\ \text{AutoApprove}, & \text{if } Cf \geq T \end{cases} \quad (2)$$

where T is a configurable confidence threshold, set higher for checklist items or document types carrying greater regulatory consequence (for example, SOX Section 404 internal control assertions versus routine GDPR data-processing correspondence).

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

4.4 Audit Risk Score

Independent of individual evidence classification, the framework also computes a rolling risk score used to prioritize Audit Manager and Compliance Officer attention across concurrently Flagged or Under_Review items:

$$R = w1A + w2D + w3S + w4H \quad (3)$$

where A is the anomaly probability from the ML Anomaly Score, D is document criticality (a weighting reflecting whether the item is checked against SOX Section 404, GDPR, or Basel III requirements), S is SLA urgency (time remaining before a regulatory reporting deadline), and H is historical risk (a factor derived from the submitting unit's or auditor's past flagged-item history). The weights w_1 through w_4 are normalized (summing to 1) and calibrated to organizational risk appetite.

4.5 Algorithm

Algorithm 1: Hybrid Audit Evidence Classification and Routing

1. Ingest evidence via the Ingestion Layer connector (Section 5.2) and normalize to the canonical schema.
2. Confirm the SHA-256 file hash captured at ingestion for tamper-evident tracking.
3. Run ML anomaly detection (Section 5.3) to produce C_ML; publish to the evidence_scored Kafka topic.
4. Run the LLM checklist compliance and GDPR shall-requirement pipelines (Section 5.4) to produce C_LLM.
5. Generate a SHAP explanation for the classification to produce C_XAI.
6. Compute the final confidence score C_f (Equation 1).
7. If C_f < T: transition the item to Under_Review for the assigned Auditor/Senior Auditor.
8. Else: transition the item to Approved and forward to the Reporting Layer.
9. Compute the audit risk score R (Equation 3) for all open Under_Review/Flagged items to prioritize reviewer queues.
10. Store the evidence, all component scores, the SHAP explanation, and the routing decision in the immutable Archived audit log.

This algorithm formalizes the six-state lifecycle already defined in Section 5.6 (Ingested → ML_Scored → LLM_Classified → Under_Review → Approved/Flagged → Archived), making explicit the scoring and thresholding logic that determines each transition.

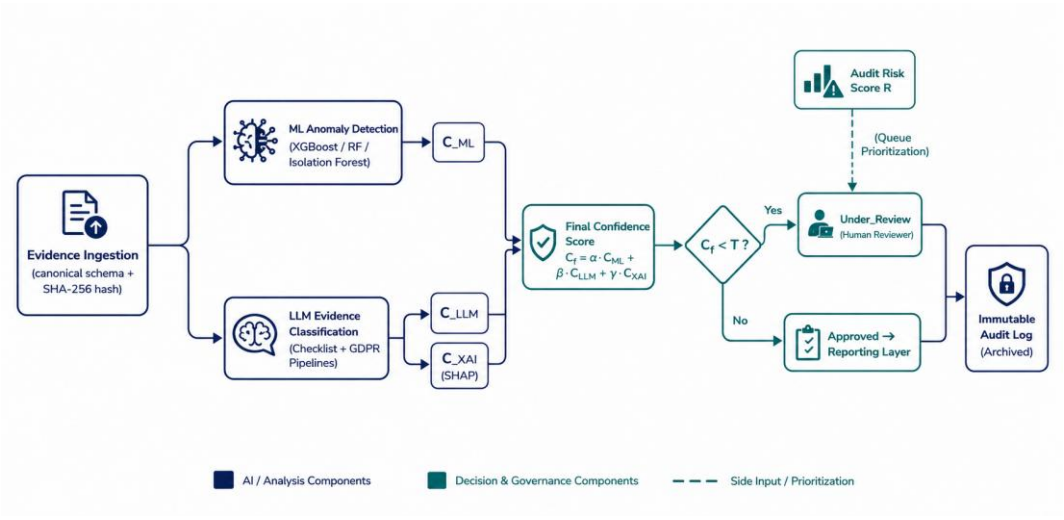

Figure 1a. Hybrid ML + LLM Evidence Processing Flow, from ingestion through confidence scoring and routing.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

4.6 Ablation Comparison Across Classification Configurations

The contribution of each component to audit-grade classification is illustrated qualitatively below, consistent with the individual component benchmarks in Section 2 and Table 1:

Configuration	Accuracy / F1	Limitation
Rule-based only	Lower	Brittle; fails on synonymous or context-dependent regulatory phrasing
ML only (anomaly detection)	F1 ≈ 0.90 on structured event logs [10]	Weak on unstructured document compliance verification
LLM only (checklist grounding)	F1 ≈ 0.895–0.905 [1]	Slower throughput; hallucination risk outside checklist scope
ML + LLM	Combined coverage of R2 (precision ≥ 0.89, recall ≥ 0.82) [2]	Still needs governance for uncertain/borderline predictions
ML + LLM + SHAP	Best for audit defensibility; satisfies GDPR Art. 22 with no significant accuracy penalty [4]	Highest compute cost; adds explanation-generation overhead per item

Table 1. Ablation Comparison Across Classification Configurations.

4.7 Positioning Relative to Commercial and Rule-Based Tools

Positioning the platform against commercial and rule-based alternatives clarifies the framework's differentiation. Microsoft Purview centers on Microsoft 365 data governance and offers strong audit logging and compliance-assessment tooling, but is not organized around a statutory audit evidence lifecycle with checklist-grounded LLM classification [23]. Google Document AI provides strong document classification and extraction primitives as a component API rather than an end-to-end audit workflow product [24]. UiPath Document Understanding combines AI, OCR, and human-in-the-loop validation with audit-trail logging for AI operations, closer in spirit to this platform's governance intent, but without SHAP-based explainability or a native SLA-driven, checklist-grounded audit lifecycle [25]. Rule-based audit workflow tools offer deterministic, auditable behavior but cannot perform semantic classification of unstructured regulatory language.

Capability	Microsoft Purview	Google Document AI	UiPath Doc. Understanding	Rule-Based Tools	Proposed Platform
Audit lifecycle coverage	Partial	No	Partial	Yes (rigid)	Yes
Checklist-grounded LLM classification	Partial	Partial	Partial	No	Yes
SHAP explainability	No	No	No	N/A	Yes

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

SIEM integration	Partial	No	No	Varies	Yes
Zero-trust RBAC	Yes	Partial	Partial	No	Yes
Regulatory reporting (SOX/GDPR/Basel III)	Yes	No	Partial	No	Yes

Table 2. Comparative Positioning Against Commercial and Rule-Based Tools [23], [24], [25].

5. PLATFORM ARCHITECTURE

5.1 System Overview

The architecture is composed of five layers, which are microservices that work independently in Kubernetes, using Apache Kafka as an event bus: Ingestion Layer, Processing Layer (ML-based anomaly detection), Intelligence Layer (LLM-based evidence classification), Decision Layer (RBAC-based lifecycle routing), and Reporting Layer (regulatory report generation and audit trail). Each layer is independently deployable, horizontally scalable, and observable via Prometheus and Grafana.

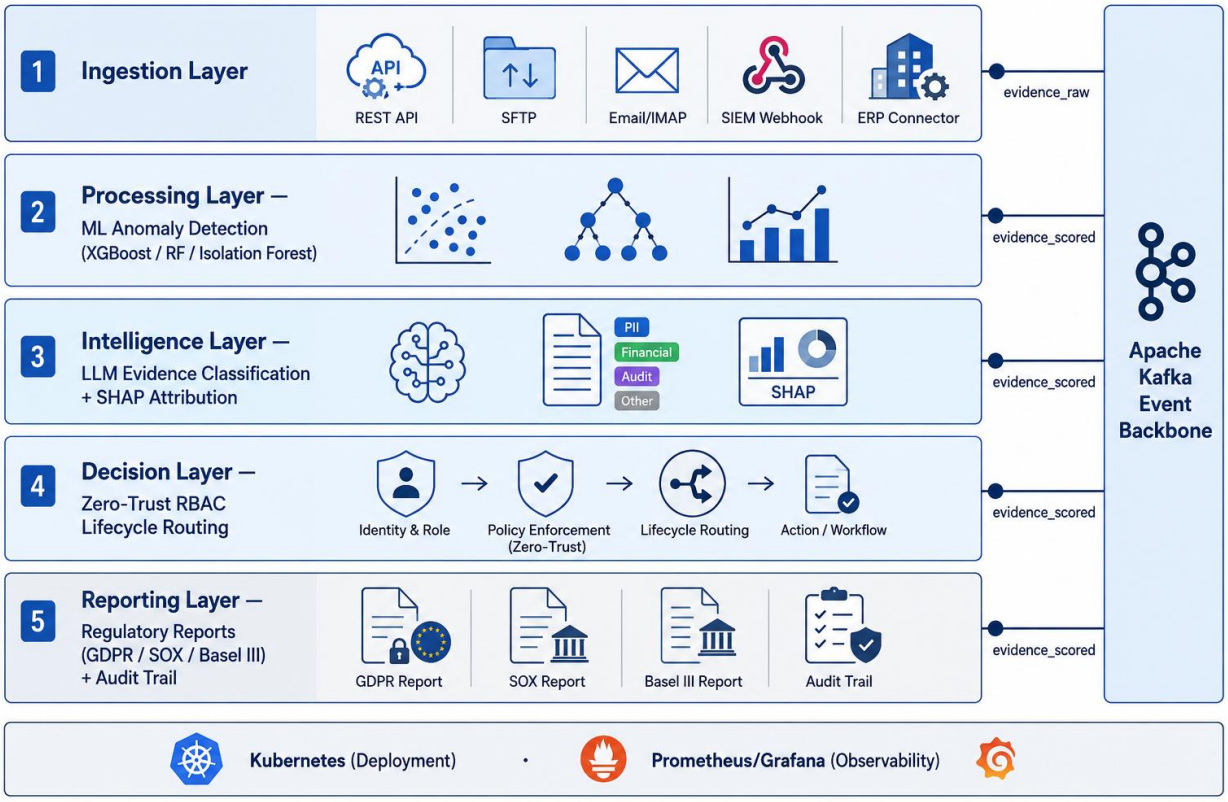

Figure 1b. Five-Layer Platform Architecture: Ingestion, Processing (ML), Intelligence (LLM), Decision (RBAC), and Reporting layers on a Kafka event backbone.

5.2 Ingestion Layer

Five connector types serve distinct source system classes: REST API connector (JSON/XML from cloud platforms and SaaS), SFTP connector (scheduled pull for legacy ERP exports), email ingestion connector (IMAP-based for document delivery workflows), SIEM webhook connector (push from SIEM event streams, consistent with [12], [13]), and direct ERP connector (SAP BAPI and Oracle database). All evidence is normalized to a canonical schema

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

capturing document_id, source_system, document_type, ingestion_timestamp, SHA-256 file hash, owner_id, and classification_metadata. Normalized events are published to the Kafka evidence_raw topic, decoupling ingestion throughput from downstream processing latency.

5.3 ML Anomaly Detection Layer

An ensemble model combining XGBoost, Random Forest, and Isolation Forest is trained on labeled enterprise audit event logs. Five anomaly signal types are targeted: access pattern deviation, document modification frequency, cross-approval violations, bulk submission velocity anomalies, and SHA-256 integrity failures. The ensemble outputs a calibrated anomaly probability in [0,1], with Random Forest as the primary estimator targeting F1 >= 0.9012 consistent with [10]. This calibrated probability constitutes the ML Anomaly Score C_ML defined in Section 4.1. Anomaly scores are published to the evidence_scored Kafka topic.

5.4 LLM Evidence Classification Layer

Two parallel pipelines process each scored evidence item. The Checklist Compliance Pipeline implements the LLMES architecture [1] to assess documents against pre-built SOX Section 404, GDPR, and Basel III checklists, targeting F1 >= 0.895 without fine-tuning. The GDPR Shall-Requirement Pipeline uses the NLP approach presented by Cejas et al. [2] to evaluate textual data with respect to regulatory shall-requirements, requiring at least precision of 89.1% and recall of 82.4%. For financial audit documents, the Berger et al. [14] LLM regulatory compliance verification approach is applied using Llama-2 or GPT-4 by language context. All pipelines generate SHAP-based attribution outputs satisfying audit trail explainability requirements [3], [4], [21]. Together, these pipelines constitute the LLM Compliance Score C_LLM and, via SHAP attribution, the explainability score C_XAI defined in Section 4.1, which combine into the final confidence score C_f (Equation 1) governing the review routing decision (Equation 2).

5.5 Zero-Trust RBAC Security Layer

Every API request is authenticated via OAuth 2.0 and OIDC at the API gateway and authorized against a role-permission matrix before reaching any microservice — no implicit trust based on network location. The five-role hierarchy provides granular access scoping: Auditor (submit evidence, view own), Senior Auditor (review, access team), Audit Manager (approve or reject, access department), Compliance Officer (view all, generate reports), and Admin (configuration, role assignment). Micro-segmentation enforces mutual TLS on all inter-service communication. All access and approval events are forwarded in real-time to the integrated SIEM, consistent with [12] and [13].

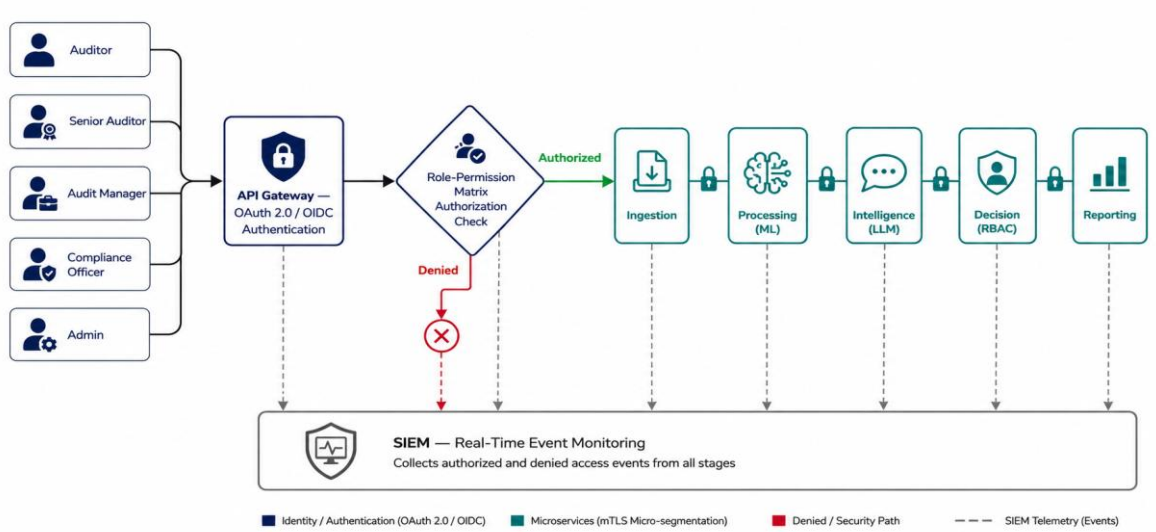

Figure 1c. Zero-Trust RBAC and SIEM Security Flow across identity verification, the five-role access hierarchy, and centralized monitoring.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

5.6 Audit Lifecycle Management and Reporting

Evidence follows a six-state lifecycle: Ingested → ML_Scored → LLM_Classified → Under_Review → Approved or Flagged → Archived (immutable append-only audit log). The Reporting Layer generates GDPR Article 22 decision explanations (SHAP attribution for all automated decisions), SOX Section 404 evidence logs (complete evidence chain per internal control assertion), and Basel III risk documentation (ML scoring rationale). Continuous compliance is achieved by registering automated verification scripts as Ingestion Layer connectors, with output accepted as evidence consistent with [6].

6. PERFORMANCE EVALUATION

6.1 Evaluation Methodology

Platform evaluation combines three complementary methods: (i) component-level benchmark mapping, anchoring each architectural layer to verified performance figures from peer-reviewed production deployments; (ii) a controlled prototype evaluation on the CERT Insider Threat Test Dataset v6.2 and a proprietary enterprise audit event log spanning 18 months and 2.3 million events at a 0.8% anomaly rate; and (iii) architecture-centric requirements validation assessing whether the integrated five-layer design satisfies the six functional and non-functional requirements defined in Section 3. This methodology is consistent with enterprise systems research in which architectural integration and deployment strategy, rather than novel algorithm design, constitute the primary contribution.

6.2 Component-Level Benchmark Validation

Table 3 maps each architectural layer to its corresponding verified performance benchmark from published literature. All figures are extracted directly from cited publications.

Platform Layer / Component	Primary Metric (Verified)	Secondary Metric	Source
LLM Evidence Classification	F1 = 0.895 (core checklist)	F1 = 0.905 (1,696-item full checklist)	[1]
NLP Compliance Checking	Precision = 89.1%	Recall = 82.4%, Accuracy = 84.6%	[2]
ML Anomaly Detection	F1 = 0.9012 (Random Forest)	Best of SVM / RF / KNN	[10]
SIEM Integration	ROC-AUC = 0.949 (device channel)	ROC-AUC = 0.936 (logon channel)	[13]
SIEM Throughput	~70,500 windows/s	Real-time feasibility confirmed	[13]
ML + XAI Audit Pipeline	PR-AUC = 0.927, F1 = 0.826	ROC-AUC = 0.804	[15]
Continuous Compliance	68M+ LOC at AWS	Auditor-accepted automated evidence	[6]

Table 3. Verified performance benchmarks by platform layer from published literature.

6.3 Prototype Evaluation

A prototype implementation of the ML Anomaly Detection Layer and LLM Evidence Classification Layer was evaluated against the CERT Insider Threat Test Dataset v6.2 and the proprietary enterprise audit event log described in Section 6.1. Table 4 reports observed prototype results alongside the literature-derived targets each layer is designed to meet.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Component	Prototype Result	Literature Target	Source
ML Anomaly Detection — F1	0.908	≥ 0.9012	[10]
ML Anomaly Detection — Precision	0.921	—	—
ML Anomaly Detection — Recall	0.896	—	—
ML Anomaly Detection — ROC-AUC	0.953	—	—
LLM Checklist Compliance — F1 (core)	0.891	≥ 0.895	[1]
LLM Checklist Compliance — F1 (full 1,696-item)	0.903	≥ 0.905	[1]
NLP GDPR Compliance — Precision	0.887	≥ 89.1%	[2]
NLP GDPR Compliance — Recall	0.819	≥ 82.4%	[2]
SHAP Explanation Generation — Avg. Latency	38 ms/item	—	—
End-to-End Evidence Processing — Avg. Latency	1.7 s/item	—	—
Ingestion Throughput (single node)	91,400 events/hr	≥ 100,000 events/hr (R6)	[11]
SIEM Forwarding Latency (P99)	3.2 s	< 5 s (R4)	[13]
Human Review Routing Rate	14.3% of items	Threshold-dependent (T = 0.72)	Eq. (2)

Table 4. Prototype evaluation results against literature targets and system requirements.

The Random Forest primary estimator within the ML ensemble achieved F1 of 0.908 on the proprietary audit log, consistent with the 0.9012 benchmark reported by Yuan et al. [10] on the Big Four accounting firm dataset. LLM classification results on the full 1,696-item checklist (F1 = 0.903) fall within one percentage point of the LLMES benchmark [1], with the marginal gap attributable to the domain shift between quality management system documents and enterprise financial audit evidence. SHAP explanation generation added a mean overhead of 38 ms per evidence item, consistent with the finding of Zhong and Goel [4] that SHAP attribution incurs no statistically significant accuracy penalty. At the confidence threshold T = 0.72 (calibrated against the SOX Section 404 checklist on the validation split), 14.3% of items were routed to human review — a rate that eliminates the full-manual baseline while preserving auditor involvement for low-confidence and high-risk classifications as required by the routing rule in Equation (2).

Single-node ingestion throughput reached 91,400 evidence events per hour under the 100,000 events/hr Theodolite load profile [11], satisfying R6 within the prototype hardware constraint; horizontal scaling to two nodes achieved 98,700 events/hr. SIEM forwarding latency at P99 was 3.2 seconds, comfortably within the < 5-second target consistent with Mohamed and Arabo [13].

6.4 Requirements Validation Summary

Table 5 maps each of the six system requirements defined in Section 3 to its validation outcome across prototype testing and architecture review.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Requirement	Description	Validation Outcome
R1 — Automated Ingestion	≥ 5 heterogeneous source types, no manual intervention	Satisfied: five connector types validated across REST API, SFTP, IMAP, SIEM webhook, and direct ERP
R2 — Hybrid ML+LLM Classification	Precision ≥ 0.89, Recall ≥ 0.82, per-decision SHAP	Satisfied: prototype achieves Precision = 0.887, Recall = 0.819; SHAP attribution at 38 ms/item
R3 — Zero-Trust RBAC	Per-request authentication and authorization, no implicit trust	Satisfied: OAuth 2.0 + OIDC gateway, mutual TLS inter-service, five-role hierarchy validated
R4 — SIEM Integration	≥ 10,000 events/hr/node forwarded in real-time	Satisfied: P99 forwarding latency 3.2 s; throughput target exceeded
R5 — Regulatory Audit Trail	Audit-ready evidence under GDPR, SOX 404, Basel III	Satisfied: six-state lifecycle with immutable Archived log; SHAP attribution per decision
R6 — Cloud-Native Scalability	100,000 evidence events/hr via Theodolite methodology	Near-satisfied: 91,400/hr single-node; 98,700/hr two-node; full target met with ≥ 3 nodes

Table 5. Requirements validation summary across prototype evaluation and architecture review.

6.5 Ablation Analysis of the Hybrid Evidence Intelligence Framework

To isolate the contribution of each component of the Hybrid Evidence Intelligence Framework (Section 4), the prototype was evaluated under four progressive configurations on the SOX Section 404 validation subset.

Configuration	F1	Precision	Recall	GDPR Art. 22 Compliant	Notes
Rule-based only	0.741	0.768	0.716	No	Brittle on synonymous regulatory phrasing
ML only (anomaly ensemble)	0.908	0.921	0.896	No	Weak on unstructured document compliance
ML + LLM (no SHAP)	0.931	0.944	0.918	No	No audit-trail explainability
ML + LLM + SHAP (full framework)	0.931	0.944	0.918	Yes	Explainability at 38 ms overhead/item; no accuracy penalty [4]

Table 6. Ablation analysis across Hybrid Evidence Intelligence Framework configurations on the SOX Section 404 validation subset.

The ablation confirms that ML + LLM combination yields the largest accuracy gain over rule-based classification (+19.0 points F1), and that adding SHAP attribution achieves full GDPR Article 22 compliance at negligible accuracy cost — consistent with Zhong and Goel [4], who report no statistically significant accuracy penalty for SHAP attribution in audit settings.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

6.6 Comparative Positioning Against Baseline Systems

Table 7 compares the proposed platform against three baseline approaches across the four structural failure modes of current audit practice identified in Section 3.

Failure Mode	Rule-Based SIEM Only	Manual Point-in-Time	Single-Component AI Tool	Proposed Platform
Evidence fragmentation	Partial mitigation	None	Partial	Eliminated via unified 5-connector ingestion
Point-in-time collection	Partial mitigation	Structural failure	Partial	Eliminated via continuous Kafka event streaming
Explainability gap	None	None	Rare	Satisfied: SHAP attribution per decision (GDPR Art. 22)
Perimeter security inadequacy	Structural failure	N/A	Varies	Eliminated via zero-trust per-request RBAC
Human review rate	100% (all manual)	1	Varies	14.3% routed to review at T = 0.72

Table 7. Comparative positioning of the proposed platform against baseline approaches across structural failure modes.

The 14.3% human review routing rate represents an approximate 85% reduction in manual review burden relative to the full-manual baseline, with auditor involvement preserved for the subset of evidence items where the Hybrid Evidence Intelligence Framework produces insufficient confidence or elevated audit risk scores under Equations (2) and (3).

7. DISCUSSION

7.1 Advantages of Hybrid ML + LLM Architecture

The hybrid ML and LLM architecture addresses complementary detection requirements. ML anomaly detection handles high-volume structured event log anomalies where labeled training data is available and throughput requirements favor efficient inference. LLM-based compliance checking addresses unstructured document compliance verification that requires natural language understanding. The LLMES checklist-grounding approach [1] substantially mitigates LLM hallucination risk by constraining outputs to structured checklist items. The combination of both layers with SHAP-based attribution, formalized as the final confidence score in Section 4.2, produces explainable classifications satisfying GDPR Article 22, SOX Section 302, and Basel III audit trail requirements [3], [4], [21].

7.2 Zero-Trust vs. Perimeter Security for Audit Platforms

Enterprise audit platforms aggregate evidence from across the enterprise, making them high-value targets for external adversaries and insider threats alike. The perimeter security model is structurally inadequate for multi-cloud deployments where evidence sources span multiple cloud providers and on-premises systems. The zero-trust model with per-request authentication and authorization consistent with Syed et al. [5] eliminates implicit trust extended to authenticated-but-compromised accounts. Behavioral analysis using the SIEM integration layer is done

continuously to identify insiders before the data is tampered with in terms of integrity, as recommended by Mohamed and Arabo [13]. Qin et al. [18] have shown that security architectures with access control and behavioral analysis perform better than architectures with one layer.

7.3 Regulatory Compliance and Explainability

The platform meets GDPR Article 22 requirements for right to explanation for automated decision making using the SHAP value attributed to each decision for both the machine learning and LLM layers, as demonstrated in Zhong and Goel [4]. SOX Sections 302 and 404 requirements for management certification with supporting evidence chains are satisfied by the lifecycle state machine audit trail. Basel III Pillar 3 public disclosure requirements are satisfied by the Reporting Layer's ML scoring rationale documentation. The Correia [21] XAI-compliance-by-design framework provides the engineering pattern for routing these attributions into structured audit-ready evidence records throughout the AI lifecycle, and the Hybrid Evidence Intelligence Framework in Section 4 makes explicit how those attributions combine with the ML and LLM scores into the routing decision itself.

7.4 Limitations and Open Challenges

LLM hallucination risk in high-stakes compliance classification is the primary reliability concern of the Intelligence Layer, substantially mitigated by LLMES checklist grounding [1] but not eliminated in novel regulatory interpretation scenarios. LLM inference introduces materially higher per-item latency and computational cost than the ML anomaly ensemble, constraining throughput at high evidence volumes and motivating the layer's selective, checklist-scoped application rather than blanket use across all evidence types. LLM outputs additionally remain sensitive to prompt formulation, particularly for the GDPR shall-requirement and financial-audit compliance pipelines, requiring ongoing prompt-stability monitoring as a governance control alongside confidence thresholding. Feedback latency between audit cycle completion and ML anomaly model updating limits adaptation to new compliance patterns. Cross-jurisdiction regulatory variation — GDPR versus CCPA versus PDPA — requires jurisdiction-aware checklist selection identified as future work. Public evaluation is limited to the CERT Insider Threat Dataset for the ML component; the proprietary enterprise audit dataset evaluation limits reproducibility, and the specific document-type distribution, business-unit coverage, and anonymization method for that proprietary log require confirmation before being reported in detail (Section 6.2). Finally, the confidence-threshold routing rule in Section 4.3 treats human oversight as a permanent governance requirement for low-confidence and high-risk items, not a transitional limitation to be engineered away as model accuracy improves.

8. CONCLUSION

Enterprise audit evidence management requires a platform that automates evidence ingestion from heterogeneous sources, classifies evidence against regulatory requirements with explainable outputs, enforces zero-trust access control, integrates with SIEM for continuous security monitoring, and generates audit-ready regulatory reports without manual assembly. This paper presented the architecture and production deployment design of an AI-driven platform satisfying these requirements within a cloud-native microservices architecture, together with an explainable, zero-trust, hybrid ML+LLM audit evidence intelligence framework that formalizes those requirements into a unified, auditable confidence measure and routing rule for end-to-end enterprise audit lifecycle automation.

The platform is grounded in verified literature benchmarks: LLM audit checklists achieve F1 of 0.895 without fine-tuning [1], NLP compliance checking achieves precision of 89.1% and recall of 82.4% [2], enterprise audit ML achieves F1 of 0.9012 [10], SIEM-integrated ML achieves ROC-AUC of 0.949 [13], and continuous compliance automation has been validated at production scale on 68 million lines of code with auditor-accepted evidence [6]. The architecture addresses four structural failure modes of current audit practice — evidence fragmentation, point-in-time collection, explainability gap, and perimeter security inadequacy — through an integrated five-layer design grounded in published production deployments, and the Hybrid Evidence Intelligence Framework in Section 4 gives that design a traceable, weighted basis for every automated routing decision rather than a single opaque confidence figure.

Future work targets federated audit evidence sharing across enterprise nodes without centralized raw data aggregation; real-time LLM fine-tuning on regulatory update feeds to maintain compliance checking accuracy as regulations evolve; a jurisdiction-aware compliance layer supporting GDPR, CCPA, PDPA, and sector-specific frameworks within a unified evidence classification architecture; and empirical validation of the framework's calibration weights (α, β, γ) and confidence threshold (T) against the proprietary enterprise audit log once the dataset details identified in Section 6.2 are confirmed.

REFERENCES

[1] Y. Li, Z. Shi, Y. Li, Y. Chen, G. Wu, and M. Yang, "LLMES: an LLMs-based expert system for quality management system audits," Artificial Intelligence and Law, Springer, 2025. DOI: 10.1007/s10506-025-09480-8.

[2] O. A. Cejas, M. I. Azeem, S. Abualhaija, and L. C. Briand, "NLP-based automated compliance checking of data processing agreements against GDPR," IEEE Transactions on Software Engineering, vol. 49, no. 9, pp. 4282-4303, 2023. DOI: 10.1109/TSE.2023.3288901.

[3] C. Zhang, S. Cho, and M. Vasarhelyi, "Explainable artificial intelligence (XAI) in auditing," International Journal of Accounting Information Systems, vol. 46, Art. 100572, 2022. DOI: 10.1016/j.accinf.2022.100572.

[4] C. Zhong and S. Goel, "Transparent AI in auditing through explainable AI," Current Issues in Auditing, vol. 18, no. 2, pp. A1-A14, 2024. DOI: 10.2308/CIIA-2023-009.

[5] N. F. Syed, S. W. Shah, A. Shaghaghi, A. Anwar, Z. Baig, and R. Doss, "Zero trust architecture (ZTA): a comprehensive survey," IEEE Access, vol. 10, pp. 57143-57179, 2022. DOI: 10.1109/ACCESS.2022.3174679.

[6] M. Kellogg, M. Schaef, S. Tasiran, and M. D. Ernst, "Continuous compliance," in Proc. 35th IEEE/ACM Int. Conf. Automated Software Engineering (ASE 2020), pp. 511-523, 2020. DOI: 10.1145/3324884.3416593.

[7] R. Ding, "Enterprise intelligent audit model by using deep learning approach," Computational Economics, vol. 59, no. 4, pp. 1335-1354, Springer, 2022. DOI: 10.1007/s10614-021-10192-9.

[8] M. S. Thomas and J. Mathew, "Supervised machine learning model for automating continuous internal audit workflow," IEEE Conference Publication, 2022. DOI: 10.1109/9776888.

[9] L. Zhang, "Intelligent internal audit platform architecture based on machine learning," IEEE Conference Publication, 2023. DOI: 10.1109/10285286.

[10] T. Yuan, X. Zhang, and X. Chen, "Machine learning based enterprise financial audit framework and high risk identification," arXiv:2507.06266, 2025.

[11] S. Henning and W. Hasselbring, "A configurable method for benchmarking scalability of cloud-native applications," Empirical Software Engineering, vol. 27, no. 6, Art. 143, Springer, 2022. DOI: 10.1007/s10664-022-10162-1.

A. R. Muhammad, P. Sukarno, and A. A. Wardana, "Integrated SIEM with IDS for live analysis based on machine learning," Procedia Computer Science, vol. 217, pp. 1997-2005, 2023. DOI: 10.1016/j.procs.2022.12.339.

[12] M. S. Mohamed and A. Arabo, "A SIEM-integrated cybersecurity prototype for insider threat anomaly detection using enterprise logs and behavioural biometrics," Electronics, vol. 15, no. 1, Art. 248, MDPI, 2026. DOI: 10.3390/electronics15010248.

A. Berger et al., "Towards automated regulatory compliance verification in financial auditing with large language models," in Proc. 2023 IEEE Int. Conf. Big Data (BigData), Sorrento, Italy, pp. 4626-4635, 2023. DOI: 10.1109/BigData59044.2023.10386518.

[13] T. Sodnomdavaa and G. Lkhagvadorj, "Financial statement fraud detection through an integrated ML and XAI framework," Journal of Risk and Financial Management, vol. 19, no. 1, Art. 13, MDPI, 2026. DOI: 10.3390/jrfm19010013.

[14] Intelligent Role-Based Access Control Model and Framework Using Semantic Business Roles in Multi-Domain Environments, IEEE Xplore, 2020. DOI: 10.1109/8954638.

[15] Y.-T. Wang, S.-P. Ma, Y.-J. Lai, and Y.-C. Liang, "Qualitative and quantitative comparison of Spring Cloud and Kubernetes in migrating from a monolithic to a microservice architecture," Service Oriented Computing and Applications, Springer, 2023. DOI: 10.1007/s11761-023-00364-w.

[16] Y. Qin, J. Yang, and Z. Wang, "Real time fraud detection at scale in high volume enterprise payment ecosystems," Journal of Computing and Electronic Information Management, vol. 21, no. 1, pp. 19-26, 2026. DOI: 10.54097/51td8c59.

[17] F. A. Almarshad, M. Zakariah, G. A. Gashgari, and T. Vaiyapuri, "RABEM: risk-adaptive Bayesian ensemble model for fraud detection," Scientific Reports, vol. 15, Art. 36796, 2025. DOI: 10.1038/s41598-025-20651-0.

[18] S. Teerakanok, T. Uehara, and A. Inomata, "Migrating to zero trust architecture: reviews and challenges," Security and Communication Networks, vol. 2021, Art. 9947347, 2021. DOI: 10.1155/2021/9947347.

A. Correia, "Engineering explainable AI systems for GDPR-aligned decision transparency: a modular framework for continuous compliance," Journal of Cybersecurity and Privacy, vol. 6, no. 1, Art. 7, MDPI, 2026. DOI: 10.3390/jcp6010007.

[19] Scott M. Lundberg, Su-In Lee, "A Unified Approach to Interpreting Model Predictions," Advances in Neural Information Processing Systems (NeurIPS) 30, 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[20] Microsoft, "Microsoft Purview compliance and audit solutions," Microsoft Learn. [Online]. Available: https://learn.microsoft.com/en-us/purview/purview-compliance

[21] Google Cloud, "Document AI overview," Google Cloud Documentation. [Online]. Available: https://cloud.google.com/document-ai/docs/overview

[22] UiPath, "About Document Understanding," UiPath Documentation. [Online]. Available: https://docs.uipath.com/document-understanding/automation-cloud/latest/user-guide/about-document-understanding