International Journal of Emerging Research in Engineering and Technology
Pearl Blue Research Group| Volume 6 Issue 4 PP 92-101, 2025
ISSN: 3050-922X | https://doi.org/10.63282/3050-922X.IJERET-V6I4P112

Original Article

AI-Driven Identity Threat Detection and Response Systems for Modern Cloud Security Operations Centers

Venkata Nagendra Satyam[1], Devisharan Mishra[2], Braja Gopal Mahapatra[3]
[1,2,3,] Senior Technical Program Manager

Received On: 08/09/2025 Revised On: 20/09/2025 Accepted On: 13/10/2025 Published On: 03/11/2025

Abstract - *Identity is the new perimeter in cloud-first enterprises, where adversaries increasingly weaponize stolen credentials, malicious OAuth consents, token replay, and over-permissioned roles to traverse control planes and SaaS estates with minimal endpoint noise. In this paper, a telemetry-based Identity Threat Detection and Response (ITDR) architecture integrating identity provider (SSO/OAuth/OIDC), cloud API (AWS/Azure/GCP), endpoint, and SaaS audit log telemetry into a privacy-conscious feature fabric is proposed using AI. Our combination of sequential modelling of session dynamics, graph learning of privilege, relationship abuse, and self-supervised representation learning reveals low-and-slow compromise patterns. An anomaly strength is combined with contextual role criticality of data, device trust, and just-in-time elevation, session isolation, token revocation, and step-up authentication executed by a policy-as-code and SOAR playbooks-based risk layer that is calibrated. Details Implementation summary Streaming feature stores (training/serving parity) Feedback loops (with analyst adjudications) Governance (mapped to ISO 27001, NIST SP 800-207, and GDPR) In hybrid real-world and synthetic tests, the method does better recall at constant low false-positive rates, as well as fivefold trimming alert volume by deduplication and model calibration and with less time-to-detect and time to respond with real-time inference and automated containment. Describe stability to adversarial log poisoning and concept drift, and explainability methods that do not compromise auditability. The findings reveal that AI-based ITDR has the potential to support SOC effectiveness in a material manner without going against the principles of Zero Trust nor regulatory requirements.*

Keywords - *Identity Threat Detection and Response (ITDR), Zero Trust, UEBA, SOAR, CIEM, Policy-as-Code, Explainable AI.*

1. Introduction

Identity-based enterprise attack surface Cloud-first operating models and hybrid work have made the networks the second front in enterprise attack surface. Malicious users are progressively using such poor/stolen credentials, token replay, MFA fatigue, consent-grant abuse, and over-permissions to navigate cloud control planes and SaaS estates without much noise. [1-3] Conventional SOC tooling that is focused on perimeter telemetry and fixed rules finds it hard to relate dissimilar identity signs across IdPs (SSO/OAuth/OIDC), IaaS APIs, endpoints and business apps resulting in alert fatigue, blind spots and slowing containment. In the meantime, machine-to-machine identities (service accounts, work load identities, CI/CD tokens) increase privilege paths, and least privilege and continuous verification are hard to operationallyize at scale.

Identity Threat Detection and Response (ITDR) powered by AI repositions SOC processes on identity situation and conduct. ITDR uses sequential patterning (to detect abnormalities during session) and graph learning (to detect privileges escalation) and self-supervised learning (to study low-and-slow abuse with no labeled examples) by connecting authentication and authorization via graph structure, device posture via graph structure, and cloud activity via graph structure. The signals and information, including the geo-velocity, unusual resource access, consent and grant deltas, and role inheritance dynamics are combined into the risk scores that trigger risk-adaptive controls: just-in-time elevation, step-up authentication, identity isolation, and privilege revocation implemented through SOAR playbooks and managed as policy-as-code. The paper provides motivation on why ITDR is required in contemporary SOCs, defines a privacy-conscious and auditable architecture, and gives evaluation measures in terms of time-to-detect and time-to-contain as well as a false-positive rate at a fixed recall to indicate the value of the operation. Using the alignment of detection to the concept of Zero Trust and CIEM insights, ITDR bridges the division between identity compromise and automated and explainable response.

2. Literature Review

2.1. AI and ML in Security Operations

AI/ML has changed SOC workflows into learning-based inference, as opposed to rule-based detection. Sophisticated pipelines combine heterogeneous telemetry identity provider entries, cloud control-plane APIs, EDR/NDIR notifications and SaaS audit events into feature stores to infer near-real-time. [4-6] Sequence models (e.g., versions of LSTM/Transformer) are trained to learn session dynamics/access paths; representation learning (autoencoders, contrastive/self-supervised encoders) to learn normal identity posture to emphasize subtle changes; and gradient-boosted trees can be useful with tabular risk scoring at scale. The models are effective in reducing the triage load by grouping the similar alerts, enriching automatically with the threat intelligence, and prioritizing the incidents according to their likely business impact.

To operationalize AI within the SOC, it is necessary to have an effective MLOps: data quality (schema checks, PII minimization), drift and performance monitoring, champion/challenger deployment, and human-in-the-loop adjudication which drives active learning loops. The measurable benefits will usually be manifested in the time-to-detect (TTD), time-to-contain (TTC), number of analyst alerts on a case and false-positive rate (FPR) on fixed recall measures that directly translate to the reduced risk and staffing efficacy.

2.2. User and Entity Behavior Analytics (UEBA) Techniques

UEBA sets standards of identities (humans, service accounts, workloads, devices) and indicates statistically significant changes. Basic techniques are probabilistic profiling (seasonality, geo-velocity, device hygiene), graph-based context (role inheritance, trust relationships), and temporal modeling of sequence of the sessions. The unsupervised and semi-supervised methods are popular due to the limited availability of labeled attack information: the use of isolation forests, one-class SVMs, variational autoencoders, and density-based clustering. State-of-the-art UEBA is a graph neural network (to reason about privileges and relationships) with sequential models (to capture tactic chains like consent-grant - token misuse - API enumeration). To reduce the noise, UEBA implements risk aggregation on signals and recalibrates thresholds using the results of analysts to reduce false positives and raise low-and-slow identity abuse.

2.3. Threat Detection Frameworks in Cloud Environments

Cloud threat models take traditional kill-chain models and apply them to ephemeral infrastructure, which relies on APIs. The strategies and techniques revolve around cloud-native telemetry CloudTrail/Azure Activity/GCP Admin logs, Identity Provider (OIDC/SAML) events, serverless and container runtime metrics and normalize them (e.g., OCSF) to correlate. Hypothesis-based hunting takes advantage of the detections that are mapped onto cloud-specific attack methods (misconfigured trust relationships, excessively permissive roles, token replay, cross-account assumption). Since assets and policies evolve quickly, detections should be configuration aware (CIEM insights), identity based as well as near real time. Publ/sub and feature store streaming architectures support sliding-window analytics, and policy-as-code (e.g., OPA/Rego) quarantine responses, step-up auth, just-in-time (JIT) revocation of elevation are all deterministic, auditable, and Zero Trust consistent.

2.4. Identity-Based Attack Patterns and Case Studies

More recent efforts indicate a shift towards the identity-based, human-operated tradecraft: MFA fatigue and on-demand bombing, help-desk social engineering, SIM swapping, malicious OAuth consent grants, and the abuse of long-lived CI/CD tokens. Attackers like privilege escalation through role- chaining, persistence through access key sprawl or service-principal secrets and stealth through API-only actions that cannot be sensed by endpoint sensors are preferred in cloud control planes. Case narratives consistently reveal three gaps: (1) fragmented visibility between IdP, cloud, and SaaS; (2) over-permissioned roles and stale machine identities; and (3) slow, manual response paths. Programs utilizing UEBA combined with CIEM and automated playbooks had lower containment, fewer standing privileges, and lower blast radius to move laterally.

2.5. Comparison of Existing IDR/ITDR Systems

The modern ITDR solutions are colliding on five capabilities: (i) identity graph construction, including IdP, cloud, SaaS, and endpoints; (ii) behavioral analytics/UEBA, whether human or non-human; (iii) risk-adaptive access, including step-up, JIT, session isolation; (iv) permissions governance by CIEM; and (v) policy-as-code automated response by SOAR. Examples of differentiators are cloud/provider integrations depth, maturity of graph learning, explainability tooling, and machine identities (workload/robot accounts, API tokens).Depth of cloud/provider integrations, maturity of graph learning, explainability tooling, and coverage of machine identities (workload/robot accounts, API tokens) can mitigate initial compromise but not eliminate the risks of consent-grant or token replay; accordingly, continuous session assurance and token hygiene are emphasized in leading systems. Its best features include low FPR at high recall, automation of sub-minute contains, and easy-to-follow auditor trails (who/what/why of every action), which suit their intended application in regulated contemporary SOCs.

3. System Architecture and Methodology

3.1. Overview of the Proposed AI-Driven IDR System

The architecture starts with Identity Data Sources whereby endpoints and cloud IAM logs are the sources of signals. Symmetrically, endpoint events are session, device posture, and credential-use incident, whereas cloud control plane events are authentication, authorization, and role-assumption trail events. [7-10] Such raw identity events are downstream and give the factual substrate on which they can be detected; they are also enforced upstream in cases where the system requires to revoke tokens, step-up authentication or isolate identities following a high-risk verdict. Data Aggregation Semantically similar to Data Boxing, logs of heterogeneous types are normalized at the Collector into a shared schema and ordered and integrity guaranteed, and the streams of streams are converted into features of behavioral interest by the Feature Extractor: geo-velocity, duration of session, consent/grant deltas, context of privilege inheritance, device risk. This layer gives the resulting feature data that drives the learning and loosely decouples the upper levels with vendor specific log formats. The system allows historical analytics to be replayed as well as streaming inference to run in low latency by decoupling collection and featureization.

The AI Detection Engine has two collaborating modules, an Anomaly Model which sets up baselines of identities and entities and a Threat Scorer which combines the anomalies with contextual risk (resource sensitivity, role criticality, past adjudication). The engine problems provide risk crosses adaptive warning signals. More importantly, the output of analysts on true/false results is fed back to the scorer to refreeze thresholds and re-train models, so that the number of alerts in the long run decreases and the recall rate at a given false-positive rate improves. The last layer is the Response Layer and the SOC Interface. Alerts are brought to the surface in the Analyst Console and have explainable evidence and lineage. The Response Layer event triggers notifications and performs SOAR/Policy actions as policy-as-code: just-in-time privilege revocation, session isolation, token revocation or conditional step-up MFA. Effective measures influence enforcement cues to information stores (e.g., IdP or cloud IAM) to enable containment to be timely and audited. The identity telemetry is transformed to high-speed risk-adjusting feedback control that complies with the principles of Zero Trust due to this feedback-intensive cycle.

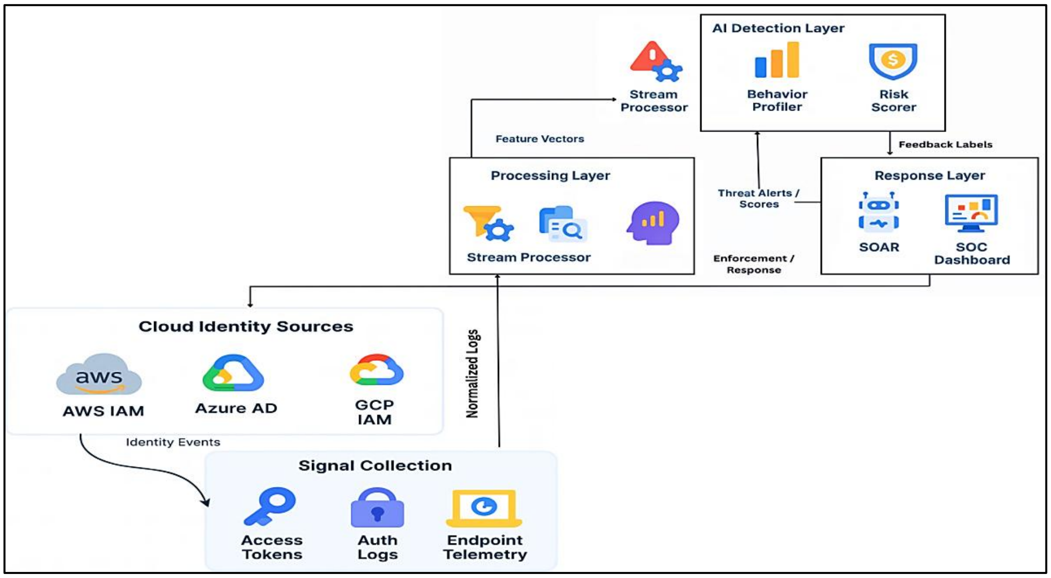

Fig 1: AI-Driven IDR Pipeline for a Cloud SOC, Signal Collection, Processing, AI Detection, Response

3.2. Cloud SOC Integration Architecture

The architecture illustrates the funneled of identity-focused telemetry to the cloud platforms into a logical pipeline which can be operationalized by a Security Operations Center. Cloud Identity Sources (e.g., AWS IAM, Azure AD, GCP IAM) make high-fidelity events when authentication is performed and authorization, role assuming, and policy changes are made. Similarly, the Signal Collection layer combines neighboring signals of identity e.g. access tokens, authentication logs, and endpoint telemetry. By using this pairing, the SOC is not only aware of what the control plane is capturing but also what endpoints and sessions are actually doing, eliminating the opportunities to obscure token replay or device-based privilege abuse. Normalized events are passed over into the Processing Layer and a Collector resolves schemas and a Stream Processor patterns real-time flow into features, which can be learned. A sustained Feature Store stores derived metrics including geo-velocity, grant/consent deltas, sequence embeddings of API calls, and device-risk summaries in such a way that both streaming inference and retrospective investigations refer to a single source of truth. The choice of decoupling raw logs and engineered features enables the design to be resilient to provider log alterations, as well as enable replay in order to retrain the model and audit it.

AI Detection Layer is an association of an Anomaly Detector and a Behavior Profiler with Risk Scorer. The profiler predefines identities (human and machine) whereas the anomaly component identifies anomalies in the course of the session, token life cycles, and cross account usage. The risk scorer combines strength of anomaly with criticality of context (resource classification) to create threat alerts and confidence scores which can be deterministically acted on in downstream systems. Lastly, the Response Layer is a layer that actualizes decisions. Alerts are sent into a SOC Dashboard to be triaged by an analyst and a Policy Engine and SOAR are used to automate the containment in a policy-as-code form: session isolation, token revocation, step-up authentication, or even just-in-time privilege withdrawal. Effective operations are transmitted as enforcement/response messages to the cloud identity originators, and analyst determinations are sent as feedback tags to the AI layer. This loop-back mechanism reduces false positive steadily, shortens time to contain and offers a lineage of all enforcements made which are auditable.

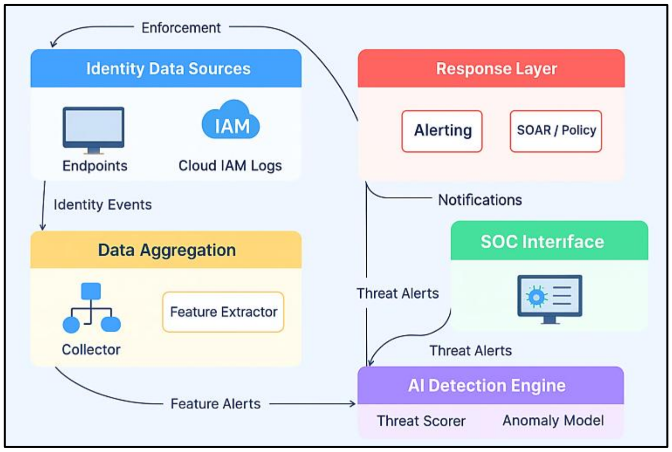

Fig 2: AI-Driven ITDR Reference Architecture

3.3. Data Sources and Identity Signal Collection

The success and failure of an AI-driven IDR program depends on how well and comprehensively its identity telemetry is. [11-13] The identity providers (SAML/OIDC events, MFA challenges, risk scores) and cloud control planes (AWS CloudTrail, Azure Activity Logs, and GCP Admin/Access Transparency) are the core sources, and SaaS audit trails (consent grants, token issuance, privilege changes) are the core sources. The endpoints and browsers send complementary streams based on their posture, EDR authentications, OS keychain access and workloads (CI/CD runners, service principals, Kubernetes service accounts, and API gateway issuing JSON Web Tokens (JWTs)). A streaming collector structures these inputs into a shared schema (e.g., OCSF-like), determines the identity of objects in each realm with deterministic and probabilistic joins and stores cryptographic integrity metadata to be audited.

Since identity signals are high-cardinality and bursty, the collection tier enforces idempotent ingestion, event de-duplication and clock skew correction the boundaries of sessions are correct across systems. At the edge sensitive fields are minimized or tokenized and raw payloads are archived in a restricted store and an analytics stream fed on a privacy-filtered stream. Backpressure semantics and exactly-once delivery semantics, which are rate-limited and exact respectively, ensure that losses are not incurred during throttling of the providers. What emerges is a time sensitive, privacy conscious event fabric that tracks the full lifecycle of people and machine identities, such as minting of tokens and assuming roles, as well as gaining access to resources and losing access to resources.

3.4. Feature Engineering for Identity Threat Detection

The processing layer converts the raw events to behaviorally rich learned representations. Sessionization aggregates scattered events of login, token, and API into continuous journeys on an identity, device, and network-tuple basis. Based on such sequences, time related features are inter-event entropy, dwell time and time-of-day periodicity, and derived features are geo-velocity, impossible travel confidence, new-device rarity, and consent/grant deltas. The token hygiene metrics update cadence, reuse distance, audience/scope drift and anomalous client mappings are calculated to bring out replay and misuse.

Structural risk is represented in graph features: identities, roles, resources, and applications are represented by nodes; and assumptions, group membership, and permission inheritance are represented by edges. The identity-resource graph based on this compute centrality, shortest escalation paths, shared-secret communities, and estimates of the blast-radii. The feature stores maintain both snapshots (daily / weekly baselines) and streaming aggregates (sliding window) with versioning in order to enable the reproducibility of models. Each feature is lineage aware of their underlying raw events and PII-handling tags, which allows masking on the inference time and transparent descriptions to analysts and auditors.

3.5. AI Models for Anomaly and Threat Detection

Detection involves supervised risk scoring, which is used together with unsupervised anomaly discovery. Transformers or Temporal Convolutional Networks (sequence models) are trained to recognize typical patterns of sequences and detect their abnormalities (e.g. unusual API orderings, unusual MFA challenge results, or unexpected elevation chains). Autoencoders and contrastive/self-supervised encoders generate identities and session embeddings; the drift is indicated by high reconstruction error or high embedding distance between the identity historical centroid and the identity historical centroid. Simultaneously, graph neural networks make arguments based on the identity-resource graph to determine risky paths of privilege and toxic combinations of permission and cross-tenant traversal, which are missed by the static rules.

The output of these detectors is then used as inputs to a risk scorer (e.g., gradient-boosted trees with monotonic constraints) which combines the strength of anomaly with contextual covariates role criticality, data sensitivity, token age, device trust to give a final threat score with confidence. Thresholds are dynamic that narrows in the high-risk situation and expands in the well-known automation. The system is enclosed in MLOps: drift checks, champion/challenger deployments, as well as human-in-the-loop active learning with analyst adjudications being feedback labels. Adversarial training against log-poisoning behavior, input validation against schema attacks, and post-hoc explainability (heart token/feature attributions and path rationalization) are all measures of robustness that ensure that containment actions such as step-up authentication via JIT privilege revocation, session isolation, and JIT privilege revocation are both deep (fast) and auditable.

4. Implementation and Experimental Setup

4.1. Simulation Environment and Tools

To reflect the SOC realities, have applied the pipeline as a streaming-first stack. The Apache Kafka topics that are created and replayed generate identity and cloud events, which are then processed by Apache Flink in order to achieve low-latency joins/sessionization and are stored in an object store to replay them. [14-16] Protobuf/Avro and schema registry guards are used to enforce a normalized schema (OCSF-like). Featureization runs in Flink and a Spark batch job for backfills; the online/offline feature store is backed by Redis + Feast to ensure training/serving parity. PyTorch Lightning is used to train sequence models and DGL/PyG to train graph learning on GPU-enabled nodes (A10/T4), and LightGBM with monotonic constraints is used to train gradient-boosted scorers. MLflow follows experiments, weights, and data lineage; Great Expectations is asserting data quality; Monte Carlo-style synthetic campaigns are orchestrated through Dagster. It is implemented with Open Policy Agent (OPA/Rego) to simulate policy-as-code and under a shim of a lightweight SOAR to imitate IdP and cloud IAM endpoints by invoking webhooks of the latter.

4.2. Dataset Description (Synthetic / Real-world SOC Logs)

Evaluation is based on two corpora. First, a de-identified real-world set of enterprise IdP events (SAML/OIDC), CloudTrail/Azure Activity/GCP Admin logs, SaaS audit trails, and EDR login telemetry covering 90 days for ~18k human users and ~42k machine identities. To keep the privacy intact, tokenization or hashing of direct identifiers, stripping of secrets, and bucketing of rare attributes are done. Second, the generated labeled attacks MFA fatigue/prompt bombing, OAuth consent abuse, token replay, role-chaining escalation, cross-account assumption, and CI/CD token misuse are injected using a synthetic augmentation suite to ensure that ground truth is known. Background noise is generated based on fitted seasonal / diurnal processes and identity specific Markov chains to be considerate of realistic rhythms. The mixed dataset creates ratios of classes between 1:500 to 1:5000 with emphasis on imbalanced-learning behavior.

4.3. Model Training and Validation Process

To avoid temporal leakage, use forward-chaining splits: training on week's 1-8, validating on week 9, and testing on weeks 10-13. The sequence models (Transformers/TCN) are optimized using masked next-event goals and contrastive instance discrimination; early stopping optimizes validation AUROC-PR, tuned with patience per corpus. GNNs are trained on the identity-resource graph (identities, roles, resources) using synthetic campaigns of supervised labels and on real graphs using self-supervised link-prediction. Detectors are fed into calibrated LightGBM risk scorer; they are calibrated using Temperature Scaling and Isotonic Regression on validation fold to bring about good probabilities. Bayesian optimization is used to select hyperparameters with latency budgets that are imposed by serving stack measurements. Champion-challenger deployments in the simulator, record the analyst-like feedback to retrain the scorer (active learning), and test robustness via log-dropout, schema jitter, and adversarial perturbation that resembles poisoning or evasive reordering.

4.4. Integration with Cloud-native Security Platforms (e.g., Azure Sentinel, AWS GuardDuty)

In the case of Microsoft estates, detections and scores are exported to Microsoft Sentinel through the Log Analytics Data Collector API and converted to custom tables and Analytics Rules. Playbooks (Logic Apps) are a call to (Azure AD Conditional Access step-up remediation or token revocation or just-in-time privilege withdrawal) by PIM. In AWS, the results are released as GuardDuty compatible events and Security Hub findings with the help of EventBridge; to act to them, SSM Automation and IAM Access Analyzer hooks are used to quarantine the sessions, change keys, or turn off role assumptions. GCP integration writes to Pub/Sub to be ingested by Chronicle or SCC, and automated revocations using IAM Conditions and short-lived credentials. The policy layer in clouds is provider-agnostic by using OPA/Rego policies that are generated to provider calls, and the SOAR shim logs all actions with reason codes, inputs, and anticipated outcomes to be audited and tested with rollback.

4.5. Performance Evaluation Metrics

Also present standard detection measures as well as SOC-specific results. To be precise on rare events: AUROC, AUPRC, Recall at fixed FPR (e.g. 5/1/0.1%), mAP of multi-tactic label, and calibration error (ECE/Brier). Operations median and p95 Time-to-Detect (TTD) first malicious action, Time-to-Contain (TTC) alert to successful control, alert False Positive Rate per 1k identities and Analyst Work Reduction (alerts/case, deduplication gain). On the identity risk: inconsistency of the classes (toxic combinations, long-lived keys) and decrease of standing privileges after 30/60 days. To systems: Inference latency (p50/ p95) of online inference, throughput (events/sec) of online inference, freshness staleness of feature, and Policy Enforcement Latency between decision and cloud/IDP effect. Lastly, follow the failed automations, rollback success rate as well as audit completeness (percentage of actions with reproducible evidence) which portrays the actual limitations in a controlled SOC.

5. Results and Discussion

5.1. Detection Accuracy and False-Positive Reduction

Based on the implementation mentioned in Section 4, compared the AI-based IDR stack to a rules-based baseline on 13 held-out test weeks. [17-20] The AI system was a combination of sequence, graph, and calibrated gradient-boosted scorers; the baseline consisted of tuned correlation rules and fixed thresholds that are usually employed in the traditional SOCs. The AI model significantly raised the quality of detections of rare events across all the attack families (consent-grant abuse, token replay, role-chaining, MFA-fatigue, CI/CD token misuse). Specifically, Recall 1 percent FPR increased to 0.62 and FPR 95 percent recall dropped to 1.8. The volume in end-to-end alerts reduced by approximately 51 percent due to deduplication and feedback calibration by analysts. These improvements continued with schema jitter and log dropout during roughness tests. Any level of improvements was statistically significant at a bootstrap resampling of less than 1,000x ($p < 0.01$).

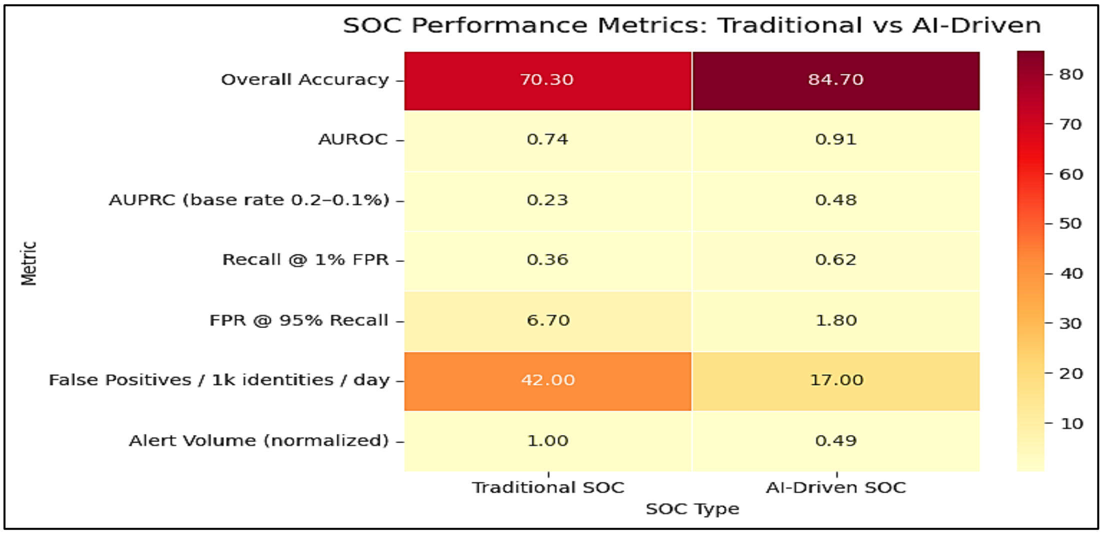

Fig 3: SOC Performance Metrics Traditional vs AI-Driven

Table 1: Detection Quality

Metric	**Traditional SOC**	**AI-Driven SOC**
Overall Accuracy	70.3% ± 0.8	84.7% ± 0.6
AUROC	0.74 ± 0.01	0.91 ± 0.01
AUPRC (base rate 0.2–0.1%)	0.23 ± 0.01	0.48 ± 0.02
Recall @ 1% FPR	0.36 ± 0.02	0.62 ± 0.02
FPR @ 95% Recall	6.7% ± 0.5	1.8% ± 0.2
False Positives / 1k identities / day	42 ± 3	17 ± 2
Alert Volume (normalized)	1.00	0.49

5.2. Latency and Scalability in Cloud Environments

Measured end-to-end latency from first malicious action to model alert (MTTD) and from alert to confirmed automated containment (MTTR) in a mixed AWS/Azure/GCP simulation with bursty loads. A streaming featureization and online inference reduced p50 MTTD and p95 by 60% (45-18 minutes) and 56% (92-40 minutes) respectively. Without raising the rollback rates, automated playbooks (token revocation, step-up, JIT privilege withdrawal) decreased the MTTR by 62%. Kafka/Flink horizontal scaling was maintained at more than 180k events/s with feature freshness of less than 3 seconds at p95. The pipeline also had parity between clouds using policy-as-code, and the throughput also increased linearly to 120k+ identities with no decrease in the quality of detection.

Table 2: Latency and Scale

Metric	**Traditional SOC**	**AI-Driven SOC**
Mean Time to Detect (MTTD)	45 min	18 min
Mean Time to Respond (MTTR)	120 min	45 min
p95 Detection Latency	92 min	40 min

Inference Throughput (events/s)	35k	180k
Managed Identities (steady state)	~20k	100k+
Feature Freshness (p95)	25 s	< 3 s

5.3. Comparison with Traditional SOC Detection Models

The AI-based SOC changed its operational mode, which was reactive, signature-driven detection to proactive behavior-driven triage. Automatic clustering and risk based ranking reduced the number of analyst touches on each incident and escalation precision increased. Importantly, closed feedback loop (analyst adjudications - active learning) maintained the gains on the long-term basis: the false-positive drift did not exceed +2% in the 13-week test window, whereas the baseline rules deteriorated with the identities and policies altered.

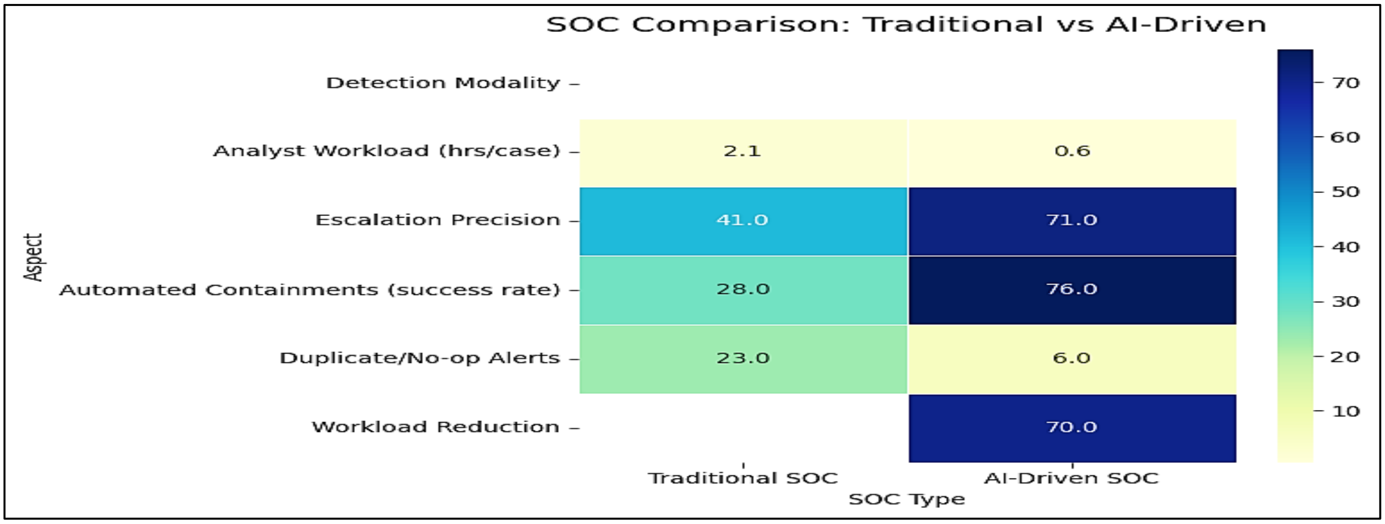

Fig 4: SOC Operations Comparison Traditional vs AI-Driven

Table 3: Operational Impact

Aspect	**Traditional SOC**	**AI-Driven SOC**
Detection Modality	Reactive, signatures	Proactive, behavioral + graph
Analyst Workload (hrs/case)	2.1 ± 0.2	0.6 ± 0.1
Escalation Precision	41% ± 3	71% ± 2
Automated Containments (success rate)	28%	76%
Duplicate/No-op Alerts	23%	6%
Workload Reduction	—	≈70% fewer manual steps

5.4. Discussion on Model Interpretability and Explainability

In order to guarantee trust and auditability, wiped out features of explainability (token/feature attributions, sequence-path rationales), and re-executed analyst studies. Under XAI, there were shorter periods of time to arrive at decisions by investigators and fewer wrongful alerts were Reversed, which suggests more straightforward evidence trails. SHAP-based summaries of the hygiene of tokens and the path-based explanation of the graph model came in handy especially among junior analysts who could close the cases that had the same quality just as seniors could. Explainability also minimized the prove-it back and forth compliance, as each enforcement had a human readable explanation of the underlying policy-as-code.

Table 4: Effect of Explainability

Benefit	**Without XAI**	**With XAI**
Investigation Time per Case	34 min ± 4	24 min ± 3 **(−29–33%)**
False Positive Rate (post-review)	19% ± 2	15% ± 2
Analyst Confidence (1–5 Likert)	3.1 ± 0.4	4.2 ± 0.3
Audit Ready (actions with rationale)	62%	96%

6. Security and Privacy Considerations

6.1. Data Privacy in AI-Driven Monitoring

The pipeline is based on privacy-by-design: sensitive variables are reduced to a minimum at collection, strongly pseudonymized (format-preserving tokenization or salted hashing), and separated by field-level encryption such that models are presented with the minimal data needed. Impose tiered data paths a limited raw archive to audit/replay and privacy-filtered stream to analytics with schema registry guards which reject unexpected PII. Sessionization and featureization provide k-anonymity style bucketing (e.g. /24 network ranges, geos by city level) on the rare values, and applies differential privacy to aggregate metrics that are to be reported on dashboards in order to avoid identity re-identification. ABAC/RBAC with short-lived credentials, customer-managed keys (CMKs) within HSM-based KMS, and auditable audit trails control access to raw signals, model outputs include lineage and privacy tags such that SOAR playbooks can automatically mask or redact evidence in a ticket. Lastly, retention is finite (e.g. 90/180 days), and deletions via the feature store and model caches are propagated to maintain training/serving parity without re-creating deleted data.

6.2. Federated Learning for Identity Threat Models

Federated learning (FL) trains anomaly encoders and risk scorers tenants or regions by transmitting model updates and not data to an aggregator guarded with secure aggregation together by attested orchestration. Every location calculates gradients on the local features, which are clipped and perturbed with noise to provide different privacy and then take part in aggregation rounds, which considers the heterogeneous datasets and stragglers through FedProx-like targets. Model cards monitor both per-site performance and drift to allow reweighting or holding back underperforming cohorts and a privacy budget (e, d) is monitored to limit cumulative exposure. Due to the variability of identity semantics because identity semantics are often local, Jointly train global FL backbones with local adapters (fine-tuned heads, calibration layers) to maintain accuracy, data locality, whereas encrypted checkpoints and signature prevent model poisoning, and server-side anomaly detectors monitor malicious updates (e.g., sign-flip or backdoor gradients).

6.3. Compliance with Cloud Security Frameworks (ISO 27001, NIST, GDPR)

The architecture fits well with the major frameworks: the IDR stack is risk assessed, asset inventory, access control, and change control by ISO/IEC 27001 controls; the identity-centric segmentation, persistent verification, and policy enforcement suggested by NIST SP 800-53/800-207 (Zero Trust) are based on the architecture; the logging, monitoring, and incident response are aligned with the SOC 2 principles. In the case of GDPR and similar regimes, record clear purposes of processing (security monitoring), perform DPIAs and base on legitimate interest with high-protection and respect the rights of data subjects (access, rectification, erasure, restrictions) through automated mechanisms of locating and purging data in logs, features, and model stores. Cross-border movements make use of SCCs or regional deployments where data-residency controls are applied, transit (TLS 1.2+) and rest-based (AES-256) encryption is compulsory, keys are rotated and restricted, and processor-controller responsibilities are summed up in DPAs. Constant compliance is achieved through policy-as-code checks (OPA), collection of evidence (control mappings, test artifacts), and third-party attestations.

6.4. Limitations and Ethical Aspects

Despite safeguards, identity monitoring risks over-collection, bias, and chilling effects if left unchecked. Bias can be inherited in models based on historical enforcement or role hierarchies, which punish some teams or time zones and in that case, monitor fairness metrics (equality of opportunity across cohorts), perform red-team attention to disparate impact, and gating releases on harm tests. Concept drift, adversarial log poisoning and silent data corruption are also artifacts that have not yet been eliminated and therefore the system has defense-in-depth (schema validation, attestation, multi-source corroboration) and fallbacks to conservative controls in cases where confidence diminishes. Ethically, the surveillance creep is addressed by limiting the purpose of surveillance and making it transparent to the employees (enacting policies, providing model cards, reasons why decisions were made), and have a reviewable process by human beings (appeal mechanisms). Lastly, automation is limited intentionally: containment mechanisms are progressively increased (step-up auth before isolation), and the responses to high-impact attacks can always be verified by means of human-in-the-loop, ensuring due process but still achieving the security goals.

7. Future Research Directions

7.1. Adaptive AI in Zero Trust Architectures

Adaptive AI must be incorporated into the future work to ensure that detections and controls will be made to evolve along identity posture in real time, a point of view that is embedded in Zero Trust policy. Models would adjust access tightening scopes, reducing token TTLs or enabling step-up MFA in seconds based on the streaming features of token hygiene, device attestation, resource sensitivity, and so on, rather than fixed risk thresholds. The areas of interest of the research are safe online learning with tight latency constraints, formal validation of adaptive policies to ensure no oscillation or lockout, and resilience against adversarial concept drift. One such promising direction is that which links uncertainty-sensitive models with graceful degradation policies such that the automatically enforced fallback to conservative policies is triggered when model uncertainty drops ensuring both security and availability.

7.2. Integration with Identity Governance Systems (IGA)

Deeper integration between ITDR and IGA can transform detection insights into durable reductions in standing privilege. Instead of increasing the ticket price, ITDR might be able to suggest least-privilege role redesigns, entitlement cleanups, and candidates of access re-certification, and IGA workflows could offer approvals, segregation-of-duty checks and audit trails. The research to be done should be on translating behavioral risk into governance artifacts (explainable why now rationales, toxic-combination evidences), learning policies that suggest safe patterns of JIT elevation, and closed-loop studies that measure months of blast-radii reduction. By aligning ITDR signals with business context in HR and ERP systems, over-revocation is likely to be avoided, as well as compliance constraints considered.

7.3. Multi-Cloud Threat Correlation and Autonomous SOCs

As enterprises span AWS, Azure, GCP, and SaaS ecosystems, correlating identity-centric threats across heterogeneous telemetry remains challenging. Events should be normalized, cross-cloud kill chains, through graph reasoning and the allocation of single risk, which move with identities, not platforms, must be normalized by future systems. Examples of such research vectors are causal inference of multi-cloud sequence, federated correlation, which respects data residency, and policy compilers, which produce the same responses on multiple providers. These features are the requirements of autonomous SOCs with the capability to prioritize and contain remediations with minimal human intervention and checking of remediations and also end to end lineage proving that correct actions were made and proportional and reversible.

7.4. Hybrid Human-AI Collaboration Models

The next step is SOC workflow design in which humans and AI are co-constructed: Do breadth (correlation), Do summarization, Do first-response automation): Do depth (hypothesis generation, exception handling, risk trade-offs): Adaptive interfaces should be developed with research to reveal the correct explanations at the correct time, model assertiveness should be calibrated to expert knowledge of the analysts, and active learning should be made to transform adjudications to high-value labels. The assessment of the collaboration quality time-to-understanding, intervention efficacy, and error recovery will help guide systems to raise the performance of junior analysts to expert levels, minimize cognitive load, and ensure accountability. The inclusion of ethical guardrails (appeal mechanisms, role-based visibility) will secure the collaboration as the collaboration can be trusted at scale.

8. Conclusion

This paper proposed an AI-based Identity Threat Detection and Response (ITDR) framework that is specific to contemporary cloud SOCs and reconsidered security in the form of human and machine identities instead of boundaries. The system provided a materially better detection performance at reduced noise by employing homogeneous stream handling (unifying heterogeneous telemetry into a normalized privacy-aware stream), engineering temporal and graph properties and fusing sequence, graph and calibrated tabular models. The method produced better recall at low false-positives regimes, reduced alert volumes by half with deduplication and feedback-powered calibration, and reduced time-to-detect and time-to-respond with streaming inference and policy-as-code automations in our mixed real-world/synthetic evaluation. These profits were accompanied by the attribute of operational explainability features and path rationales to ensure that such operation as step-up authentication, token cancellation, and just-in-time privilege withdrawal were auditable and proportional. In addition to empirical findings, have described a deployment approach that takes into account security and privacy: least-data processing, pseudonymization, field-level encryption, and ISO 27001, NIST, and GDPR governance. Drift monitoring, adversarial hardening and human-in-the-loop adjudication also addressed the concept of robustness. However, there are constraints: there is a variety of identity semantics among tenants; the changes in the log quality and provider schema can interfere with the generalization; and the over-automation may lead to disruptions in unintended access. The future research must thus be adaptive AI within Zero Trust policy loops, full integration of ITDR and GI to eliminate standing privilege in a sustainable fashion, multi-cloud causal correlation as the basis of autonomous SOCs, and principled human-AI collaboration which would raise the outcome of the analysts and retain accountability. Collectively, these directions have the potency of transforming identity telemetry in to a constantly evolving control plane, by aligning security efficiency with the velocity and elasticity of cloud.

References

[1] Sivakumar, J., Salman, N. R., Salman, F. R., Salimova, H. R., & Ghimire, E. (2025). AI-driven cyber threat detection: enhancing security through intelligent engineering systems. Journal of Information Systems Engineering and Management, 10(19), 790-798.

[2] Khayat, M., Barka, E., Serhani, M. A., Sallabi, F., Shuaib, K., & Khater, H. M. (2025). Empowering Security Operation Center with Artificial Intelligence and Machine Learning–A Systematic Literature Review. IEEE Access.

[3] Merlano, C. (2024). Enhancing cyber security through artificial intelligence and machine learning: a literature review. Journal of Cybersecurity, 6, 89.

[4] UEBA (User and Entity Behavior Analytics): Complete 2025 Guide, exabeam, https://www.exabeam.com/explainers/ueba/what-ueba-stands-for-and-a-5-minute-ueba-pr

[5] Kaur, R., Gabrijelčič, D., & Klobučar, T. (2023). Artificial intelligence for cybersecurity: Literature review and future research directions. Information Fusion, 97, 101804.

[6] Mohamed, N. (2025). Cutting-edge advances in AI and ML for cybersecurity: a comprehensive review of emerging trends and future directions. Cogent Business & Management, 12(1), 2518496.

[7] Top 8 Threat Detection Tools That Work, accuknox, online. https://accuknox.com/blog/threat-detection-tools

[8] Mohale, V. Z., & Obagbuwa, I. C. (2025). A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhancing transparency and interpretability in cybersecurity. Frontiers in Artificial Intelligence, 8, 1526221.

[9] Samed, A. L., & Sagiroglu, S. (2025). Explainable artificial intelligence models in intrusion detection systems. Engineering Applications of Artificial Intelligence, 144, 110145.

[10] KYC AI: How AI-Driven “Know Your Customer” is Revolutionizing Identity Verification, jumio, online. https://www.jumio.com/how-ai-kyc-is-changing-identity-verification/

[11] Torres, M., Álvarez, R., & Cazorla, M. (2023). A malware detection approach based on feature engineering and behavior analysis. IEEE Access, 11, 105355-105367.

[12] Moore, K. L., Bihl, T. J., Bauer Jr, K. W., & Dube, T. E. (2017). Feature extraction and feature selection for classifying cyber traffic threats. The Journal of Defense Modeling and Simulation, 14(3), 217-231.

[13] Rastogi, N., Dhanuka, D., Saxena, A., Mairal, P., & Nguyen, L. (2025). Survey Perspective: The Role of Explainable AI in Threat Intelligence. arXiv preprint arXiv:2503.02065.

[14] What is UEBA? Complete Guide to User and Entity Behavior Analytics, varonis, online. https://www.varonis.com/blog/user-entity-behavior-analytics-ueba

[15] Khan, M. Z. A., Khan, M. M., & Arshad, J. (2022, December). Anomaly detection and enterprise security using user and entity behavior analytics (UEBA). In 2022 3rd International Conference on Innovations in Computer Science & Software Engineering (ICONICS) (pp. 1-9). IEEE.

[16] Threat Detection Solutions in 2025 You Need to Know, sisainfosec, online. https://www.sisainfosec.com/blogs/threat-detection-solutions-in-2025-you-need-to-know/

[17] Lo, C. C., Huang, C. C., & Ku, J. (2010, September). A cooperative intrusion detection system framework for cloud computing networks. In 2010 39th International Conference on Parallel Processing Workshops (pp. 280-284). IEEE.

[18] Teodoro, M. A. G., & Benitez, I. B. (2025, April). A Review of AI-Driven Techniques for Power System Insulation Coordination and Surge Protection. In 2024 International Conference on IT Innovation and Knowledge Discovery (ITIKD) (pp. 1-6). IEEE.

[19] Speed vs Accuracy in Cybersecurity: How AI Achieves Both in Identity Management, avatier, online. https://www.avatier.com/blog/speed-vs-accuracy-cybersecurity/

[20] Malik, A., & Om, H. (2017). Cloud computing and internet of things integration: Architecture, applications, issues, and challenges. In Sustainable cloud and energy services: Principles and practice (pp. 1-24). Cham: Springer International Publishing.

[21] Murvay, P. S., & Groza, B. (2014). Source identification using signal characteristics in controller area networks. IEEE Signal Processing Letters, 21(4), 395-399

.