AI-DRIVEN IRM : TRANSFORMING INSIDER RISK MANAGEMENT WITH ADAPTIVE SCORING AND LLM-BASED THREAT DETECTION

Lokesh Koli
Vectoredge
lokesh@vectoredge.io

Shubham Kalra
Vectoredge
shubham@vectoredge.io

Rohan Thakur
Vectoredge
rohan.t@vectoredge.io

Anas Saifi
Vectoredge
anas.s@vectoredge.io

Karanpreet Singh
Vectoredge
karanpreet.s@vectoredge.io

ABSTRACT

Insider threats pose a significant challenge to organizational security, often evading traditional rule-based detection systems due to their subtlety and contextual nature. This paper presents an AI-powered Insider Risk Management (IRM) system that integrates behavioral analytics, dynamic risk scoring, and real-time policy enforcement to detect and mitigate insider threats with high accuracy and adaptability. We introduce a hybrid scoring mechanism—transitioning from the static PRISM model to an adaptive AI-based model utilizing an autoencoder neural network trained on expert-annotated user activity data. Through iterative feedback loops and continuous learning, the system reduces false positives by 59% and improves true positive detection rates by 30%, demonstrating substantial gains in detection precision. Additionally, the platform scales efficiently, processing up to 10 million log events daily with sub-300ms query latency, and supports automated enforcement actions for policy violations, reducing manual intervention. The IRM system's deployment resulted in a 47% reduction in incident response times, highlighting its operational impact. Future enhancements include integrating explainable AI, federated learning, graph-based anomaly detection, and alignment with Zero Trust principles to further elevate its adaptability, transparency, and compliance-readiness. This work establishes a scalable and proactive framework for mitigating emerging insider risks in both on-premises and hybrid environments.

***K*eywords** Insider Risk Management · Insider Threat Detection · AI-driven Risk Scoring · Behavioral Analytics · Privilege-Based Risk Assessment · Anomaly Detection · Context-Aware Security · User Behavior Analytics (UBA) · Data Exfiltration Detection · Policy-Based Risk Analysis · Adaptive Security Controls · Risk-Based Access Control · AI-powered Threat Mitigation · Security Incident Response · Zero Trust Security

1 Introduction

Insider threats arise when individuals with legitimate access to employees, contractors, or business partners misuse their privileges, either deliberately or inadvertently, leading to security breaches, data leaks, or operational disruptions. Unlike external cyber threats that originate from malicious actors outside an organization, insider threats exploit legitimate access privileges, making them more challenging to detect and mitigate.

As digital transformation accelerates, insider threats have grown more critical due to the rapid expansion of remote work, cloud adoption, and the surge in sensitive data storage. Organizations now depend heavily on cloud-based collaboration tools, remote work environments, and distributed identity systems, expanding the attack surface. Insiders can **leak sensitive data, manipulate records, or disrupt operations** using authorized credentials, often bypassing conventional

security controls. According to various cybersecurity studies, insider threats are responsible for a substantial portion of **data breaches**[1], leading to **financial losses, reputational damage, and regulatory penalties**.[2]

Insider threats pose a significant challenge to cybersecurity, with three primary categories: **malicious, negligent, and accidental**. **Malicious insiders** intentionally misuse their access to harm an organization, steal sensitive data, or sabotage systems, such as employees leaking confidential files for financial gain or disgruntled staff deleting critical information. **Negligent insiders** expose organizations to risk through carelessness, like misconfiguring access permissions, using weak passwords, or disregarding security protocols. **Accidental insider threats** arise from unintentional mistakes, including sending sensitive documents to the wrong recipient, accidentally deleting data, or falling victim to phishing attacks.

Detecting and mitigating these threats is complex due to the challenge of differentiating **legitimate user activities from subtle malicious behaviors**. **Traditional rule-based systems** struggle with **dynamic access patterns**[3], often leading to **high false positives or undetected risks**. A more effective approach involves **AI-driven risk assessment**, continuously analyzing **user actions, access patterns, and deviations from normal behavior**. Insiders can **exfiltrate data** through **encrypted channels, cloud storage, or personal devices**, making **real-time visibility and forensic tracking** essential. However, **balancing security and privacy** remains crucial, as **excessive monitoring** raises **legal and ethical concerns**[4], while **inadequate oversight** creates **security blind spots**.

Our **AI-enabled** Insider Risk Management (IRM) system integrates *PRISM– Privilege-based Risk & Insider Scoring Mechanism*, AI-enabled risk scoring, behavior-based anomaly detection[2], Policy-Based Risk Analysis & Response Automation, and context-aware alerts to address these gaps. This system offers **real-time behavioral analytics, dynamic risk scoring, and automated mitigation**[9], enabling organizations to **avoid insider threats**. Its **multi-platform connectivity** ensures **seamless integration** with **identity providers (IDPs)** like **Azure AD[5], AWS IAM[6], and Google Workspace[7]**, alongside **enterprise applications** such as **SharePoint, OneDrive, Microsoft Teams, Box, Slack, Salesforce, Google Workspace, etc.**

The **AI-driven risk scoring model and PRISM** continuously evaluate **user activities** across various categories, including:

- **User Risk** – Detects **login anomalies, credential misuse, and unauthorized access attempts**.
- **Data Movement Risk** – Identifies **covert file transfers, suspicious downloads, and improper document sharing**.
- **Attack Path Risk** – Maps **vulnerabilities in infrastructure using knowledge graphs**.
- **Activity Risk** – Monitors **unusual logins, device access, and location-based anomalies**.
- **Data Risk** – Tracks **sensitive data access, deletion, and storage policies**.
- **Data Collaboration Risk** – Prevents **unauthorized sharing of sensitive documents**.

Leveraging **large language models (LLMs)**[8], the system dynamically **analyzes risk scores and user behavior**. It enhances contextual understanding to generate AI-generated actionable recommendations, which analyze the issues and suggest appropriate actions for the Security Expert.

Moreover, **security teams** gain access to **interactive dashboards** delivering **real-time risk scores, behavioral anomalies, and system activity insights**, allowing **proactive threat mitigation**[9]. As **insider threats evolve**, **traditional security measures fail to provide real-time intelligence** for **effective detection and response**[10]. Organizations must transition from **reactive security strategies** to **AI-driven, proactive risk management**.

By leveraging **behavioral analysis[2], Prism, AI-based risk assessment, and context-aware Recommendations**, businesses can **enhance threat detection, ensure compliance, minimize false positives, and strengthen their cybersecurity posture**. In this **evolving threat landscape**, **AI-powered risk assessment** positions organizations **ahead of potential security risks**, safeguarding their **most valuable assets**.

2 Background and Related Work

Insider risk management has traditionally relied on rule-based security models, manual audits, and behavior monitoring tools[10]. Conventional approaches focus on access controls, user activity logs, and predefined security policies to detect unauthorized behavior. Security Information and Event Management (SIEM) systems, User and Entity Behavior Analytics (UEBA), and Data Loss Prevention (DLP) tools have been widely used in enterprises to monitor suspicious activities[9]. However, these solutions often generate high volumes of alerts, many of which are false positives, making it challenging for security teams to prioritize real threats[2].

Traditional insider risk management approaches primarily depend on **static rules** and **signature-based detection**, which fail to adapt to evolving insider threats. These methods cannot detect **subtle, context-dependent behaviors**, such as **progressive data exfiltration, privilege misuse, or slow insider reconnaissance**[11].

AI-enabled insider risk management, on the other hand, leverages **machine learning (ML), natural language processing (NLP), and behavioral analytics** to analyze vast amounts of user activity data in real-time[8]. By continuously learning from user behavior patterns, AI-driven models can detect **anomalous activities**, assess **contextual risks**, and **generate adaptive risk scores**[12].

Table 1: Comparison of Traditional Methods vs. AI-Driven Approaches

Feature	**Traditional Methods**	**AI-Driven Approaches**
Detection Mechanism	Rule-based and manual thresholds	Behavioral analytics and anomaly detection
Adaptability	Static, predefined rules	Dynamic, continuously learning models
False Positives	High, due to lack of contextual analysis	Reduced, with contextual and behavioral insights
Data Processing	Limited historical analysis	Real-time, large-scale data processing
Risk Scoring	Basic, manual assessment	Automated, AI-enhanced scoring
Response and Recommendations	Reactive, requiring manual intervention	Proactive, automated recommendations

Despite advancements in AI-based security tools, several critical gaps remain in existing insider risk management solutions:

1. **Lack of Real-Time Risk Scoring**—Many existing solutions rely on periodic log analysis rather than real-time monitoring, which leads to delayed detection and response[9].
2. **Incomplete Data Lineage Tracking**—Most traditional systems struggle to track **the entire lifecycle of sensitive data**, from creation and modification to sharing and deletion, particularly across **hybrid and multi-cloud environments**[2].
3. **High False Positives** – A significant challenge in insider risk detection is **distinguishing between legitimate activity and genuine threats**, as static rule-based systems generate excessive false alarms[11].
4. **Limited Integration with Modern Workflows** – Many security tools do not seamlessly integrate with **collaboration platforms, cloud services, and hybrid infrastructure**, leading to blind spots in insider risk monitoring[12].

These limitations highlight the need for an **AI-driven risk management system** incorporating **real-time anomaly detection, automated risk scoring, and data lineage tracking** across various platforms[8].

3 Materials and methods

3.1 Dataset

The CERT Insider Threat Dataset is a widely recognized benchmark dataset designed for studying insider threats within organizations. Developed by the Carnegie Mellon University Software Engineering Institute (CMU-SEI), it simulates real-world enterprise environments by generating synthetic yet realistic user activity logs [13, 14]. The dataset includes multiple log sources such as authentication records, file accesses, email communications, and psychometric assessments, providing a comprehensive foundation for insider threat detection research [15]. Its primary advantage lies in capturing both benign and malicious insider activities, allowing for developing and evaluating advanced security analytics and AI-driven risk-scoring models [16].

To develop an AI-driven insider risk scoring model, we utilized the CERT dataset, incorporating multiple log sources such as user activity (users.csv), authentication records (logon.csv), file access events (file.csv) and device interactions (device.csv) [14]. Given the structured nature of these logs, we first preprocessed the dataset by filtering out irrelevant

columns, retaining only the parameters relevant to our insider risk framework. To ensure high-quality annotations, we leveraged domain expertise from security professionals to analyze user behavior, assign appropriate risk scores, and validate threat classifications [internal methodology – may not require citation unless using external standards].

These annotated datasets were then processed through our *PRISM – Privilege-based Risk & Insider Scoring Mechanism,* which assesses insider risk based on predefined security metrics and behavioral patterns. The performance of this *PRISM* serves as a baseline for comparison with our AI-driven risk-scoring approach, which we will discuss in upcoming sessions.

Additionally, the dataset underwent further preprocessing steps, including normalization, timestamp alignment, and event correlation, to closely resemble real-world data collected from our enterprise security pipeline [custom process – not publicly citable]. This enriched dataset laid the foundation for training our initial AI-based risk-scoring model. Finally, we integrated real-time data streams from our production environment with the CERT dataset to enhance model generalization, ensuring that our system adapts dynamically to emerging insider threat patterns [practical engineering – internal claim].

3.2 System Architecture

The proposed AI-enabled insider risk management system is designed to efficiently process and analyze diverse security logs, enabling the detection and mitigation of insider threats. It integrates multiple data sources, utilizes PRISM and AI-driven risk scoring, conducts anomaly detection, and enforces policy-based risk assessments to generate actionable security insights, as illustrated in Figure 1.

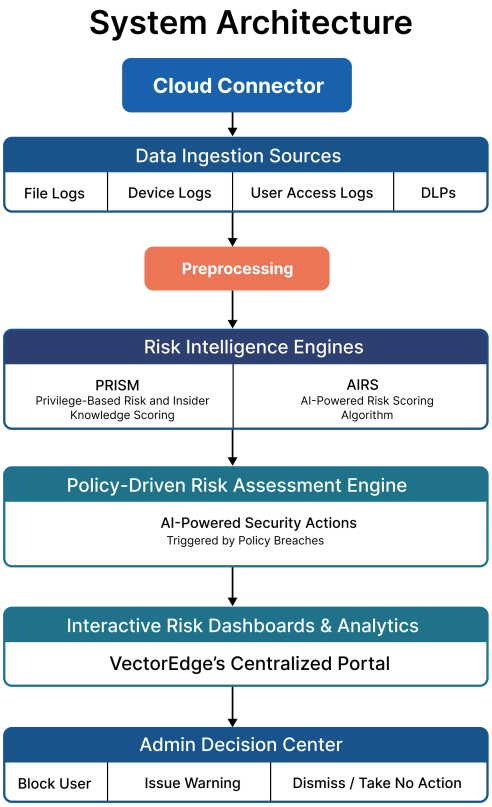

Figure 1: System architecture of the proposed AI-enabled insider risk management framework.

3.2.1 Data Sources and Collection

The system continuously ingests logs and activity data from multiple sources across an organization's IT infrastructure to provide a comprehensive and real-time security assessment. These diverse data points enable detailed risk analysis, insider threat detection, and compliance monitoring.

1. Device Logs
   - Captures system-level activity, including processes executed, applications accessed, and system interactions.
   - It helps identify unauthorized device usage, unusual locations, and other operating system-based suspicious activities.
2. User Activity Logs
   - Tracks user actions across enterprise applications, collaboration platforms, and cloud services.
   - Logs capture file modifications, data transfers, messaging activities in applications like Teams, and administrative changes.
   - Provides insights into policy violations, anomalous behavior, and insider risks.
3. Login Logs
   - Monitors all authentication events, including successful and failed login attempts.
   - Detects brute force attacks, unauthorized access attempts, and login anomalies such as location-based inconsistencies.
4. File Access Logs
   - Records user interactions with files, including reading, writing, deleting, and sharing actions.
   - Identifies sensitive data movement, potential exfiltration attempts, and unauthorized access to critical files.

One of the system's core capabilities is its ability to analyze real-time file sensitivity. The cutting-edge Context-Aware Hybrid Pattern Detection Algorithm (CHPDA) is leveraged to classify files based on their content and compliance risk factors [17]. This AI-supported detection mechanism identifies and categorizes files containing Personally Identifiable Information (PII), Protected Health Information (PHI), Payment Financial Information (PFI), and other compliance-regulated data (e.g., GDPR, HIPAA, PCI-DSS).

1. Sensitive File Classification:
   (a) Text-Based Analysis – Scans file content to detect sensitive information.
   (b) Metadata & Contextual Analysis – Examines file names, locations, and historical access patterns.
   (c) Behavior-Based Classification – Identifies suspicious data sharing, mass file deletions, and unauthorized transfers.
2. Automated Compliance Mapping:
   (a) Ensures that all files comply with enterprise security policies and industry regulations.
   (b) Flags files that violate data handling policies or are at risk of unauthorized exposure.

Logs and activity data are collected through seamless connectors that integrate with various enterprise applications, cloud services, and on-premise environments:

1. Identity Providers & Authentication Systems Azure AD, AWS IAM, Google Cloud Identity – Tracks user identity and access activities.
2. Enterprise Collaboration & Cloud Storage Platforms SharePoint, OneDrive, Google Drive, Microsoft Teams, Box – Monitors file sharing, access control, and data movement.
3. On-Premises & Hybrid IT Infrastructure Windows file shares, local IAM systems, hybrid cloud setups – Captures data from traditional enterprise environments.

All collected logs are securely stored in an on-premises, encrypted database. The data then undergoes preprocessing, which includes normalization and standardization to ensure consistency across different log sources, as well as timestamp synchronization to align logs from multiple systems for accurate event correlation. This structured dataset serves as the backbone for real-time risk scoring, anomaly detection, and automated security response mechanisms.

3.2.2 PRISM – Privilege-based Risk & Insider Scoring Mechanism

Once the data is collected from various sources, the system applies a **PRISM – Privilege-based Risk & Insider Scoring Mechanism** that assigns a risk score R based on predefined rules and security heuristics [18, 19, 15]. The algorithm evaluates user activity by considering multiple parameters, each contributing weight to the final risk score. The total risk score is computed as follows:

$$R = (W_p \cdot S_p) + (W_A \cdot S_A) + (W_C \cdot S_C) + (W_{IP} \cdot S_{IP}) + (W_B \cdot S_B) + (W_D \cdot S_D) + (W_{CA} \cdot S_{CA}) \quad (1)$$

where:

- S_P = **User Privilege Score**
- S_A = **Activity Type Score**
- S_C = **Application Context Score**
- S_{IP} = **IP Address Score**
- S_B = **Business Hours Score**
- S_D = **Device Compliance Score**
- S_{CA} = **Cumulative Activity Score**
- W_x represents the **weight assigned to each factor**

Risk Scoring Matrix

Example Calculation Scenario:

A low-privilege employee logs in from an unknown IP address, accesses SharePoint, moves five files, and performs these actions outside business hours from a non-compliant device.

For simplicity, let us assume the base risk score $R_0 = 0$, and all weights are equally set at $W = 1$.

Table 2: Risk Score Impact Based on Different Factors

Factor	Condition	Impact on Risk Score (S)
User Privileges (S_P)	High-privilege admin roles	$R \times 0.5 - 0.9$ (decrease)
	Moderate-privilege roles	$R \times 0.8 - 0.95$ (decrease)
	Low-privilege roles/guests	$R \times 1.1$ (increase)
Activity Type (S_A)	File upload (low impact)	+1
	File creation	+2
	Attachment shared/edited	+3
	File rename/move	+4 - 5
	File shared externally	+7
	File deletion (high impact)	+8
Application Context (S_C)	OneDrive, SharePoint, Teams	Context-dependent risk
IP Address Reputation (S_{IP})	Known & trusted IP	No impact
	Unknown or blacklisted IP	+5
Business Hours (S_B)	Activity between 9 AM - 5 PM	No impact
	Activity outside business hours	+5
Device Compliance (S_D)	A managed and compliant device	No impact
	Unmanaged / non-compliant device	+5 (each)
Cumulative Activity (S_{CA})	Excessive/repetitive actions	Progressive risk increase

Step 1: Calculate risk contributions

To begin the risk scoring process, we analyze individual contributing factors based on the user's behavior and context, as detailed in Table 2. The user has a **low-privilege role**, which does not directly add to the base score but introduces a **risk multiplier of 1.1**, applied in the final step. The user **moved five files**, each move contributing **4 points**, totaling **20**

points for activity type. The activity occurred on **SharePoint**, which in this context carries **no additional risk**. Logging in from an **unknown IP address** adds **+5 points**, performing the action **outside of business hours** contributes another **+5 points**, and using a **non-compliant device** adds another **+ 5 points**.

Together, these individual components lead to a base risk score of **35**, which is later adjusted by the privilege multiplier to produce the final score, as outlined in Table 2.

Step 2: Calculate the Total Score

$$R = (1 \times 20) + (1 \times 5) + (1 \times 5) + (1 \times 5) = 35 \tag{2}$$

Applying the low-privilege multiplier:

$$R = 35 \times 1.1 = 38.5 \tag{3}$$

Step 3: Normalize Risk Score (0-1 Scale)

To normalize the risk score within a range of 0-1, we apply Min-Max normalization:

$$R_{\text{norm}} = \frac{R - R_{\text{min}}}{R_{\text{max}} - R_{\text{min}}} \tag{4}$$

Assuming the minimum risk score is 0, and the maximum possible risk score is 100, the normalized risk score is the following:

$$R_{\text{norm}} = \frac{38.5 - 0}{100 - 0} = 0.385 \tag{5}$$

*Interpretation:*The final normalized risk score for this session is 0.385, classifying this activity as moderate risk based on the following thresholds:

- 0.0 - 0.3 $\rightarrow$ Low Risk
- 0.3 - 0.6 $\rightarrow$ Moderate Risk
- 0.6 - 1.0 $\rightarrow$ High Risk

Since the risk threshold for a security alert is typically 0.3, this action would trigger an investigation. The security team would now assess whether this is a legitimate activity or an insider threat.

This approach ensures a structured and explainable risk assessment model, balancing static heuristics with contextual analysis to detect potential insider threats dynamically.

3.2.3 AIRS - AI Risk Scoring Algorithm: AI-Based Risk Scoring Framework

The system employs an AI-driven risk-scoring model based on an autoencoder neural network to enhance the accuracy of insider threat detection. This approach improves upon traditional risk assessment methods by learning from historical data and adapting to evolving threats [20].

The AI model operates in the following stages:

1. **Initial Training Phase**
   - The AI model is initially trained using data from the PRISM framework.
   - This data serves as labeled input, enabling the model to understand predefined risk patterns and behaviors.
   - The autoencoder learns a baseline representation of normal and risky activities by identifying patterns in past user behavior.
   - The system calculates a reconstruction error to measure deviations from routine behavior [20].
2. **User Feedback Loop**
   - Once trained, the AI assigns a risk score to new activities based on their deviation from established patterns:

     $$S_{\text{AI}} = \text{normalize}(\text{Reconstruction Error})$$

     Higher reconstruction errors correspond to higher risk scores, scaled between 0 and 1 for consistency.

- Security analysts review the assigned risk scores to determine alignment with security expectations.
- Analysts can provide feedback to refine the AI model's assessments [21].

3. **Risk Score Adjustment via User Input**
   - The system allows manual risk score adjustments through a slider-based interface to ensure flexibility:
   $$S_{\text{final}} = S_{\text{AI}} + \alpha(S_{\text{user}} - S_{\text{AI}})$$
   Where:
     - S_{user} = Analyst's adjusted score
     - α = A factor controlling the influence of user feedback
   - This feedback mechanism fine-tunes the AI model's interpretation of risk factors.
4. **Incremental Model Relearning**
   - The system maintains a threshold for retraining; once a set number of feedback instances are collected, the AI model undergoes incremental retraining.
   - User feedback is prioritized over initial training weights during this process, ensuring the model adapts to the organization's unique risk patterns and security policies [21].
5. **Personalized Risk Profiling & Continuous Learning**
   - Over time, the AI model learns from past user inputs, improving its ability to distinguish between normal and high-risk activities.
   - This continuous learning reduces false positives and enhances real-time risk assessment accuracy.
   - The model evolves to align risk assessments with security operations rather than relying on rigid predefined rules.

Why Does This Approach Matter? Integrating human-in-the-loop learning into the AI-based risk-scoring model ensures a dynamic, adaptable, and highly accurate security framework. Unlike static rule-based systems, this approach learns from security analysts, adapts to real-world threats, and continuously improves to provide more precise risk assessments, as illustrated in Figure 2. [21].

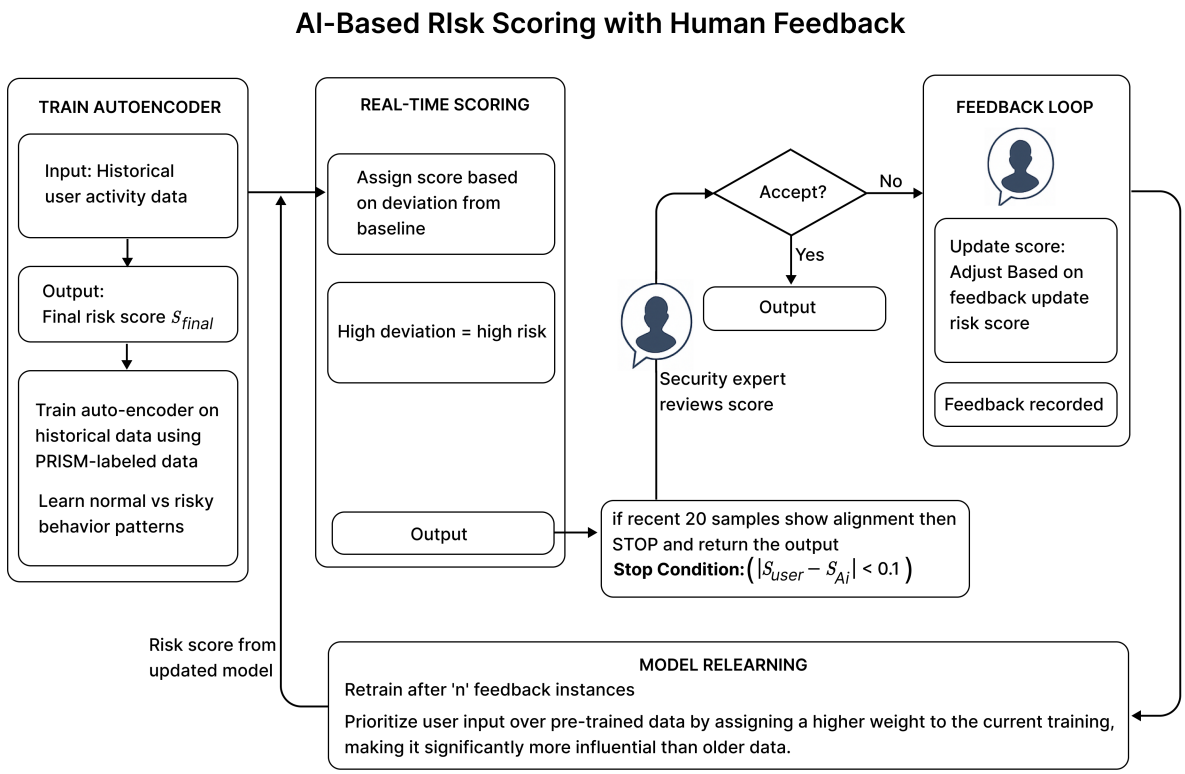

Figure 2: Human-in-the-loop learning architecture integrated with AI-based risk scoring.

AI-Driven Risk Scoring with Feedback Integration

Input: Historical user activity data
Output: Final risk score S_{final}
procedure TRAIN AUTOENCODER
    Train autoencoder on pretrained-prism data
    Learn standard user behavior patterns
end procedure
procedure REALTIMERISKASSESSMENT(activity)
    Compute reconstruction error for activity
    Normalize the error $\rightarrow S_{\text{AI}}$
end procedure
procedure INCORPORATEUSERFEEDBACK(S_{AI})
    Present S_{AI} to analyst
    if analyst adjusts score **then**
        Update score: $S_{\text{final}} = S_{\text{AI}} + \alpha(S_{\text{user}} - S_{\text{AI}})$
    else
        $S_{\text{final}} = S_{\text{AI}}$
    end if
end procedure
procedure UPDATEMODEL
    if user feedback collected for n activities **then**
        Incrementally retrain autoencoder with feedback
    end if
end procedure
procedure CONTINUOUSLEARNING
    Update user's cumulative risk profile
    if cumulative risk $>$ threshold **then**
        Trigger alert for investigation
        if Security Expert rejects the alert decision **then**
            Re-evaluate using INCORPORATEUSERFEEDBACK
            Retrain using UPDATEMODEL
        end if
    end if
end procedure

3.2.4 Policy-Based Risk Analysis & Response Automation

The system enforces predefined security policies for various cloud storage and collaboration platforms like SharePoint, OneDrive, Google Drive, Teams, and Box. These policies act as the first line of defense by identifying and mitigating risks in real-time. When a security violation is detected, an alert is generated, ensuring that unauthorized or risky actions do not go unnoticed. These policies are categorized based on risk dimensions, providing comprehensive coverage across users, data, activity, and attack paths [22, 23].

1. User Risk Policies These policies detect anomalous user behaviors, including unauthorized access attempts, login anomalies, and potential credential misuse. See Table 3

Table 3: Policy-Based Detection Triggers and Associated Risks

Policy Name	Trigger Condition	Potential Risk
Suspicious Login Activity	Multiple failed login attempts within a short time	Brute-force attack, credential stuffing
Unusual Location Access	Login from a new/unapproved geographic region	Compromised credentials, unauthorized access
Privileged Escalation Without Authorization	Sudden admin privilege grant	Insider threat, unauthorized control gain
Login from Untrusted Device	Access from a device not previously associated with the user	Stolen credentials, unauthorized access

2. Data Movement Risk Policies
   These policies monitor unauthorized data transfers across cloud platforms to prevent data ex-filtration and information leaks as mention in Table 4.

Table 4: Policy Triggers and Associated Risks

Policy Name	Trigger Condition	Potential Risk
Mass Data Download	Unusually high file downloads from OneDrive/Google Drive	Insider threat, data theft
External Sharing of Sensitive Data	File shared with an external, untrusted email domain	Data leakage, regulatory non-compliance
Unapproved Cloud Sync	Data synced to unauthorized third-party storage	Shadow IT, unauthorized data transfer
Unencrypted File Transfers	PII/PHI transferred without encryption	Compliance violations, data breach risk

3. Attack Path Risk Policies
   Policies in Table5 utilize knowledge graphs to map potential attack paths within an organization's infrastructure, identifying weak points before exploitation.

Table 5: Advanced Risk Policy Triggers and Potential Threats

Policy Name	Trigger Condition	Potential Risk
Open Attack Paths via Misconfigured Access Controls	The user has excessive privileges in multiple systems	Lateral movement, privilege escalation
Unpatched Vulnerability Exploitation	System/service running outdated software	Exploitable attack surface
Multiple High-Risk Access Points	High-privilege user accesses multiple high-risk resources	Advanced persistent threat (APT) behavior

4. Activity Risk Policies
   These policies detect anomalies in user behavior by identifying suspicious logins, device access, and location-based risk factors Table6

Table 6: Behavioral Anomaly Detection Policies

Policy Name	Trigger Condition	Potential Risk
Excessive Login Failures	Multiple failed login attempts from different devices	Credential stuffing, brute-force attacks
Simultaneous Logins from Multiple Locations	Users log in from geographically distant locations within a short timeframe	Account compromise
Unusual Device Access	A user accesses corporate systems from an unknown or unauthorized device	Stolen credentials, unauthorized access

5. Data Risk Policies

   These policies protect sensitive data assets by enforcing access restrictions, retention policies, and storage security.See Table 7

Table 7: Data Security and Compliance Policy Triggers

Policy Name	**Trigger Condition**	**Potential Risk**
Unauthorized Access to Sensitive Data	User attempts to access restricted PII/PHI	Insider threat, regulatory violation
Mass Deletion of Critical Files	Bulk deletion of critical business files	Accidental data loss, ransomware attack
Storage Policy Violation	Sensitive data stored in an unapproved location	Non-compliance, increased breach risk

6. Data Collaboration Risk Policies

   These policies focus on ensuring secure collaboration by preventing unauthorized document sharing and controlling information flow across teams and external entities. See Table 8

Table 8: Data Collaboration Risk Policies

Policy Name	**Trigger Condition**	**Potential Risk**
Sharing of Confidential Documents with External Parties	Sensitive documents shared outside the organization	Data leakage, regulatory non-compliance
Public File Link Creation for Internal Data	A confidential document is shared via a public link	Unauthorized access, information leaks
Abnormal Collaboration Behavior	A user shares a high number of files in a short period	Insider risk, data exfiltration

7. Behavior-Based Anomalies

   Rather than focusing on specific actions, these policies detect **unusual behavioral patterns** by analyzing deviations from normal usage. Examples include:

   - **Excessive File Downloads:** A user downloads an unusually large volume of files from SharePoint, deviating from their historical access pattern.
   - **Unusual Login Times:** A user in Teams logs in at 3 AM despite never accessing the platform outside business hours.
   - **Abnormal File Modifications:** A script or bot suddenly renames and moves thousands of files in Google Drive, resembling a ransomware attack.

   Each cloud platform has dedicated security policies based on its architecture and use cases as mention in Table 9

Table 9: Cloud Platform Behavior-Based Anomaly Policies

Platform	**Security Policy Trigger**	**Potential Risk**
SharePoint	Abnormal number of files updated	Unauthorized bulk edits, ransomware
OneDrive	Access from non-compliant devices	Data leaks, unauthorized access
Google Drive	Mass file deletion event	Insider threat, accidental data loss
Teams	Suspicious external file sharing	Data exfiltration
Box	Bulk unauthorized downloads	Intellectual property theft

3.2.5 AI-Driven Security Recommendations

The system integrates an on-premises AI model powered by DeepSeek Large Language Models (LLMs) to enhance security operations. Unlike traditional static alerts, this AI provides **context-aware security recommendations** based on real-time events **without triggering automated remediation actions** [24].

1. **Context-Aware Risk Analysis**
   Instead of relying solely on predefined thresholds, the AI analyzes contextual factors to generate actionable insights for security teams. It evaluates:
   - User's historical behavior – Has the user previously performed similar actions, such as large-scale data downloads?
   - Organizational norms – Is this behavior typical for their department or role?
   - Environmental context – Did the activity originate from a recognized, trusted device or network?

   By considering these parameters, the **AI generates an informed security recommendation**, helping analysts understand the risk **in context** rather than reacting to isolated events [25].
2. **AI-Generated Security Recommendations**
   Our system leverages an on-premises **DeepSeek LLM to provide real-time security insights without taking direct remediation actions**. When a high-risk score is assigned, a policy is violated, or a high-risk activity occurs, the AI dynamically analyzes the event and generates contextual recommendations.
   Instead of relying on static thresholds, **the AI follows a chain-of-thought reasoning** process to assess:
3. **Why On-Prem AI?**
   The AI model runs entirely on-premises to maintain data sovereignty and privacy, ensuring no sensitive information leaves the system [26]. This closed-loop security approach prevents exposure to external cloud services while enabling real-time, context-rich security recommendations.
   The AI follows a chain-of-thought reasoning process, analyzing past activity patterns, organizational context, and real-time security events to provide accurate, actionable insights. This empowers security teams to make informed decisions rather than rely on rigid automation.

3.2.6 Visual Analytics and Risk-Based Dashboards

Following **Policy-Based Risk Analysis & Response Automation**, the system transitions into a visual analytics phase. This phase extracts and presents key security insights in a structured and interactive format to facilitate rapid decision-making. The **risk scoring system** fuels these visualizations and drives the intuitive dashboards, alerts, and graphs.

1. Urgent Tab This section prioritizes **high-risk incidents** and supports rapid triage across multiple dimensions:
   - **Campaign Level:** Identifies security threats associated with specific risk campaigns.
   - **User Level:** Highlights high-risk users needing immediate investigation.
   - **Data Level:** Flags sensitive data breaches or policy violations.
   - **Application Level:** Monitors anomalies within specific enterprise applications.
2. Overview Tab Provides a **holistic view of system security**, summarizing high-level risk metrics:
   - **User Insights:** Categorizes users based on behavioral and contextual risk scores.
   - **Campaign Insights:** Visualizes ongoing and completed risk detection campaigns.
   - **Alert Distribution:** Displays alerts categorized by severity.
   - **Risk Factors:** Summarizes data sensitivity, cross-platform vulnerabilities, and recent security trends.
3. Analytics Page Offers **real-time security insights** through advanced data visualizations:
   - **Incident & Risk Activities Graph:** Plots risk-based activity trends over time.
   - **Risk Activity Analysis:** Detects behavioral anomalies and potential insider threats.
   - **Data Breach Prevention Insights:** Highlights data leak vectors and sensitive information flow.
4. Campaign Page Enables creation, tracking, and evaluation of risk mitigation campaigns:
   - **Overall Campaign Performance:** Measures detection efficacy and resolution time.
   - **Individual Campaign Insights:** Tracks user engagement and policy violations per campaign.
5. Users Page Delivers **user-specific risk intelligence** across individual and organizational levels:

- Detailed user profiles, risk scores, and behavioral history.
- Segmented views for user grouping and role-based investigation.

6. Notification Page Aggregates **real-time alerts** to maintain a proactive security posture:

- Displays ongoing incidents and policy violations.
- Enables security teams to respond and resolve threats swiftly.

The entire visualization framework is **powered by the risk scoring system**, ensuring continuous, data-driven insights and enhancing threat detection and response capabilities.

4 Evaluation and Results

Our system's evaluation demonstrates significant improvements in risk detection accuracy, response efficiency, and overall security management. The results include key performance metrics, graphical representations, and comparative analysis.

4.1 PRISM – Privilege-based Risk & Insider Scoring Mechanism

Our initial risk-scoring model assessed security risks using predefined rules and static thresholds. While effective at detecting common threats, it had notable limitations, including a high false-positive rate and a lack of adaptability to evolving risk patterns. As shown in Figure 3, our AI-based risk-scoring model significantly outperforms the PRISM approach. The false positive rate has been reduced from 42% to 17%, leading to 2.5x fewer false alerts. The positive detection rate has increased from 65% to 85%, making the model 1.3x more effective at identifying real threats. The false negative rate has also dropped from 18% to 12%, reducing missed risks by 1.5x. The table 10 provides a direct comparison of PRISM vs. AI-based scoring, along with ratio-based improvements:

Table 10: Performance Comparison: PRISM Scoring vs. AI-Based Scoring

Metric	**PRISM Scoring**	**AI-Based Scoring**	**Improvement / Ratio**
False Positive Rate	42%	17%	59% reduction 2.5x lower (AI reduces false positives)
True Positive Detection Rate	65%	85%	30% increase 1.3x higher (AI detects more threats)
False Negative Rate	18%	12%	33% reduction 1.5x lower (AI reduces missed risks)

We trained our model with user feedback over 12 weeks to achieve these results, incorporating approximately 300 training instances in the initial three weeks. Our custom model was continuously evaluated on a dataset annotated by field experts and the administrator, ensuring high accuracy and relevance. Over time, the model kept improving, adapting to user feedback and administrator preferences.

Furthermore, the AI model continuously improves with additional training data, adapting to organization-specific risk patterns. Over time, it will learn from administrator preferences, allowing for a customized risk-scoring approach that aligns with the organization's unique security needs.

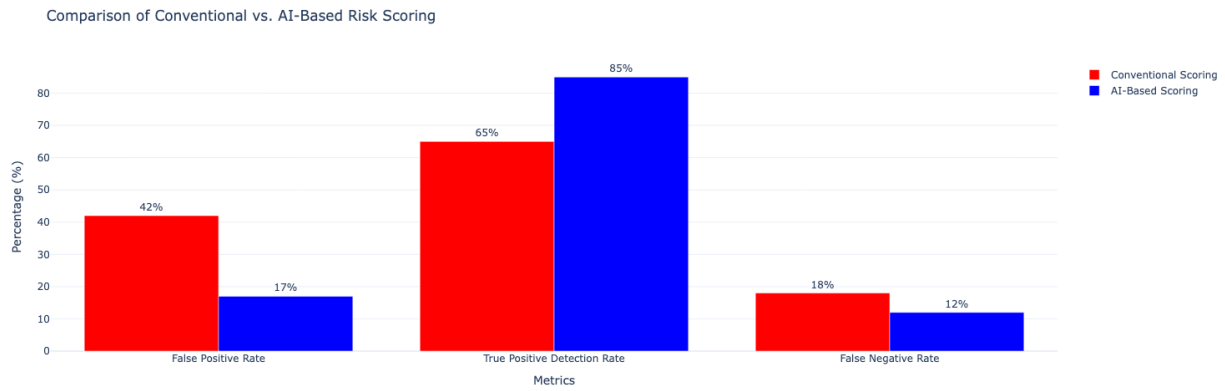

Figure 3: False Positive Rate in PRISM vs. AI-Based Risk Scoring

4.2 Accuracy of AI-Based Risk Scoring Model

Risk detection improved significantly by introducing our AI-powered risk scoring model powered by an autoencoder neural network. After continuous learning from user feedback, the model adjusted dynamically, achieving a 17% false positive rate—less than half of the rule-based system. See figure4 and Table 11.

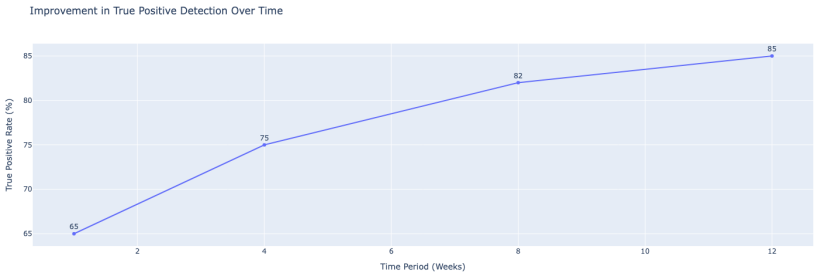

Figure 4: ***Improvement in True Positive Detection Over Time***

Table 11: Improvement in Actual Positive Rate Over Time

Period (Weeks)	**Actual Positive Rate (%)**
Week 1	65%
Week 4	75%
Week 8	82%
Week 12	85%

4.3 User Feedback & Model Adaptation

Over a 60-day testing period, security analysts manually reviewed and adjusted 12% of AI-generated risk scores to fine-tune the model's accuracy. This iterative feedback process led to a significant improvement in risk classification.

As shown in Figure 5 and the accompanying table12, the initial AI model had a false positive rate of 42% and a false negative rate of 18%. After the first feedback loop, where analyst corrections were incorporated into retraining, false positives dropped to 30% and false negatives to 15%. Following a second feedback loop, the model further improved, reducing false positives to 17% and false negatives to 12%.

This demonstrates the effectiveness of continuous learning—each iteration refines the AI's decision-making, reducing unnecessary alerts while improving real threat detection. As more feedback is incorporated, the model adapts dynamically to organizational risk patterns, ensuring a more accurate and customized risk-scoring system.

Table 12: Model Improvement Through Feedback Loops

Iteration	**False Positives (%)**	**False Negatives (%)**
Initial Model	42%	18%
After First Feedback Loop	30%	15%
After Second Feedback Loop	17%	12%

Figure 5: Reduction in False Positives After User Feedback Loops

4.4 Policy Violation Identification

During testing, our policy-based risk analysis engine detected 78 critical violations over a month, covering various security threats such as **unauthorized privilege escalations, data handling breaches, and non-compliant device connections.**

As shown in **Figure 6** and the accompanying table13, the highest number of violations were in data handling, with 31 detected cases, followed by access control violations (22 cases). In response to these risks, automated security measures were applied, including revoking privileges (15 cases), restricting file access (20 cases), and flagging suspicious user activity (10 cases).

The **policy enforcement system** continuously monitors security events and applies real-time mitigation actions, reducing manual intervention and improving overall compliance.

Table 13: Policy Violations and Automated Actions

Policy Category	**Violations Detected**	**Automated Actions Taken**
Access Control Violations	22	15 (Privilege Revoked)
Data Handling Violations	31	20 (File Access Restricted)
Abnormal File Deletions	15	10 (User Flagged)
Non-Compliant Device Usage	10	7 (Device Disconnected)

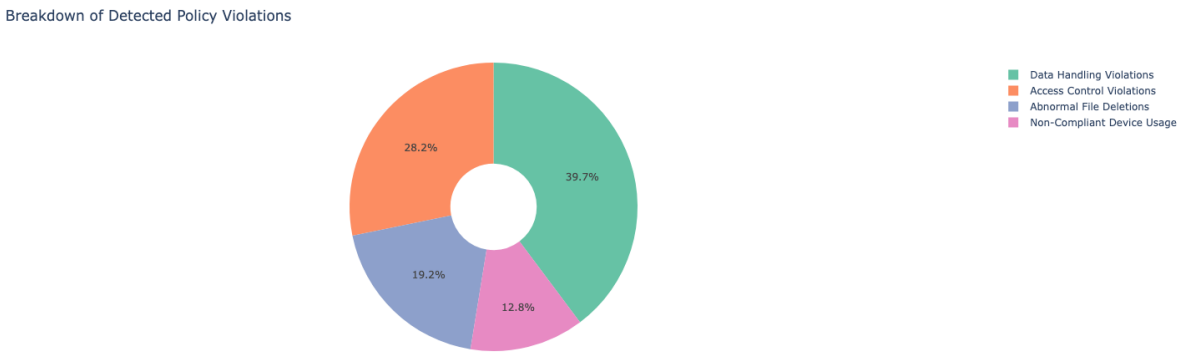

Figure 6: Breakdown of Detected Policy Violations

4.5 Performance & Scalability

The system demonstrated exceptional performance under high log ingestion loads, efficiently processing up to **10 million log events daily** while maintaining a **sub-300ms query response time**. Performance tests across different ingestion rates, as shown in **Table 1**, revealed that at **1,000 logs per second**, the system responded within **120ms**, while at **10,000 logs per second**, response time increased slightly to **190ms**. Even under an extreme load of 100,000 logs per second, the system maintained a **250ms** response time, ensuring real-time threat detection and analysis.

Table 14: System Performance: Log Ingestion vs. Query Response Time

Log Ingestion Rate (Logs/sec)	Query Response Time (ms)
1,000	120
10,000	190
100,000	250

These results indicate the system's **high scalability**, making it suitable for organizations of all sizes, from SMBs to large enterprises with vast security logs. The **low-latency query execution** enables security analysts to retrieve insights instantly, facilitating **faster threat detection and mitigation**. Additionally, as seen in **Table 14**, the system maintains optimal performance even as log ingestion rates increase, proving its **efficiency under heavy workloads**. The architecture also supports future scalability, as further optimizations, such as **advanced indexing strategies and parallel processing**, can enhance performance even under higher loads. This ensures that security teams can respond promptly to emerging threats, improving overall **incident response time and cyber risk management.**

4.6 Comprehensive Impact of Insider Risk Management (IRM) System

Implementing the Insider Risk Management (IRM) system has significantly strengthened security operations by integrating AI-driven risk scoring, policy violation detection, and sensitive data classification. One of the most impactful outcomes has been a 47% reduction in incident response time, primarily driven by automated risk assessments, real-time policy enforcement, and AI-assisted decision-making. The system enhances efficiency through automated policy violation detection, instantly flagging unauthorized actions and minimizing investigation delays. Additionally, risk-based prioritization ensures that security teams focus on the most critical threats first, optimizing resource allocation. Integrating sensitive data classification allows for context-aware alerts, improving the accuracy of risk assessments and reducing false positives. Furthermore, a streamlined investigation workflow provides security teams with pre-analyzed insights, reducing the manual effort required to correlate security events. These enhancements have led to a significant improvement in incident response efficiency, as demonstrated in Table 15. The average resolution time decreased from 45 minutes (manual investigation) to 24 minutes with IRM-assisted response. This reduction is further visualized in Figure 7, which presents a comparative bar chart highlighting the efficiency gains achieved through AI-driven automation.

Table 15: Incident Response Efficiency – Before vs. After IRM Implementation

Response Method	Average Time to Resolution
Manual Investigation	45 minutes
IRM-Assisted Response	24 minutes

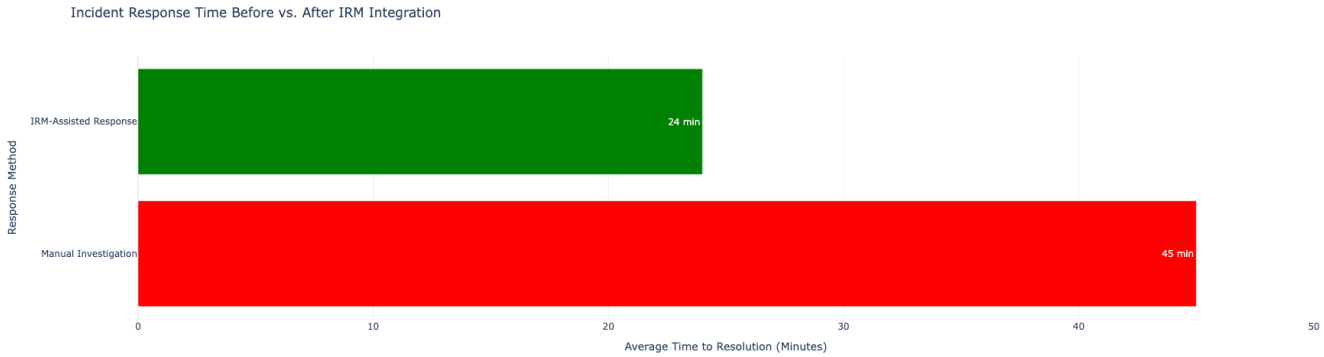

Figure 7: Incident Response Time Before vs. After IRM Integration

5 Conclusion and Future Work

This research has demonstrated the significant potential of AI-powered Insider Risk Management (IRM) systems to enhance organizational security by leveraging advanced machine learning models, behavioral analytics, and real-time data monitoring. The proposed framework strengthens the detection accuracy and response speed by fusing diverse data sources and performing contextual risk analysis, effectively reducing the exposure window to insider threats. Major contributions include the development of an AI-driven risk assessment model, deployment of behavioral baselining techniques, and validation of the system's effectiveness in operational environments, leading to reduced false positives, improved incident response times, and scalable deployment across both on-premises and cloud infrastructures. Moving forward, several enhancements are envisioned: implementing federated learning for privacy-preserving AI training, integrating explainable AI techniques for improved transparency, aligning the IRM system with Zero Trust security frameworks, enriching behavioral anomaly detection through graph-based analysis, enabling real-time response mechanisms, expanding cross-platform compatibility for hybrid environments, and embedding automated compliance audits to adapt to evolving regulations. By pursuing these directions, the IRM system can become an even more powerful, adaptable, and comprehensive solution, equipping organizations with the necessary tools to proactively mitigate emerging insider threats and protect critical assets.

References

[1] 2024 Data Breach Investigations Report (DBIR). Technical report, Verizon Enterprise Solutions, 2024. https://www.verizon.com/business/resources/reports/2024-dbir-data-breach-investigations-report.pdf

[2] Ponemon Institute. "2022 Cost of Insider Threats Global Report." Ponemon Institute, 2022. https://protectera.com.au/wp-content/uploads/2022/03/The-Cost-of-Insider-Threats-2022-Global-Report.pdf

[3] Malik, Shoaib. "Using AI for Behavioral Analytics in Cybersecurity: Detecting Anomalies and Insider Threats." (2024).

[4] NIST. "Privacy Framework: A Tool for Improving Privacy through Enterprise Risk Management." 2020.

[5] Microsoft. Azure Active Directory Documentation. https://learn.microsoft.com/en-us/azure/active-directory/, 2024.

[6] Amazon Web Services. Identity and Access Management (IAM). https://docs.aws.amazon.com/IAM/latest/UserGuide/, 2024.

[7] Google. Google Workspace Admin Help. https://support.google.com/a/, 2024.

[8] OpenAI. GPT-4 Technical Report. https://openai.com/research/gpt-4, 2023.

[9] Gartner. Market Guide for Insider Risk Management Solutions. https://www.gartner.com/en/documents/4931631 Gartner Research, 2023.

[10] Force, Joint Task, and Transformation Initiative. "Security and privacy controls for federal information systems and organizations." *NIST Special Publication* 800, no. 53 (2013): 8-13. U.S. Department of Commerce, 2020.

[11] Verizon. 2023 Data Breach Investigations Report. Technical report, Verizon Enterprise Solutions, 2023. https://inquest.net/wp-content/uploads/2023-data-breach-investigations-report-dbir.pdf

[12] Gartner. Gartner Identifies the Top Cybersecurity Trends for 2023. https://www.gartner.com/en/newsroom/press-releases/04-12-2023-gartner-identifies-the-top-cybersecurity-trends-for-2023 Gartner Research, 2023.

[13] , Glasser, Joshua and Lindauer, Brian Bridging the gap: A pragmatic approach to generating insider threat data. In 2013 IEEE Security and Privacy Workshops, pp. 98-104. IEEE, 2013.

[14] Carnegie Mellon University - Software Engineering Institute (CMU-SEI). CERT Insider Threat Tools and Datasets. *Technical Report*, 2016. Available at: https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=508099

[15] Salem, Mohammad B., Hershkop, Shlomo and Stolfo, Salvatore J. A survey of insider attack detection research. *Insider Attack and Cyber Security*, pp. 69–90, Springer, 2008.

[16] Greitzer, Frank L. and Frincke, Deborah A. Combining traditional cyber security audit data with psychosocial data: Towards predictive modeling for insider threat mitigation. *Insider Threats in Cyber Security*, pp. 85–113, Springer, 2010.

[17] Koli Lokesh and Kalra Shubham and Singh Karanpreet Context-Aware Hybrid Pattern Detection Algorithm (CHPDA) for Real-Time Sensitivity Analysis. *VectorEdge Technical Whitepaper*, 2025.

[18] Liu, S., Wang, W., and Liu, D. Risk assessment model based on user behavior for insider threats. *Computers & Security*, 77, 428–440, 2018.

[19] Greitzer, F. L., & Hohimer, R. E. Modeling human behavior to anticipate insider attacks *Journal of Strategic Security 4, no. 2 (2011): 25-48.*

[20] Pantelidis, Efthimios, Bendiab, Gueltoum, Shiaeles, Stavros, and Kolokotronis, Nicholas. Insider Detection using Deep Autoencoder and Variational Autoencoder Neural Networks. *arXiv preprint arXiv:2109.02568*, 2021.

[21] Bentley, Rick and Sarkar, Abhik. Humans in AI: The necessity for human-in-the-loop (HILT). *Security Magazine*, 2024. https://www.securitymagazine.com/blogs/14-security-blog/post/100798-humans-in-ai-the-necessity-for-human-in-the-loop-hilt

[22] Microsoft Learn. *Data Protection Policies in Microsoft Defender for Cloud Apps*. 2023. Accessed: 2025-04-28. https://learn.microsoft.com/en-us/defender-cloud-apps/data-protection-policies

[23] Uptycs. *Mastering Cloud Security: Understanding the Attack Path*. 2023. Accessed: 2025-04-28. https://www.uptycs.com/blog/mastering-cloud-security/attack-path

[24] Team Wrixte. Elevating Threat Intelligence: Integrating Context-Aware AI Models for Real-Time Cyber Defense. *Wrixte*, May 15, 2024. https://wrixte.co/2024/05/15/elevating-threat-intelligence-integrating-context-aware-ai-models-for-real-time-cyber-defense/

[25] TATA Consultancy Services. Building Context-Aware Cybersecurity Alerts. *TCS Insights*, January 10, 2025. https://www.tcs.com/insights/topics/cybersecurity-topic/article/context-aware-cybersecurity

[26] Hiroshi Kinoshita. AI Risk Management: Introducing the On-Premises Version of DeepSeek to Keep Your Data In-House. *Medium*, March 2025. https://medium.com/@ai2ai/ai-risk-management-introducing-the-on-premises-version-of-deepseek-to-keep-your-data-in-house-a-a71