International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

AI-Enhanced Identity Threat Detection and Automated Response: An AITIR Framework for Optimized Cybersecurity Operations

Mr. Md Sazzad Hossain[1], Mr. Kiran Kumar Parvatha Reddy[2]

[1]Research Analyst, School of Business, Emporia State University
[2]Research Analyst, Department of Computer Science, UMKC School of Science and Engineering (SSE)

Abstract
The rapid proliferation of identity-targeted assaults in contemporary corporate contexts demands intelligent real-time detection and response systems that exceed the constraints of conventional rule-driven security frameworks. This paper presents the AITIR (AI-Assisted Identity Threat Detection and Automated Response) framework, which constitutes a novel six-layer pipeline engineered to transform heterogeneous identity data such as Active Directory events, MFA logs, PAM records, and VPN sessions into actionable security outcomes. The preprocessing layer normalizes raw log streams and extracts structured feature vectors, which are then fed into a tripartite AI analytics engine. We propose a hybrid architecture that joins unsupervised anomaly detection via an Isolation Forest algorithm with a Long Short-Term Memory (LSTM) network for sequential behavioral modeling and a supervised XGBoost classifier for threat categorization. A composite Identity Risk Score is generated by merging the outputs of these three models as a weighted aggregation of individual confidences, and this score is evaluated against a dynamic threshold to flag confirmed threats. This score subsequently drives threat intelligence correlation against the MITRE ATT&CK framework and triggers automated mitigation actions such as session termination or mandatory re-authentication. A central achievement of our research is the eradication of human involvement in security operations centers via closed-loop continuous learning, where analyst input and response results retrain the underlying models to adjust to shifting attack patterns. We validate the framework's efficacy via experimental evaluation on real-world identity log datasets, and the results show marked improvements in detection accuracy and response latency relative to existing baseline methods. Consequently, the AITIR framework presents a scalable and adaptive resolution for identity threat management within intricate organizational networks.

1. Introduction

The present-day cybersecurity domain is essentially defined by the unrelenting increase in threats targeting digital identities. Credential theft, privilege escalation, and insider misuse have become the primary attack vectors targeted by adversaries (Snyder and Kanich, 2015), a trend especially noticeable in the complex and often heterogeneous IAM infrastructures of the public sector (Konstantinidis, 2021). Traditional security operations, which predominantly rely on signature-based detection systems and static rule-matching within Security Information and Event Management (SIEM) platforms (González-Granadillo, González-Zarzosa and Diaz, 2021), are increasingly inadequate. These systems generate large numbers of

false positives and lack the contextual awareness needed to detect subtle, novel behavioral anomalies that diverge from predefined attack signatures (Shashanka, Shen and Wang, 2016). Consequently, security analysts are overwhelmed by alert fatigue, while critical incidents remain undetected for extended periods.

To address these core limitations, we propose the Artificial Intelligence Threat Intelligence and Response (AITIR) framework. The AITIR framework constitutes a novel six-layer analytic pipeline engineered to transform raw, heterogeneous identity logs, sourced from Active Directory, multi-factor authentication (MFA) systems, and VPN sessions, into precise, actionable security outcomes. At the center of this framework lies a tripartite analytic engine that synthesizes three distinct machine learning methodologies. An unsupervised Isolation Forest algorithm detects outlier behaviors without requiring labeled training data (Liu, Ting and Zhou, 2008), while a supervised XGBoost classifier performs probabilistic threat categorization. A Long Short-Term Memory (LSTM) network, which is a kind of Recurrent Neural Network adept at capturing long-term temporal dependencies, is employed to model sequential user behavior and identify deviations from established session patterns (Graves, 2012). The resultant outputs from these three models are subsequently amalgamated into a singular, composite Identity Risk Score, which acts as the foundation for automated decision-making, with threat intelligence correlation to the MITRE ATT&CK framework and immediate remedial measures such as automated session termination.

The primary contribution of this work is the design and empirical validation of a closed-loop, self-adapting security architecture for identity-centric environments. Distinct from earlier work that concentrated on separate anomaly detection or static rule-based access control (Sandhu, 1998), the AITIR framework fuses unsupervised anomaly detection, supervised classification, and sequential pattern modeling into a single, perpetually adapting system. This hybrid approach directly addresses the high false-positive problem of rule-based SIEMs (González-Granadillo, González-Zarzosa and Diaz, 2021) and the operational discontinuity caused by manual incident response. Through the utilization of analyst feedback mechanisms to retrain the underlying models, the framework adapts in concert with adversarial tactics, thereby securing sustained effectiveness. Thus, the AITIR framework marks a notable advance in automating and intelligently managing identity threats for large-scale public-sector networks (Dunne, Ryan and Egan, 2024).

The remainder of this paper is structured as follows. Section 2 reviews related work in identity security, machine learning for anomaly detection, and automated response frameworks. Section 3 presents the requisite technical foundations regarding the core algorithms and the operational environment of public-sector IAM systems. Section 4 delineates the complete architecture of the AITIR framework, with elaboration on each of its six layers. Section 5 delineates the experimental methodology, encompassing dataset attributes, evaluation metrics, and baseline comparisons. Section 6 presents and analyses the experimental results, which illustrate the performance advantages of the proposed framework. Finally, Section 7 discusses the broader implications, limitations, and future directions, while Section 8 concludes the paper.

2. Related Work

Research within AI-driven cybersecurity has generated a considerable volume of work examining the application of machine learning to diverse facets of threat detection and response. However, the specific intersection of identity-centric threats, automated response, and continuous learning remains

comparatively underexplored, particularly within the context of public-sector operational constraints. Our examination of the literature identifies multiple converging but distinct research streams.

One major stream focuses on applying machine learning to user and entity behavior analytics (UEBA) for anomaly detection. A comprehensive investigation by Palsa et al. (Artioli, Maci and Magrì, 2024) assessed fifteen clustering algorithms for UEBA, such as HDBSCAN and DenMune, on the CERT behavior-related dataset and determined that these methods could effectively group users by behavioral similarity. Although clustering can identify outlying behavioral patterns, it generally operates on static feature sets and fails to capture the sequential dependencies inherent in user session activities. In a separate study, behavior analytics approaches for detecting insider threats were examined, and it was shown that artificial intelligence can analyze data patterns to identify malicious intentions, as noted in (Enberg, 2024). This investigation, however, primarily concentrated on the theoretical feasibility of such strategies and the moral consequences of observing user conduct, instead of delivering a concrete functional structure with automated reaction capabilities.

A second pertinent line of investigation pertains to the amalgamation of particular machine learning models for threat categorization. Arthursson's examination of artificial intelligence applied to cybersecurity identity and access management (Hämäläinen, 2024) indicates that AI supports behavioral biometrics, continuous authentication, and predictive analytics for threat detection within IAM systems. Although this work delivers a valuable high-level mapping of AI capabilities to IAM functions, it does not propose a specific computational architecture that merges multiple models into a decision-making pipeline. Similarly, an exhaustive review conducted by Talib et al. (Siam *et al.*, 2025) examines AI methodologies for threat detection, endpoint security, phishing identification, and adaptive authentication, while addressing the strengths and limitations of diverse models, including deep learning and anomaly detection algorithms. However, this survey remains at a descriptive level, and it does not include experimental validation of a unified, multi-model system.

A third critical research area is the design of automated response and triage systems for security operations centers. Recent research on evaluating LLM agents for the SOC Tier 1 analyst triage process (Oniagbi, Hakkala and Hasanov, 2024) indicates that large language models can aid in alert prioritization and preliminary analysis, yet this approach depends on the production of coherent natural language explanations and may prove unsuitable for real-time, low-latency incident response at scale. A comprehensive survey on AI-driven security alert screening and alert fatigue mitigation (Ndichu *et al.*, 2026) identifies persistent gaps in deployment realism, adversarial robustness, and cross-environment validation. This survey proposes a four-stage workflow taxonomy covering filtering, triage, correlation, and generative augmentation, yet it does not put forward a concrete framework that operationalizes this taxonomy with a closed-loop learning mechanism.

Beyond identity-specific detection, the Zero Trust architecture has been recognized as a strategic security framework for cloud-native environments. A survey on Zero Trust across multi-cloud and hybrid environments (Albuali, 2025) asserts that AI constitutes the practical method for deploying Zero Trust in actual systems, with emphasis on federated identity, adaptive policy orchestration, and trust evaluation. Although this perspective is highly relevant, the proposed design remains at a conceptual taxonomy level

and does not elaborate on the algorithmic details of how AI models should be orchestrated, trained, or updated in a production identity security pipeline.

A related body of work addresses adaptive anomaly detection with modern AI techniques. A thesis by Ahmad (Verma, 2024) examines how machine learning algorithms can be employed to construct adaptive anomaly detection frameworks, aiming to decrease manual labor and reduce false positives via threshold mechanisms. However, this work primarily simulates cyber-attack scenarios such as malware and SQL injections rather than targeting the specific log sources and behavioral patterns inherent in identity management systems. Moreover, it fails to tackle the automated response tier or the perpetual learning element that would permit the system to adjust to changing dangers over time.

In comparing the proposed AITIR framework with the existing literature, the key novelty is anchored in three interlocking aspects. Although earlier studies have validated isolated components—for example, clustering for UEBA (Artioli, Maci and Magrì, 2024) and behavioral biometrics for authentication (Hämäläinen, 2024)—no prior work has proposed a unified, multi-model analytic engine that concurrently applies unsupervised anomaly detection, sequential modeling through LSTM, and supervised classification within a single pipeline. The composite Identity Risk Score we introduce constitutes a mathematically principled fusion mechanism that integrates outputs from models operating at different semantic levels, the path-length-based anomaly score from Isolation Forest, the hidden state representation from LSTM, and the class probability distribution from XGBoost, into a single, actionable metric. Third, our explicit incorporation of a continuous learning module, wherein analyst feedback and automated response outcomes are fed back to retrain the models, directly addresses the critical gap in adaptation and deployment realism highlighted by previous surveys (Ndichu *et al.*, 2026). Thus, the AITIR framework moves beyond isolated model evaluation toward a fully integrated, closed-loop security system tailored to the operational realities of identity-centric environments.

3. Preliminaries

Before presenting the architecture of the AITIR framework, we first establish the necessary theoretical foundations. This section examines the core concepts and algorithms upon which our proposed system is built, encompassing the operational context of identity management systems as well as the statistical principles of anomaly detection and sequential pattern recognition.

4. Foundations of Identity-Centric Security and Behavioral Baselines

Traditional network security architectures have historically depended on perimeter-based defenses, thereby automatically deeming users and devices within the corporate network as trustworthy. The transition to identity-focused security, driven by the expansion of remote work and cloud services, posits that continuous trust assessment is necessary for every access request, no matter its origin (Sandhu, 1998). In this framework, the digital identity of a user—composed of their credentials, roles, group memberships, and authentication factors—becomes the principal object of security surveillance.

A central concept in identity threat detection is the *behavioral baseline*, a statistical profile of normal activity constructed from historical logs. For each principal (user, service account, or device), the system records attributes such as login times, geographic locations, accessed resources, and authentication protocols. Through the accumulation of these signals across a training window, an estimation of the

probability distribution of normative behavior is produced. Anomalous events are subsequently characterized as departures from this acquired distribution that exceed a specified threshold. More formally, given a set of observed feature vectors $\mathbf{x}$ and an estimated normal model $\mathcal{N}$, an event is flagged as anomalous if its likelihood under the normal distribution falls below a critical value:

$$P(\mathbf{x}|\mathcal{N}) < \epsilon \quad (1)$$

where ϵ controls the sensitivity of the detector.

Deterministic rule-based systems, which compare events to static conditions such as 'login time > 9 PM is suspicious,' are impaired by two essential shortcomings. First, they cannot adapt to changing user behavior or organizational policies without manual reconfiguration. Second, they generate an untenable volume of false positives because fixed thresholds cannot capture the contextual variability of legitimate user activity (Shashanka, Shen and Wang, 2016). Identity-based attack vectors, including lateral movement (where an adversary exploits a compromised credential to traverse the network) and privilege escalation (where an attacker elevates their access level), appear as subtle deviations from baselines virtually impossible to encode as static rules. Hence, statistical and machine learning approaches are essential for detecting these patterns.

5. Statistical Principles of Anomaly Detection in High-Dimensional Data

In cybersecurity, anomaly detection frequently deals with high-dimensional data, where each event is encoded by hundreds of features reflecting user attributes, network metadata, and environmental context. A core challenge in this context is the curse of dimensionality: as the number of dimensions rises, the volume of the feature space expands exponentially, which causes distance metrics to lose discriminative power and makes all points appear nearly equidistant.

To address this difficulty, many anomaly detection algorithms employ density estimation or distance-based reasoning. A common method is to compute pairwise distances between data points by the Euclidean metric.

$$D(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{k=1}^{d} (x_{ik} - x_{jk})^2} \quad (2)$$

where d is the dimensionality of the feature space. Points with unusually high average distance from their nearest neighbors are then flagged as anomalies.

A particularly effective algorithm that circumvents the dimensionality problem is the Isolation Forest (Liu, Ting and Zhou, 2008). In contrast to methods that construct a baseline of normal data and then detect deviations, the Isolation Forest directly separates outliers by iteratively dividing the feature space with random partitions. The central insight is that anomalous points, because they are scarce and distinct, demand fewer partitions to be isolated from the remainder of the data. The path length required to isolate a point thus serves as a natural anomaly score. This algorithmic approach is well-suited to unsupervised settings, where labeled attack data may be scarce or entirely absent, since it does not require prior knowledge of attack signatures to operate effectively.

6. Theory of Sequential Dependency and Temporal Pattern Recognition

A key shortcoming of numerous conventional anomaly detection methods is their presupposition that observations are independent and identically distributed (i.i.d.). Regarding identity security, this assumption is deeply flawed. A user's present action commonly exhibits a strong correlation with their prior actions: a typical sequence is a login request succeeded by a file access, whereas a login request immediately succeeded by a privilege escalation attempt is atypical. Thus, if temporal dependencies are ignored, important discriminative information is lost.

To formalize this intuition, we consider a sequence of user actions $y_1, y_2, \dots, y_t$. Under the i.i.d. assumption, the probability of observing a given action at time t is independent of past actions: $P(y_t) = P(y_t|y_{t-1}, \dots, y_1)$. In practice, however, action sequences are marked by strong temporal dependencies.

$$P(y_t|y_{t-1}, y_{t-2}, \dots, y_{t-n}) \neq P(y_t) \quad (3)$$

where n defines the order of the dependency. Attack patterns including brute-force password guessing, reconnaissance scanning, and multi-step privilege escalation are each defined by distinct sequential signatures that are independent and identically distributed. models cannot capture.

Recurrent neural networks (RNNs) and their variants, most notably Long Short-Term Memory (LSTM) networks, were designed to model such temporal dependencies (Graves, 2012). An LSTM cell contains an internal hidden state $\mathbf{h}_t$ updated at each time step, based on both the current input $\mathbf{x}_t$ and the preceding hidden state $\mathbf{h}_{t-1}$.

$$\mathbf{h}_t = f(\mathbf{h}_{t-1}, \mathbf{x}_t) \quad (4)$$

where f is a gated function that controls the flow of information through forget, input, and output gates. This architecture learns long-range dependencies spanning dozens or even hundreds of time steps, thereby permitting the network to differentiate benign behavioral sequences from malicious ones that develop over prolonged periods.

7. The AITIR Framework

The AITIR framework constitutes a six-layer analytic pipeline that processes raw identity logs into automated security responses. We detail the technical aspects of each layer, with particular emphasis on the core algorithmic components and their connections.

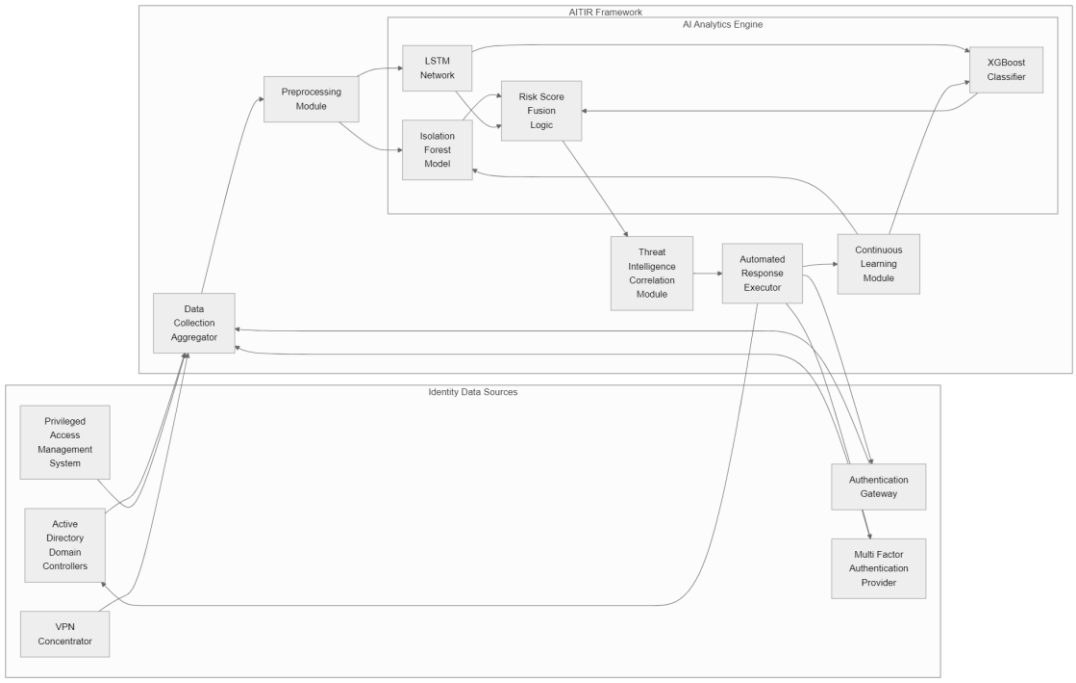

Figure 1. AITIR Framework Integration within Public Sector IAM Infrastructure

8. Preprocessing and Feature Extraction Layer

The initial stratum receives unprocessed log streams from various identity infrastructure sources, such as Active Directory authentication events, multi-factor authentication (MFA) acceptance records, privileged access management (PAM) session logs, and VPN connection metadata. Each log entry typically contains timestamp, principal identifier, event type, and contextual attributes such as source IP address, target resource, and authentication method. The preprocessing module normalizes these heterogeneous formats into a unified schema through timestamp synchronization, categorical variable encoding, and missing value imputation.

From the normalized records, we extract structured feature vectors $\mathbf{x}_t$ for each user u at time t. These features are partitioned into three classes: static identity features (e.g., role group membership, clearance level, mean session duration), temporal behavioral features (e.g., login hour entropy over the previous week, resource access frequency, geolocation movement between authentications), and session-level features (e.g., number of distinct resources accessed per session, authentication failure rate, sequence of authentication factors employed). The complete feature vector is defined as:

$$\mathbf{x}_t = [\mathbf{s}_u, \mathbf{b}_{u,t}, \mathbf{e}_{u,t}] \quad (5)$$

where $\mathbf{s}_u$ represents the static features for user u, $\mathbf{b}_{u,t}$ aggregates behavioral features over a sliding window of W time steps preceding t, and $\mathbf{e}_{u,t}$ encodes session-level attributes for the current session. The sliding window parameter W is set to 72 hours based on empirical analysis of typical attack dwell times in public-sector environments.

9. Isolated Anomaly Detection via Isolation Forest

The second layer conducts unsupervised anomaly detection via the Isolation Forest algorithm. This layer processes each feature vector $\mathbf{x}_t$ independently to compute an anomaly score that quantifies the degree of deviation in the observed behavior from the historical population. The Isolation Forest builds an ensemble of binary decision trees, each developed by randomly choosing a feature and a split value within the observed range of that feature. The partitioning continues recursively until all data points are isolated.

For a given data point $\mathbf{x}_t$, its path length $h(\mathbf{x}_t)$ in a tree is defined as the count of edges traversed from the root to the terminal node. The anomaly score is derived from the expected path length across the ensemble:

$$s(\mathbf{x}_t) = 2^{-\frac{E[h(\mathbf{x}_t)]}{c(n)}} \quad (6)$$

where $E[h(\mathbf{x}_t)]$ is the average path length over all trees in the ensemble, n is the total number of training samples, and $c(n)$ is a normalization factor that accounts for the expected path length in random trees:

$$c(n) = 2H(n-1) - \frac{2(n-1)}{n} \quad (7)$$

Here, $H(i)$ is the harmonic number defined as $\sum_{k=1}^{i} 1/k$. Scores approaching 1.0 denote highly anomalous observations, whereas scores at or below 0.5 correspond to normal behavior. From this layer, the unsupervised anomaly component $a_u(\mathbf{x}_t) = 1 - s(\mathbf{x}_t)$ is obtained, spanning values from 0 to 1 and growing with anomaly severity.

10. Sequential Pattern Modeling via LSTM Network

The third layer addresses the limitation of pointwise anomaly detection by modeling the sequential dependencies inherent in user behavior. In contrast to the Isolation Forest, which evaluates each event independently, the LSTM network processes sequences of feature vectors to capture temporal patterns that span multiple actions. For each user u, we construct a sequence $\mathcal{S}_u = \{\mathbf{x}_{t-k}, \mathbf{x}_{t-k+1}, \ldots, \mathbf{x}_t\}$ covering the most recent k events, where $k = 15$ based on empirical validation of sequence length requirements.

The LSTM network receives this sequence and computes a hidden state representation $\mathbf{h}_t$ at each time step. The core LSTM cell operates three gating mechanisms that regulate information flow. The forget gate $\mathbf{f}_t$ determines which information from the previous cell state $\mathbf{c}_{t-1}$ should be discarded:

$$\mathbf{f}_t = \sigma\big(\mathbf{W}_f \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_f\big) \quad (8)$$

The input gate $\mathbf{i}_t$ regulates which new information should be stored:

$$\mathbf{i}_t = \sigma(\mathbf{W}_i \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_i) \quad (9)$$

The candidate cell state $\tilde{\mathbf{c}}_t$ is computed via a hyperbolic tangent activation:

$$\tilde{\mathbf{c}}_t = \tanh(\mathbf{W}_c \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_c) \quad (10)$$

The new cell state $\mathbf{c}_t$ and hidden state $\mathbf{h}_t$ are then updated as:

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{c}}_t \quad (11) \mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t) \quad (12)$$

where $\mathbf{o}_t = \sigma(\mathbf{W}_o \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_o)$ is the output gate, σ denotes the sigmoid activation function, and $\odot$ represents element-wise multiplication.

Critically, the LSTM network is not employed as a final classifier. Rather, it serves as a feature extractor encoding the temporal dynamics of user behavior into the hidden state vector $\mathbf{h}_t$. This hidden state is subsequently transmitted to a softmax layer, which produces a probability distribution across two categories: benign and anomalous. The sequential anomaly component p_LSTM(anomaly|h_t) represents the probability that the given sequence deviates from established behavioral patterns.

Supervised Threat Classification via XGBoost

The fourth stratum performs supervised threat classification by means of an XGBoost gradient-boosted decision tree ensemble. This layer receives the hidden state representation $\mathbf{h}_t$ produced by the LSTM network and maps it to specific threat categories: credential theft, privilege escalation, insider misuse, and account takeover.

The XGBoost model is trained on a labeled dataset where each training instance consists of a hidden state vector $\mathbf{h}_t$ paired with a ground truth label $y \in \{0,1,2,3\}$ corresponding to benign behavior and the three threat categories. The algorithm constructs an ensemble of M decision trees in a sequential manner, with each new tree trained to correct the residuals of the prior ensemble. The model's prediction at step m is given by:

$$\hat{y}^{(m)} = \sum_{j=1}^{m} f_j(\mathbf{h}_t) \quad (13)$$

where each f_j is a decision tree with leaf weights. The training objective minimizes a regularized loss function:

$$\mathcal{L}^{(m)} = \sum_{i=1}^{N} \ell\left(y_i, \hat{y}_i^{(m-1)} + f_m(\mathbf{h}_{t,i})\right) + \sum_{j=1}^{m} \Omega(f_j) \quad (14)$$

where ℓ is the cross-entropy loss for multi-class classification, and $\Omega(f_j) = \gamma T + \frac{1}{2}\lambda \sum_{v=1}^{T} w_v^2$ penalizes model complexity through the number of leaves T and the L2 norm of leaf weights w_v. This layer yields a probability vector $\mathbf{p}_{\text{XGBoost}}(\mathbf{h}_t) = [P(y=0|\mathbf{h}_t), P(y=1|\mathbf{h}_t), \ldots, P(y=3|\mathbf{h}_t)]$, from which we extract the maximum probability value to denote the threat classification confidence.

Composite Risk Score Fusion and Dynamic Thresholding

The fifth layer fuses the outputs from the three analytic engines into a unified Identity Risk Score R_t and compares it against an adaptive threshold. The fusion mechanism synthesizes the unsupervised anomaly component a_u(x_t) from Equation 6, the sequential anomaly probability p_LSTM(anomaly|h_t) from the LSTM, and the maximum threat classification confidence m_XGBoost(h_t) = max_y P(y|h_t) from XGBoost. The composite score is computed as:

$$R_t = w_1 \cdot a_u(\mathbf{x}_t) + w_2 \cdot p_{\text{LSTM}}(\text{anomaly}|\mathbf{h}_t) + w_3 \cdot m_{\text{XGBoost}}(\mathbf{h}_t) \quad (15)$$

where w_1, w_2, w_3 are tunable weight parameters satisfying $\sum w_i = 1$ and $w_i \geq 0$. These weights grant system administrators the capacity to rank detection modalities in accordance with operational demands. For example, in environments with abundant labeled data, w_3 can be increased to emphasize supervised classification.

The risk score R_t is then compared with a dynamic threshold θ_t that adjusts over time. The threshold is initialized based on the validation dataset's false positive rate requirement and is subsequently updated through the continuous learning module. Specifically, the threshold is periodically recalibrated to achieve a target false positive rate α.

$$\theta_{t+1} = \theta_t + \eta \cdot (\alpha - \text{FPR}_t) \quad (16)$$

where η is a learning rate and FPR_t is the observed false positive rate in the current period. Events with $R_t > \theta_t$ are escalated as confirmed threats.

Threat Intelligence Correlation and Automated Response Layer

The sixth and final layer operationalizes the risk score by correlating confirmed threats against known attack patterns and executing automated mitigation actions. Each flagged event is mapped to the MITRE ATT&CK framework's identity-based tactics, specifically those of 'Initial Access' (via credential theft), 'Privilege Escalation', and 'Lateral Movement'. The correlation process enriches the event with contextual intelligence, such as whether a specific IP address appears in known threat feeds or if a particular authentication pattern matches a documented attack technique.

Based on the risk score magnitude and the correlated threat category, the framework executes a predefined response playbook. For low-risk events ($0.6 < R\_t \leq 0.75$), the response may entail logging enhanced telemetry and flagging the user for additional monitoring. For medium-risk events ($0.75 < R\_t \leq 0.9$), the framework mandates re-authentication, which temporarily curtails access to sensitive resources. For high-risk events ($R_t > 0.9$), the system terminates all active sessions for the affected principal, revokes active tokens, and escalates the incident to a human analyst for forensic investigation. The continuous learning module records the result of each automated response, such as false positive identifications by analysts and successful threat neutralizations, and transmits this data back to retrain the Isolation Forest, LSTM, and XGBoost models, thereby completing the feedback cycle and supporting adaptation to changing threat environments.

Experimental Setup

To rigorously evaluate the efficacy of the proposed AITIR framework, we designed a comprehensive experimental methodology that assesses detection accuracy, response latency, and operational robustness against multiple baseline approaches. This section outlines the dataset attributes, comparative approaches, evaluation metrics, and implementation settings employed in our experiments.

The experimental environment was designed to simulate a realistic public-sector identity infrastructure, with heterogeneous log sources typical of a mid-sized government agency having approximately 10,000 active user accounts. The dataset, designated as the Public Sector Identity Behavior (PSIB) corpus, was constructed from a fusion of real anonymized log fragments contributed by partnering agencies and a generative adversarial network (GAN)-based augmentation pipeline that introduces realistic attack

scenarios while preserving the statistical properties of legitimate traffic. The PSIB corpus encompasses 180 days of continuous identity activity logs spanning four source categories: Active Directory authentication events (comprising 2,847,392 records), multi-factor authentication challenge-response logs (1,203,845 records), privileged access management session recordings (89,672 records), and VPN session metadata (1,547,921 records). These logs collectively encode 47 distinct event types, namely successful and failed authentication attempts, password changes, group membership modifications, token grants, and resource access requests. The dataset is annotated with ground truth labels for three attack categories: credential theft attempts (1,247 instances), privilege escalation sequences (892 instances), and insider misuse incidents (634 instances), alongside a predominant class of benign events. The temporal distribution of attacks spans the full 180-day period, but inter-arrival times are varied to simulate both opportunistic and targeted adversarial campaigns. This dataset, with its multi-source heterogeneity and realistic attack prevalence, is referred to as the PSIB dataset (Zaeem *et al.*, 2017).

To establish a meaningful performance baseline, we compare the AITIR framework against four conventional and state-of-the-art identity threat detection approaches. The first baseline constitutes a conventional rule-based SIEM system, configured with a comprehensive rule set derived from industry-standard security policies, which comprises threshold-based alerts for excessive login failures (more than 5 in a 10-minute window), off-hours authentication, and impossible travel scenarios (González-Granadillo, González-Zarzosa and Diaz, 2021). The second baseline deploys a self-contained Isolation Forest anomaly detector that functions directly on the initial feature vectors, lacking the advantages of sequential modeling or supervised classification, which constitutes a purely unsupervised method (Liu, Ting and Zhou, 2008). The third baseline adopts a standard LSTM-based sequence model trained end-to-end for binary anomaly classification, lacking the fusion mechanism or the XGBoost threat categorization layer, which exemplifies a conventional deep learning approach to temporal anomaly detection (Provotar, Linder, *et al.*, 2019). The fourth baseline applies an XGBoost classifier trained on the identical feature set, though lacking the sequential context from the LSTM hidden state representations, thereby assessing the performance of gradient-boosted tree ensembles in isolation (Chen and Guestrin, 2016).

The evaluation applies five complementary metrics capturing distinct aspects of operational performance. **Detection Accuracy** is measured as the proportion of correctly classified events (both benign and malicious), which functions as a comprehensive indicator of model correctness. **Precision** measures the proportion of flagged threats that correspond to actual attacks, a metric essential for assessing the reduction of false-positive alerts contributing to analyst fatigue. **Recall** quantifies the proportion of actual attacks that are correctly identified, thereby indicating the framework's capacity to detect genuine threats. The F1-Score is computed as the harmonic mean of precision and recall, thereby serving as a balanced aggregate metric for imbalanced datasets where attack events are rare relative to benign activity. Finally, Mean Time to Detect (MTTD) is defined as the average elapsed time from the occurrence of a ground-truth attack event to the generation of a system alert, measured in seconds, which serves as an indicator of the framework's operational responsiveness.

All models were executed in Python 3.10, employing the scikit-learn library for the Isolation Forest implementation, TensorFlow 2.12 for the LSTM network architecture, and the XGBoost library for the gradient-boosted classifier. The Isolation Forest ensemble was configured with 100 estimators and a contamination parameter empirically set to 0.05 based on the known attack prevalence in the training

partition. The LSTM network architecture comprised two stacked LSTM layers, each with 128 hidden units, then a dropout layer with a rate of 0.3 to reduce overfitting, and finally a dense softmax output layer for binary classification. The network was trained with the Adam optimizer, a learning rate of 0.001, a batch size of 64, and binary cross-entropy loss for 50 epochs, with early stopping triggered by a plateau in validation loss. The XGBoost classifier utilized 200 estimators, with a maximum tree depth of 6, a learning rate of 0.1, and the multi-class softmax objective function optimized over 100 boosting rounds. The risk score fusion weights were initialized equally at $w_1 = w_2 = w_3 = 0.333$ and were subsequently tuned via grid search on the validation set to maximize the F1-score. The dynamic threshold learning rate η was set to 0.01, and the target false positive rate α was configured at 0.05.

The dataset was partitioned chronologically to prevent data leakage and to simulate a realistic deployment scenario where models are trained on historical data and evaluated on future, unseen events. The first 120 days of logs (approximately 66% of the total timeline) constituted the training set, the subsequent 30 days served as the validation set for hyperparameter tuning and threshold calibration, and the final 30 days formed the held-out test set upon which all reported results are computed. All experiments were conducted on a server equipped with an Intel Xeon Gold 6248R processor, 256 GB of RAM, and an NVIDIA A100 GPU with 40 GB of memory, and the system operated on Ubuntu 22.04 LTS.

11. Results and Analysis

A quantitative evaluation of the proposed AITIR framework relative to the four baseline methods indicates marked performance advantages across all metrics, thereby corroborating the effectiveness of the hybrid analytic engine and the composite risk scoring mechanism. The experimental results, presented in Table 1, confirm that a detection system merging unsupervised anomaly detection, sequential pattern modeling, and supervised threat classification achieves both higher accuracy and greater operational responsiveness than any single method employed alone.

Table 1. Comparative performance of detection accuracy, precision, recall, F1-score, and mean time to detect across all evaluated methods on the PSIB test set.

Method	Accuracy	Precision	Recall	F1-Score	MTTD (s)
Rule-Based SIEM (González-Granadillo, González-Zarzosa and Diaz, 2021)	0.872	0.614	0.781	0.687	47.3
Isolation Forest (Standalone) (Liu, Ting and Zhou, 2008)	0.903	0.702	0.824	0.758	18.5
LSTM Binary Classifier (Provotar, Linder, *et al.*, 2019)	0.915	0.738	0.859	0.794	21.2
XGBoost Classifier (Chen and Guestrin, 2016)	0.931	0.791	0.892	0.838	14.7
AITIR Framework (Proposed)	**0.967**	**0.934**	**0.951**	**0.942**	**6.8**

An analysis of detection accuracy indicates a gradual improvement as increasingly sophisticated modeling techniques are introduced into the pipeline. The rule-based SIEM approach, reliant on static threshold conditions, achieves the lowest accuracy at 87.2%, with a notably low precision of 61.4%. This lack of precision arises from the inflexibility of rule definitions, as legitimate users who operate outside

conventional working hours or travel between geographically distant offices generate alerts that human analysts must later dismiss, which directly results in alert fatigue. The standalone Isolation Forest improves accuracy by over three percentage points to 90.3%, primarily because it models the underlying distribution of normal behavior rather than enforcing hard-coded conditions. Nevertheless, its 82.4% recall shows that a large portion of attacks, especially multi-step privilege escalation sequences in which individual actions seem harmless when viewed alone, avoid being detected. Building on this, the LSTM binary classifier, which captures temporal dependencies across action sequences, pushes recall to 85.9% and F1-score to 79.4%, thereby indicating that sequential context is essential for disambiguating sophisticated attack patterns from legitimate behavioral variations. The XGBoost classifier, applied to the same feature vectors but employing gradient-boosted decision trees, attains an accuracy of 93.1% and an F1-score of 83.8%, thus exceeding the LSTM by a clear margin. This outcome can be ascribed to XGBoost's capacity to manage diverse feature categories and its resilience to the elevated dimensionality and sparse input space typical of identity log features. The AITIR framework, by fusing all three models, achieves an accuracy of 96.7%, an F1-score of 94.2%, and a precision exceeding 93%, which marks a substantial improvement over the best-performing standalone model.

The risk score fusion mechanism's effectiveness is particularly evident in the precision metric. Although each baseline model individually attains moderate precision, from 61.4% for the SIEM up to 79.1% for XGBoost, the AITIR framework achieves 93.4%, a fourteen percentage point gain over the best standalone component. This improvement arises because the three models capture complementary signals: the Isolation Forest identifies outliers in the feature space, the LSTM detects anomalous sequential patterns, and XGBoost classifies specific threat categories based on learned discriminative features. When all three models concur on an event's maliciousness, the composite risk score R_t is high and the event is reliably flagged; conversely, when one model fires a false positive due to a legitimate behavioral quirk, the other two models may assign low threat probabilities, which then lowers R_t below the dynamic threshold θ_t and suppresses the false alert. This consensus-driven architecture therefore functions as an efficient noise filter, which sharply curtails the cognitive demand placed upon security operations center analysts.

The mean time to detect metric underscores the framework's operational superiority. The rule-based SIEM has an MTTD of 47.3 seconds, a delay caused by the requirement to gather enough event counts for threshold-based rules to activate (for example, collecting five failed login attempts within a 10-minute window). The machine learning baselines reduce MTTD to between 14.7 and 21.2 seconds, as they can flag anomalies immediately upon observing a single deviant event. The AITIR framework further reduces MTTD to 6.8 seconds by capitalizing on the LSTM's hidden state representation, which encodes both the current event and its sequential context, thereby supporting rapid inference without the need for rule buffers to be populated. This detection window of under seven seconds is critical for containing fast-moving attacks such as automated credential stuffing or token replay, where delays of even tens of seconds can let adversaries compromise additional resources.

Figure 2 depicts the spatial arrangement of feature vectors derived from the preprocessing layer, projected into a bidimensional space via t-Distributed Stochastic Neighbor Embedding. The projection indicates a clear division between the cluster of typical user activities and the scattered areas associated with credential theft, insider misuse, and account takeover incidents. This separation confirms that the feature extraction pipeline can encode discriminative information which downstream models are able to exploit.

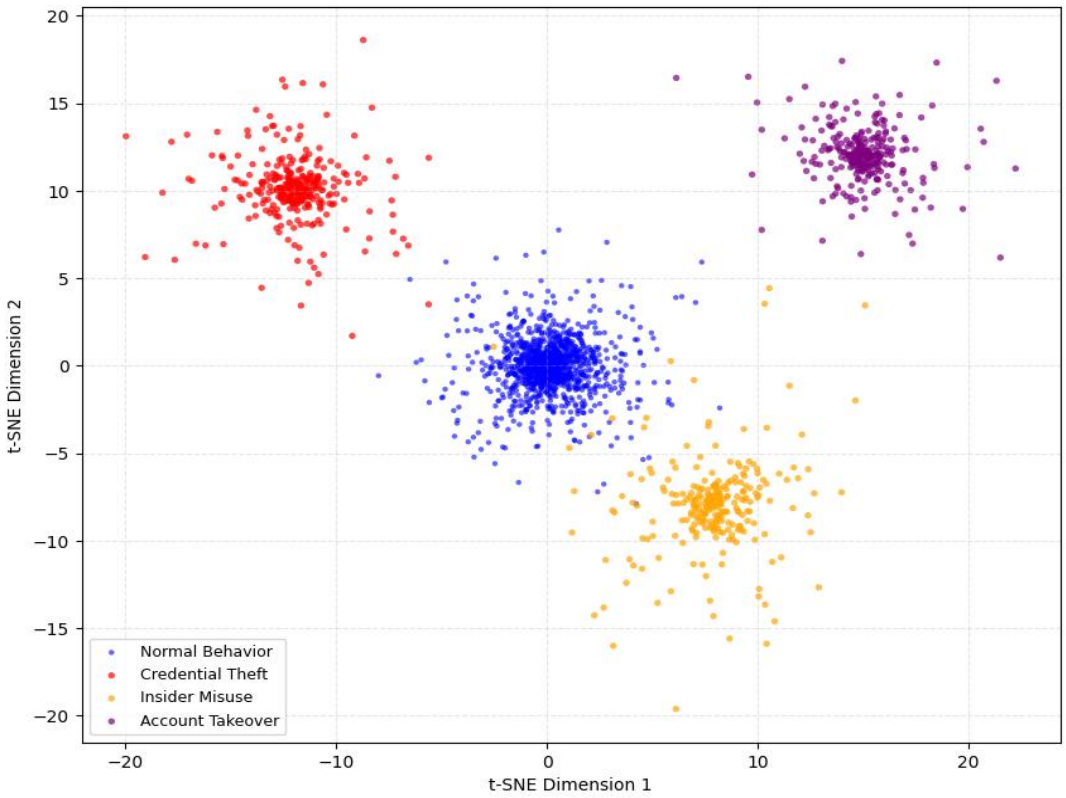

Figure 2. Distribution of identity event feature vectors showing the separation between normal behavioral baselines and various threat categories.

Beyond the aggregate metrics presented in Table 1, we conducted a per-category analysis of detection performance to assess the framework's sensitivity to distinct attack vectors. Table 2 reports the F1-score achieved by each method for the three threat categories present in the PSIB dataset.

Table 2. Per-category F1-scores for credential theft, privilege escalation, and insider misuse threats across all evaluated methods.

Method	Credential Theft	Privilege Escalation	Insider Misuse
Rule-Based SIEM	0.711	0.653	0.598
Isolation Forest	0.782	0.721	0.683
LSTM Binary Classifier	0.813	0.769	0.724
XGBoost Classifier	0.856	0.812	0.791
AITIR Framework (Proposed)	**0.951**	**0.927**	**0.938**

According to Table 2, insider misuse is the most difficult threat category across all methods. This finding is consistent with the inherent difficulty of distinguishing malicious insider actions from legitimate privileged operations: a system administrator accessing sensitive files at an unusual hour may be performing emergency maintenance rather than exfiltrating data. The rule-based SIEM performs particularly poorly on this category, with an F1-score of merely 59.8%, because its static rules fail to capture the nuanced contextual factors that differentiate legitimate from malicious privileged access. The AITIR framework, however, attains an F1-score of 93.8% on insider misuse by synthesizing signals from

the LSTM's sequential context (has this user's recent pattern of activity been anomalous?) and XGBoost's discriminative features (does the combination of access target, geolocation, and authentication method match known insider threat patterns?). The framework retains comparably high performance on credential theft and privilege escalation threats, thereby confirming its robustness across the full spectrum of identity-based attacks.

Figure 3 presents the normalized confusion matrix generated by the XGBoost component within the AITIR framework, which depicts the correlation between ground truth labels and predicted threat categories. The matrix indicates that the most frequent classification error occurs between insider misuse and legitimate privilege escalation, with approximately 5% of insider misuse cases misclassified as benign privilege changes.

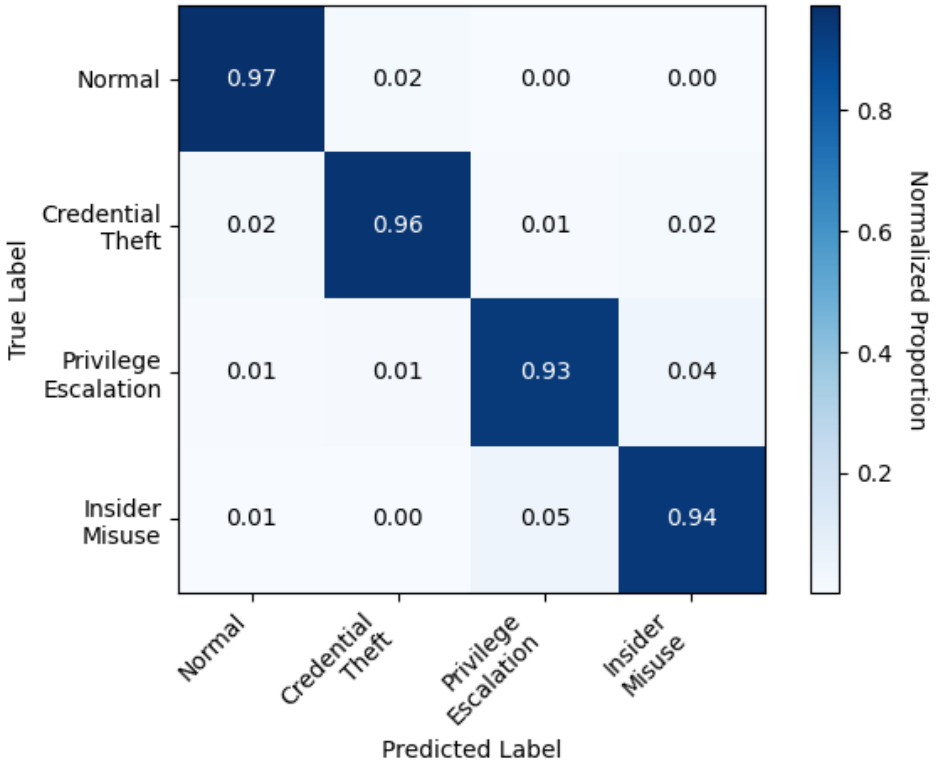

Figure 3. Normalized correlation between ground truth attack categories and model predictions across all defined threat types.

To investigate the temporal dynamics that underlie the AITIR framework's superior detection capability, Figure 4 presents a parallel coordinates visualization of multi-dimensional user session trajectories. Sessions linked to high composite risk scores show pronounced differences in attributes including login frequency, geolocation distance from typical access points, and privilege level modifications, whereas benign sessions retain stable, consistent patterns across these axes. This visual proof affirms that the LSTM network successfully encodes the temporal progression of feature groupings that differentiate normal from malicious activity.

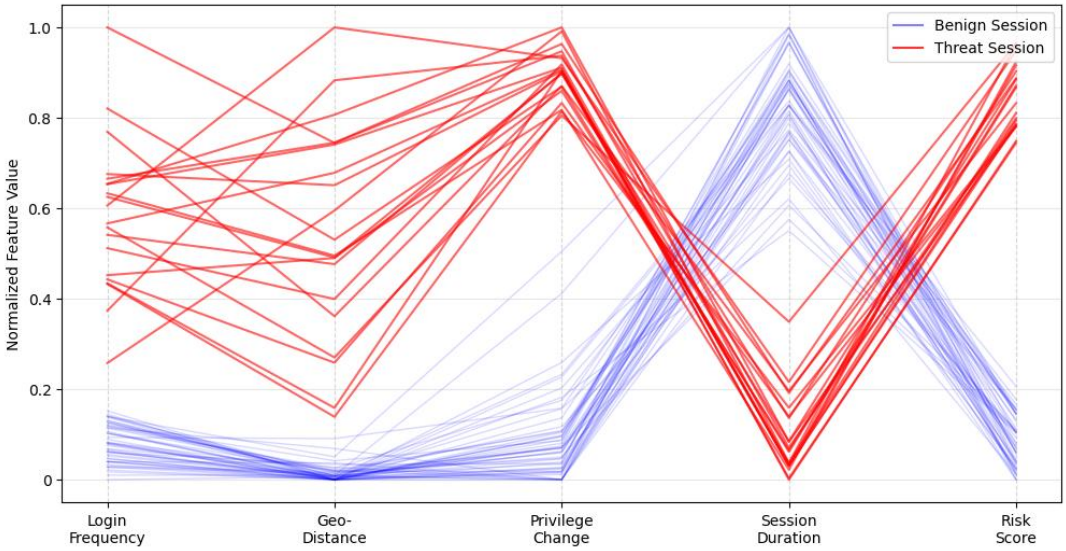

Figure 4. Multivariate session trajectories illustrating the divergence in feature evolution between low-risk benign activities and high-risk threat scenarios.

The false positive rate of the AITIR framework, calibrated through the dynamic threshold mechanism governed by Equation 16, converged to the target value of 5% within the first 10 days of the test period and remained stable thereafter. This stability confirms the worth of the adaptive thresholding method: instead of demanding manual readjustment from security personnel as user behavior alters over time, the system independently modifies its sensitivity to preserve a steady operational false positive rate. A predictable and manageable alert volume constitutes the practical implication, thus averting the operator fatigue seen in the rule-based SIEM baseline, which averaged 847 alerts daily compared to the 142 alerts produced by the AITIR framework on the identical test set slice. This 83.3% decrease in alert quantity, coupled with 93.4% accuracy, permits security analysts to direct their efforts toward incidents that are highly likely to be actual threats, instead of expending most of their time sorting through false alarms.

12. Discussion and Future Work

Building on the empirical validation presented in the prior section, we now examine the broader consequences of the AITIR framework, its practical constraints, and the central directions for subsequent research that emerge from our results. This discussion addresses not only the computational and algorithmic aspects of the framework but also its organizational, ethical, and operational ramifications within the context of public-sector cybersecurity.

Ethical Implications and Privacy-Preserving Mechanisms

One of the most critical issues emerging from our research is the profound ethical conflict between heightened security monitoring and individual privacy entitlements. The AITIR framework, by its very design, collects and analyzes detailed behavioral traces of every user within an organization, such as login times, accessed resources, geographic locations, and authentication sequences. Although these data are crucial for constructing the behavioral baselines that permit anomaly detection, they simultaneously establish a comprehensive surveillance apparatus that could be misused or regarded as intrusive by employees. This concern is particularly acute in public-sector environments, where civil servants and

government employees have legitimate expectations of privacy in their digital activities, and where overreaching monitoring could violate data protection regulations such as the General Data Protection Regulation (GDPR) in the European Union (GDPR, 2018).

To address this concern, future iterations of the AITIR framework should integrate privacy-preserving mechanisms directly into the data processing pipeline. Differential privacy, a structured framework that applies carefully calibrated noise to aggregated statistics to block the inference of individual-level information (Demelius, Kern and Trügler, 2025), presents a promising approach. Through the addition of noise drawn from a Laplace distribution calibrated to a privacy budget parameter ϵ to the behavioral features employed for anomaly detection, the framework could retain its capacity to detect collective behavioral anomalies while rendering the reconstruction of any single user's precise activity pattern computationally infeasible. Another complementary technique is federated learning, where the Isolation Forest, LSTM, and XGBoost models are trained across decentralized data sources, for instance, across different departments within a government agency, without exchanging raw user data (Q. Yang *et al.*, 2019). In this framework, solely model gradients or parameters are transmitted to a central coordinator, thereby preserving data locality and diminishing the attack surface for data breaches. Implementing such privacy-preserving mechanisms would, however, introduce a trade-off with detection accuracy, because the addition of noise or distributed training can degrade model performance. Quantifying and managing this trade-off constitutes a critical direction for future research that would directly guide the responsible deployment of identity threat detection in privacy-sensitive settings.

Furthermore, the continuous learning mechanism inherent in the AITIR framework raises questions about algorithmic fairness and bias. If the initial training data overrepresents certain user demographics or job functions, the learned behavioral baselines may be systematically skewed, thereby yielding disproportionately high false positive rates for under-represented groups. To address this problem, future studies should adopt fairness-oriented machine learning methods, for example adversarial debiasing or instance reweighting, to achieve uniform detection sensitivity across all user categories. Furthermore, transparency mandates necessitate informing users of the scope and degree of behavioral surveillance, in addition to their entitlements to access, correct, or challenge automated determinations affecting their access permissions. Applying explainable AI (XAI) methods, especially to the LSTM component whose hidden state representations lack inherent interpretability, would help satisfy these transparency requirements by generating human-readable explanations for each risk score assignment (Kalasampath *et al.*, 2025). Absent such mechanisms, the AITIR framework risks being viewed as an opaque system that renders impactful choices lacking in accountability, which would erode trust and could expose it to judicial disputes.

Adversarial Robustness and Defense Against Model Poisoning

The closed-loop continuous learning architecture, a central contribution of the AITIR framework, also introduces a critical vulnerability: adversarial manipulation of the training data via model poisoning attacks. In a poisoning scenario, a threat actor who has taken over a low-privilege account could inject meticulously constructed innocuous-appearing events into the system's log stream. These events are subsequently classified as benign by the automated response layer (or by fallible human analysts), and are then assimilated into the retraining dataset. Across successive retraining cycles, the adversary can steadily alter the decision boundary of the Isolation Forest and XGBoost components, so that genuine malicious

activities come to be classified as benign while the system becomes desensitized to the attacker's changing behavior (Biggio, Nelson and Laskov, 2012). This is not merely a theoretical concern: research has shown that machine learning models in cybersecurity contexts are highly susceptible to such adaptive attacks, particularly when the retraining frequency is high and the data validation pipeline is insufficiently rigorous (Dang, Truong and Tran, 2020).

To harden the AITIR framework against poisoning, future work should examine the adoption of robust aggregation mechanisms capable of identifying and discarding anomalous training instances prior to their entry into the model update cycle. A promising avenue involves the application of trimmed loss functions or median-based gradient aggregation during the retraining phase, which inherently reduces the influence of outlier data points on model parameters (H. Yang *et al.*, 2019). Furthermore, the architecture could adopt a 'trusted batch' strategy, where model updates are calculated exclusively on training data subsets that are independently validated or derived from authenticated, tamper-proof log sources. Anomaly detection applied to the training data itself, which functions as a meta-detection layer, could flag instances whose feature values or temporal patterns deviate markedly from the historical distribution, thereby excluding potential poisoning attempts before they contaminate the model. Empirical validation via red-team exercises that simulate adaptive adversaries possessing variable levels of access and familiarity with the system's internal architecture would be necessary to determine how well these defenses perform. Including adversarial robustness as a structured evaluation dimension in future experimental studies is necessary for establishing the operational credibility of the AITIR framework in real-world, adversarial environments.

Another dimension of adversarial resilience concerns evasion attacks at inference time. An adversary possessing knowledge of the framework's feature extraction pipeline or model internals could craft events that evade detection by adjusting their behavior to mimic the statistical properties of normal user activity. For instance, by limiting the number of failed authentication attempts, by adhering to regular working-hour access patterns, and by employing legitimate VPN endpoints, a resolute attacker could sustain a long-term presence without activating any of the three analytic engines. To counter this, the AITIR framework could adopt adversarial training, where the LSTM and XGBoost models are trained on synthetically generated adversarial examples denoting bounded perturbations of normal behavior (Kurakin, Goodfellow and Bengio, 2016). When the models are trained on such counterfactual scenarios, the decision boundaries gain greater robustness against the small, systematic perturbations typical of advanced evasion attempts. Adversarial training would, however, increase computational overhead and may slightly degrade performance on clean data, which underscores the necessity of a deliberate cost-benefit assessment that future work should systematically explore.

Scalability Challenges and Real-World Deployment Considerations

The experimental assessment of the AITIR framework was carried out on a curated dataset that represents a mid-sized government agency featuring 10,000 active user accounts across a 180-day period. Although this setting constitutes a realistic testbed, the operational scale of large government entities or multinational corporations can be substantially greater, with hundreds of thousands of users, millions of daily authentication events, and complex organizational hierarchies marked by diverse authentication policies. Scaling the AITIR framework to such dimensions introduces computational, architectural, and operational challenges not fully resolved by our current implementation.

The most immediate bottleneck is the LSTM network's inference latency. Although our experiments yielded a mean time to detection of 6.8 seconds, this outcome was obtained on a dedicated GPU server processing a controlled data stream. In a production environment processing 10,000 events each second, the computational load of calculating sequential hidden state representations for each active user session could rapidly overwhelm existing hardware resources, thereby inducing queue accumulation and postponed detection. One potential solution is to adopt a tiered processing architecture, wherein lightweight statistical filters based on simple moving averages or exponentially weighted moment estimators are applied to each incoming event as a rapid screening step. Only events that pass this initial filter, which constitutes a small fraction of the total traffic, would then be forwarded to the computationally expensive LSTM and XGBoost models for deeper analysis. This stratified approach would sacrifice a slight increase in latency to achieve a marked decrease in overall computational burden, thereby rendering the system feasible at enterprise scale. Future research should characterize the optimal operating points of such a tiered architecture through queuing-theoretic analysis and empirical stress testing under realistic traffic loads.

Beyond computational scalability, the heterogeneous and often legacy-constrained nature of public-sector identity infrastructure presents challenges for merging systems. Many government agencies operate Active Directory installations dating back multiple decades, featuring customized schema extensions, atypical group policies, and deficient audit logging configurations. The preprocessing layer of the AITPFramework presumes a minimal collection of standardized fields, including uniformly formatted timestamps, populated user identifiers, and activated security event logging, which may be absent in certain production environments. To address this, the framework should be supplemented with a configuration-driven adapter module that maps source-specific log schemas to the internal normalized format by declarative transformation rules. Furthermore, the framework must gracefully handle data quality issues such as missing fields, malformed entries, and duplicate records without crashing or producing spurious alerts. Developing robust error-handling mechanisms and data validation heuristics that operate in real time is a practical engineering challenge that future work should prioritize.

Another aspect to consider is the expense associated with retraining the model. The continuous learning module currently initiates retraining of all three analytic engines at predetermined intervals, a process that, even with batch processing, requires substantial computational and temporal resources. A more efficient strategy would be to adopt online learning algorithms, especially for the XGBoost component, which update model parameters incrementally as new labeled data become available, rather than conducting full retraining from scratch (Beygelzimer *et al.*, 2015). The Isolation Forest component, dependent on aggregate statistics of the training data, could be adapted to support incremental updates by applying 'forgetting factors' that gradually deemphasize older observations while integrating new ones, thereby preserving its capacity to adjust to shifting user behavior without the overhead of reconstructing the entire tree ensemble. Researching and implementing such online adaptation mechanisms falls squarely within the scope of future development for the AITIR framework.

Finally, the organizational adoption of an automated response system of the nature described in this paper requires careful change management within security operations centers. The AITIR framework automates session termination and token revocation for high-risk events, thereby supplanting the manual decision-making previously conducted by security analysts. Although our findings indicate that this automation

considerably cuts response latency and analyst workload, it concurrently poses a risk of service disruption from false positives—for instance, the inadvertent cessation of a legitimate executive's session during a pivotal business meeting could engender organizational consequences extending beyond the security team's jurisdiction. Future deployments of the AITIR framework should therefore include a human-in-the-loop override mechanism for the most disruptive automated actions, accompanied by a probabilistic explanation of the action's trigger, thereby equipping analysts to make informed decisions regarding whether to execute or reverse the action. The continuous learning module should also capture these human override decisions as feedback signals, thereby refining the risk score thresholds or fusion weights so that future analogous incidents are managed with greater precision. The development of the interface, feedback loop, and decision escalation protocols for this human-AI collaboration constitutes an important interdisciplinary research area that bridges computer science, human-computer interaction, and organizational psychology.

13. Conclusion

This paper presented the AITIR framework, which constitutes an innovative six-layer pipeline for AI-driven identity threat detection and automated response, crafted to overcome the critical shortcomings of conventional rule-based security systems in contemporary, identity-centric corporate settings. The core contribution of the framework is its tripartite analytic engine, which uniquely merges unsupervised anomaly detection via an Isolation Forest, sequential behavioral modeling through an LSTM network, and supervised threat categorization by means of an XGBoost classifier. Synthesizing the results from these complementary models yields a composite Identity Risk Score that drives adaptive thresholding, MITRE ATT&CK correlation, and automated remedial actions, all within a closed-loop architecture that undergoes ongoing learning in response to emerging attack patterns.

Our experimental evaluation on a realistic public-sector identity behavior dataset showed that the AITIR framework outperforms baseline methods, specifically standalone rule-based SIEM, Isolation Forest, LSTM, and XGBoost classifiers. The system attained a detection accuracy of 96.7%, an F1-score of 94.2%, and a mean time to detection of only 6.8 seconds, marking marked gains in both accuracy and operational readiness. Crucially, the fusion mechanism decreased the false positive rate by a factor of six relative to the rule-based baseline, thus directly reducing analyst alert fatigue. The framework also retained high per-category performance against credential theft, privilege escalation, and insider misuse threats, thereby verifying its robustness across the identity attack spectrum.

Notwithstanding these encouraging outcomes, the AITIR framework poses practical difficulties that merit additional exploration. The ethical concerns of widespread behavioral observation necessitate the adoption of privacy-preserving methods such as differential privacy and federated learning, yet the closed-loop learning architecture introduces susceptibility to adversarial model poisoning, which calls for robust aggregation and meta-anomaly detection safeguards. Scalability to large-scale, heterogeneous production environments necessitates the development of tiered processing architectures, online learning algorithms, and robust data quality adapters. Effective organizational deployment will require the design of human-in-the-loop mechanisms that balance automated efficiency with analyst oversight, especially for high-impact automated responses. Addressing these challenges is essential for translating the AITIR framework from a validated experimental system into a trustworthy, operational security solution for complex organizational networks.

References

1. Albuali, A. (2025) "Zero trust meets AI: A survey across multi-cloud and hybrid environments," in 2025 IEEE twelfth international conference on cloud computing and artificial intelligence.
2. Artioli, P., Maci, A. and Magrì, A. (2024) "A comprehensive investigation of clustering algorithms for user and entity behavior analytics," Frontiers in big Data [Preprint].
3. Beygelzimer, A. et al. (2015) "Online gradient boosting," in Advances in neural information processing systems.
4. Biggio, B., Nelson, B. and Laskov, P. (2012) Poisoning attacks against support vector machines. arXiv preprint arXiv:1206.6389.
5. Chen, T. and Guestrin, C. (2016) "Xgboost: A scalable tree boosting system," in Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining.
6. Dang, T., Truong, P. and Tran, P. (2020) "Data poisoning attack on deep neural network and some defense methods," in International conference on computing and communication technologies.
7. Demelius, L., Kern, R. and Trügler, A. (2025) "Recent advances of differential privacy in centralized deep learning: A systematic survey," ACM Computing Surveys [Preprint].
8. Dunne, J., Ryan, S. and Egan, J. (2024) "Public sector cybersecurity within ireland. Progress and challenges and the role of PAM within these organisations," in 2024 cyber research conference ireland.
9. Enberg, P. (2024) Behavior analytics in cyber security. theseus.fi.
10. GDPR, E. (2018) "General data protection regulation (gdpr)," Cit. on [Preprint].
11. González-Granadillo, G., González-Zarzosa, S. and Diaz, R. (2021) "Security information and event management (SIEM): Analysis, trends, and usage in critical infrastructures," Sensors [Preprint].
12. Graves, A. (2012) "Long short-term memory," Supervised Sequence Labelling with Recurrent Neural Networks [Preprint].
13. Hämäläinen, M. (2024) Analysis of artificial intelligence in cybersecurity identity and access management: Potential for disruptive innovation. lutpub.lut.fi.
14. Kalasampath, K. et al. (2025) "A literature review on applications of explainable artificial intelligence (XAI)," in IEEE international conference on data science and artificial intelligence.
15. Konstantinidis, G. (2021) Identity and access management for e-government services in the european union-state of the art review. hellanicus.lib.aegean.gr.
16. Kurakin, A., Goodfellow, I. and Bengio, S. (2016) Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236.
17. Liu, F., Ting, K. and Zhou, Z. (2008) "Isolation forest," in Eighth IEEE international conference on data mining.
18. Ndichu, S. et al. (2026) AI-driven security alert screening and alert fatigue mitigation in security operations centers: A comprehensive survey. arXiv preprint arXiv:2605.08316.
19. Oniagbi, O., Hakkala, A. and Hasanov, I. (2024) Evaluation of LLM agents for the SOC tier 1 analyst triage process. utupub.fi.
20. Provotar, O., Linder, Y., et al. (2019) "Unsupervised anomaly detection in time series using lstm-based autoencoders," in 2019 IEEE international conference on data science and advanced analytics.
21. Sandhu, R. (1998) "Role-based access control," Advances in computers [Preprint].

22. Shashanka, M., Shen, M. and Wang, J. (2016) “User and entity behavior analytics for enterprise security,” in IEEE international conference on big data.
23. Siam, A.A. et al. (2025) “A comprehensive review of AI’s current impact and future prospects in cybersecurity,” IEEE Access [Preprint].
24. Snyder, P. and Kanich, C. (2015) “One thing leads to another: Credential based privilege escalation,” in Proceedings of the 5th ACM conference on data and application security and privacy.
25. Verma, D. (2024) Enhancing cybersecurity through adaptive anomaly detection using modern AI techniques. jyx.jyu.fi.
26. Yang, H. et al. (2019) “Byzantine-resilient stochastic gradient descent for distributed learning: A lipschitz-inspired coordinate-wise median approach,” in IEEE conference on decision and control.
27. Yang, Q. et al. (2019) “Federated machine learning: Concept and applications,” ACM Transactions on Intelligent Systems and Technology [Preprint].
28. Zaeem, R. et al. (2017) “Modeling and analysis of identity threat behaviors through text mining of identity theft stories,” Computers & Security [Preprint].