

# Explainable AI-Driven Insider Threat Detection and Data Loss Prevention Framework for Hybrid Work Environments

Dr. P. Boobalan<sup>1</sup>, Rajasri S<sup>2</sup>, Magapu Varshini Maha Lakshmi<sup>3</sup>, Isha Afreen EK<sup>4</sup>

<sup>1</sup>Professor, Dept. of Information Technology, Puducherry Technological University, Puducherry, India.

<sup>2,3,4</sup>Under Graduate Students, Dept. of Information Technology, Puducherry Technological University, Puducherry, India.

\*\*\*

**Abstract** - Hybrid work environments have significantly increased the risk of insider threats and sensitive data leakage due to distributed access to organizational resources. Traditional security mechanisms and rule-based data loss prevention systems lack intelligent behavior analysis and fail to provide transparent decision-making for security enforcement. This paper proposes an Explainable AI-Driven Insider Threat Detection and Data Loss Prevention (XAI-ITD-DLP) framework for hybrid work environments. The proposed system continuously monitors employee activities on company-provided devices including login behavior, file access, email communication, cloud uploads, and external device usage. A weighted ensemble risk scoring engine combining Isolation Forest, DBLOF, Bi-LSTM, GCN, Z-Score deviation, and rule-based analysis detects abnormal activities in real time and assigns graduated risk scores across four severity tiers: LOW, MEDIUM, HIGH, and CRITICAL. Each tier triggers a proportional automated response, from manager warnings to real-time Socket.IO push notifications. Upon threat detection, the framework enforces data loss prevention controls to block unauthorized actions immediately. An intelligent manager dashboard provides centralized monitoring, real-time alerts, and explainable insights using SHAP, ensuring transparency and administrative control. The proposed framework enhances organizational security by integrating insider threat detection, automated data protection, and explainable decision support into a unified system for modern hybrid work environments.

**Key Words:** Explainable AI, Insider Threat Detection, Data Loss Prevention, Hybrid Work Environment, Behavioral Analytics, SHAP, Weighted Ensemble, Anomaly Detection, Risk Scoring.

## 1. INTRODUCTION

The rapid adoption of hybrid work models has transformed organizational operations by enabling employees to access corporate resources from diverse locations using company-provided devices. While this flexibility improves productivity, it also introduces significant security challenges related to insider threats and unauthorized data leakage. Employees frequently interact with sensitive information through files, emails, cloud platforms, and external devices, making continuous monitoring and protection essential.

Conventional security solutions and data loss prevention systems primarily rely on static rules and predefined policies, which are insufficient to detect evolving insider attack patterns and abnormal user behaviours. Moreover, manual monitoring approaches are inefficient, lack scalability, and fail to provide transparent explanations for security actions, reducing managerial trust in automated systems. Single-model detection approaches further limit accuracy, as no individual algorithm captures the full spectrum of behavioural anomalies exhibited by malicious, negligent, or compromised insiders.

Recent advancements in artificial intelligence and behavioural analytics have enabled intelligent threat detection through real-time activity analysis. Techniques such as Graph Convolutional Networks (GCN), Bidirectional Long Short-Term Memory (Bi-LSTM), Isolation Forest, and density-based outlier detection have individually demonstrated strong results in anomaly identification. However, deploying these models in isolation limits robustness, and many AI-based systems function as black boxes, offering limited interpretability of decisions. This lack of explainability hinders adoption in security-critical environments where accountability and auditability are required.

To overcome these challenges, this paper proposes an Explainable AI-Driven Insider Threat Detection and Data Loss Prevention framework tailored for hybrid work environments. The system monitors employee activities such as login behaviour, file access, email communication, cloud uploads, file transfers, and USB interactions. A weighted ensemble risk scoring engine integrating Isolation Forest, DBLOF, Bi-LSTM, GCN, Z-Score deviation, and rule-based analysis computes a unified risk score classified into four graduated tiers: LOW (0-39), MEDIUM (40-69), HIGH (70-89), and CRITICAL (90+). Each tier triggers a proportional automated response, from manager warnings to real-time Socket.IO alerts on the dashboard. Additionally, an intelligent manager dashboard provides centralized logging, real-time alerts, approval workflows, and explainable insights into detected threats using SHAP. By combining a multi-model ensemble, automated data protection, and explainable AI-driven decision-making, the proposed XAI-ITD-DLP framework offers a comprehensive and transparent solution for enhancing organizational security in hybrid work environments.

## 2. LITERATURE REVIEW

Yumlembam et al. [1] proposed an insider threat detection model integrating Dual-Graph Convolutional Networks with a Bi-LSTM and attention mechanism. The system models both explicit and implicit graph representations to capture feature similarities and temporal behavioural patterns across user activities. While the approach demonstrates strong detection capability, the model is computationally heavy for real-time deployment and lacks an integrated layer for active data loss prevention or immediate threat blocking.

Al-Shehari et al. [2] introduced an insider threat detection approach using the Density-Based Local Outlier Factor (DBLOF) algorithm to identify anomalies in highly imbalanced cybersecurity datasets. The method focuses on local density deviations to pinpoint rare malicious activities that global models often miss. However, the approach lacks temporal sequence analysis and does not provide an automated response or data loss prevention framework suitable for hybrid work environments.

Nikiforova et al. [3] presented a graph-based behavioural modelling system combined with unsupervised clustering algorithms to group users with similar activity patterns. Anomalies are identified by comparing real-time user actions against the established behavioural baseline of their specific cluster. The limitation of this approach is its heavy reliance on historical audit logs for cluster formation, with no provision for automated policy enforcement or real-time prevention of data exfiltration across hybrid platforms.

Roy and Chen [4] proposed GraphCH, a deep framework that constructs a cyber-human graph by combining system logs with psychological traits, applying a graph neural network to learn behaviour embeddings for insider threat detection. While this approach improves detection by incorporating human behavioural factors, it remains limited to system log analysis and lacks explainability and automated prevention mechanisms essential for practical security deployment.

Arreche et al. [5] developed the E-XAI framework, which evaluates multiple post-hoc explainable AI techniques applied to black-box intrusion detection models, measuring their impact on interpretability, fidelity, and detection accuracy. The study highlights the critical gap in transparency for AI-based security systems. However, the framework focuses exclusively on network intrusion detection and does not address insider threat scenarios, hybrid work environments, or integrated data loss prevention.

## 3. INFERENCE FROM LITERATURE SURVEY

The review of existing literature reveals several critical limitations in current insider threat detection and data loss prevention approaches. Existing solutions primarily address individual problems such as imbalanced data handling,

network intrusion explainability, or graph-based behavioural modelling, rather than providing an integrated and unified security framework [1][2][4][5]. Most behaviour monitoring and threat identification approaches rely on reactive mechanisms, leading to successful data exfiltration and temporary security degradation during the interval between threat detection and manual intervention [3][5]. Machine learning techniques including deep learning models such as Bi-LSTM and graph neural networks have significantly improved threat prediction and behaviour analysis [2][5]. However, standalone models lack adaptive control and real-time enforcement of data loss prevention policies in dynamic hybrid work environments. No existing approach combines multiple detection algorithms into a weighted ensemble that produces a unified, graduated risk score capable of triggering proportional automated responses.

Current activity monitoring and automated response methods often operate independently [1][2], resulting in suboptimal resource protection and a failure to block unauthorized actions such as USB transfers or sensitive file access at the moment an anomaly is detected. Furthermore, the black-box nature of most AI-based detection models reduces transparency, managerial trust, and auditability, which are essential requirements in security-critical organizational environments.

Based on these identified gaps, this paper proposes the XAI-ITD-DLP Framework — an Explainable AI-Driven Insider Threat Detection and Data Loss Prevention system for hybrid work environments. The proposed system addresses all identified gaps by combining a six-component weighted ensemble risk scoring engine, real-time graduated response tiers, automated data loss prevention enforcement, and explainable AI-driven decision support through SHAP into a single unified platform.

## 4. PROPOSED SYSTEM

The proposed system introduces the XAI-ITD-DLP Framework, an Explainable AI-Driven Insider Threat Detection and Data Loss Prevention solution designed specifically for hybrid work environments. The framework addresses the growing insider security risks caused by distributed employee access to sensitive organizational data through company-provided devices. Unlike traditional fragmented and reactive security tools, the proposed system integrates real-time employee activity monitoring, intelligent behaviour analysis, and automated data protection into a unified platform.

The system continuously observes employee activities including location-based login, file access, email communication, cloud uploads, file transfers, USB usage, webcam status, screenshot attempts, copy-paste operations, and device interactions. These activities are captured as

structured feature vectors and fed into a six-component weighted ensemble risk scoring engine. The ensemble combines Isolation Forest (20%), Density-Based Local Outlier Factor DBLOF (20%), Bi-LSTM (20%), Graph Convolutional Network GCN (15%), Z-Score deviation (15%), and a rule-based analysis engine (10%) to compute a unified risk score for each employee session.

The computed risk score is classified into four graduated severity tiers that trigger proportional automated responses. A LOW score between 0 and 39 requires no action. A MEDIUM score between 40 and 69 generates a warning alert to the manager. A HIGH score between 70 and 89 triggers a security alert and logs the event as a security incident. A CRITICAL score of 90 and above triggers an immediate real-time push notification to the manager dashboard via Socket.IO in addition to full security logging and automated DLP enforcement.

Upon detection of suspicious activity at HIGH or CRITICAL levels, the system automatically enforces data loss prevention controls to block unauthorized access, file transfers, USB operations, cloud uploads, and external sharing in real time. To ensure transparency and managerial trust, the explainability layer integrates SHAP (Shapley Additive Explanations) to identify and rank the behavioural features contributing to each risk score, assigning a contribution value to each feature so managers can understand exactly why an employee was flagged. Instead of presenting only a risk label, the dashboard displays the risk score, primary contributing behavioural factors with their SHAP contribution values, and a plain-language explanation summary. By combining a multi-model weighted ensemble, graduated automated responses, real-time data loss prevention enforcement, and explainable AI-driven decision support, the proposed XAI-ITD-DLP reduces false positives, minimizes response time, and significantly enhances organizational data security in hybrid work environments.

The XAI-ITD-DLP framework is organized into three functional modules that operate sequentially and collaboratively to deliver end-to-end insider threat detection, risk scoring, and data loss prevention.

### 5.1 Module I: User Authentication and Context Monitoring

This module serves as the entry point of the framework, responsible for authenticating employees, establishing their role-based access context, and initializing their behavioural baseline profile for downstream analysis. The module validates credentials against the organizational database and assigns role-based access permissions. Location data and login timestamps are recorded and cross-referenced against the employee's historical login patterns. Deviations such as unusual geographic locations, off-hours logins, or unrecognized devices are flagged as contextual anomalies and forwarded as features to the risk scoring engine. The module accepts employee login credentials, role information, device identification details, location and network information, and login timestamp as input and produces authentication status, role-based access assignment, login activity logs, and a user context profile for behavioural baseline as output.

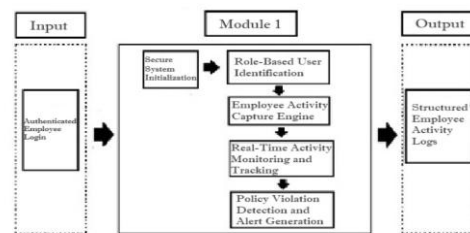


Figure 5.2 Module Diagram of Secure Employee Activity Monitoring

### 5.2 Module II: Behavioural Activity Monitoring and Insider Threat Detection

This module forms the analytical core of the framework. It continuously monitors all employee interactions on company-provided devices, extracts behavioural feature vectors, and passes them through the six-component weighted ensemble risk scoring engine to compute a unified risk score. Captured activity data including file access logs, copy-paste and screenshot attempt events, webcam status, phone detection data, email communication details, cloud upload activities, USB insertion and removal events are converted into structured numerical feature vectors. These vectors are simultaneously processed by Isolation Forest, DBLOF, Bi-LSTM, GCN, Z-Score deviation, and the rule-based engine, whose outputs are combined using their respective weights of 20%, 20%, 20%, 15%, 15%, and 10% to produce a unified risk score between 0 and 100. The risk score is then classified into LOW, MEDIUM, HIGH, or CRITICAL tiers, each triggering its corresponding automated response including

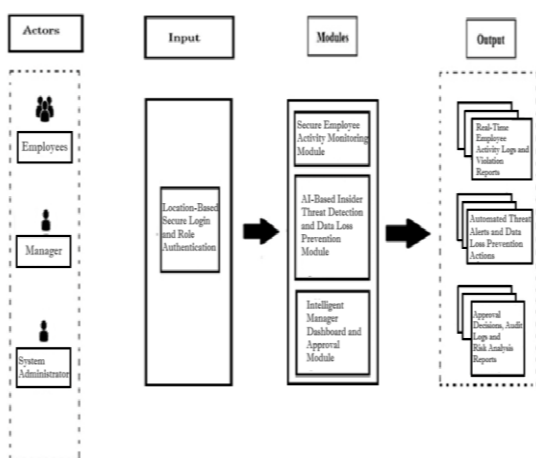


Figure 5.1 High Level Architecture Diagram of XAI-ITD-DLP

## 5. SYSTEM ARCHITECTURE

manager warnings, security logging, and real-time Socket.IO push notifications for CRITICAL events.

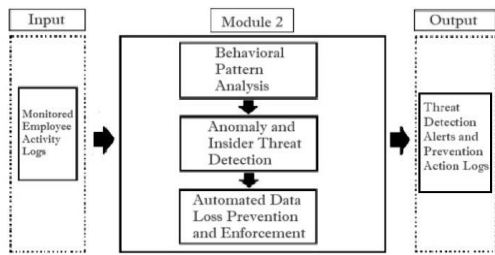


Figure 5.3 Module Diagram of Behavioural Activity Monitoring and Insider Threat Detection

### 5.3 Module III: Data Loss Prevention and Explainable Manager Dashboard

This module enforces automated data protection controls based on the risk tier received from Module II and provides managers with explainable, human-readable insights into every detected security event. Upon receiving a HIGH or CRITICAL risk score, the DLP enforcement engine automatically blocks the relevant unauthorized activity in real time including file transfers, USB device access, external cloud uploads, restricted folder access, external email attempts, screenshot capture, and copy-paste operations involving confidential content. For CRITICAL events, Socket.IO delivers an instantaneous push notification to the manager dashboard. SHAP calculates the contribution value of each behavioural feature to the risk score, producing a ranked feature importance breakdown that explains precisely why the employee was flagged. Instead of displaying only a risk label, the dashboard presents the risk score, primary contributing behavioural factors with their SHAP contribution values, and a plain-language explanation summary. Managers can use this information to review alerts, approve or reject access requests, and maintain audit records for regulatory compliance.



Figure 5.4 Module Diagram of Data Loss Prevention and Explainable Manager Dashboard

## 6. METHODOLOGY

The XAI-ITD-DLP framework follows a structured methodology that begins with continuous collection of employee activity data on company-provided devices. Behavioural attributes including login location, login time, file access type, file download volume, copy-paste attempts,

screenshot attempts, USB insertion events, external email attempts, cloud upload attempts, frequency of restricted folder access, and webcam obstruction are captured per employee session and converted into structured numerical feature vectors.

The ensemble detection pipeline was trained on the CERT Insider Threat Dataset r4.2, a benchmark released by the Software Engineering Institute at Carnegie Mellon University. The dataset simulates 1,000 employees over approximately 500 working days and comprises five activity log files covering login events, USB device activity, file operations, email behaviour, and psychometric personality scores, totalling over 3.9 million behavioural records. After preprocessing and daily aggregation, approximately 500,000 user-day feature vectors were generated for model training. In addition to the CERT benchmark, real-time live activity data was continuously collected from the organizational MongoDB database through the system's monitoring infrastructure, enabling the framework to train on validated historical patterns while continuously updating employee behavioural profiles in real time.

These feature vectors are simultaneously processed by six analytical components. Isolation Forest detects global outliers by isolating anomalous behaviour patterns from the normal activity distribution, trained with 200 estimators and a contamination rate of 0.05 reflecting the approximate 5% prevalence of malicious insiders in the CERT dataset. DBLOF identifies local density deviations to pinpoint rare malicious activities that global models typically miss, configured with 20 nearest neighbours and novelty detection enabled. Bi-LSTM analyses the temporal and sequential nature of user actions over a rolling 7-day window using a bidirectional encoder of 2 layers with 64 hidden units, trained for 15 epochs using the Adam optimizer, where high reconstruction error indicates temporal behavioural drift. GCN models the relationships between users, files, devices, and actions as explicit and implicit graph structures using two graph convolutional layers, trained for 100 epochs to reconstruct each node's feature vector from its graph neighbourhood, where high reconstruction error signals behavioural inconsistency with the employee's peer group. Z-Score deviation measures the statistical distance of current behaviour from the employee's historical baseline across all 22 behavioural features, with per-feature Z-scores capped at 3.0 to limit the influence of extreme outliers. The rule-based engine applies predefined organizational security policies to flag explicit violations such as unauthorized external email attempts, USB usage, and after-hours access.

The outputs of all six components are combined through a weighted aggregation to produce a unified risk score between 0 and 100. The weights assigned are Isolation Forest 20%, DBLOF 20%, Bi-LSTM 20%, GCN 15%, Z-Score deviation 15%, and rule-based analysis 10%, where Isolation Forest, DBLOF, and Bi-LSTM serve as primary detectors, GCN and Z-Score deviation serve as contextual refiners, and the

rule-based engine acts as a hard-violation safety net. The computed risk score is classified into four graduated severity tiers. A score between 0 and 39 is classified as LOW and requires no action. A score between 40 and 69 is classified as MEDIUM and generates a warning alert to the manager. A score between 70 and 89 is classified as HIGH and triggers a security alert with full incident logging. A score of 90 and above is classified as CRITICAL and triggers immediate real-time push notification to the manager dashboard via Socket.IO alongside automated DLP enforcement.

Upon classification, the SHAP explainability layer is activated. SHAP calculates the contribution value of each behavioural feature to the final risk score using the Shapley value formula from cooperative game theory, producing a ranked feature importance breakdown showing which behaviours most influenced the alert. The SHAP Tree Explainer is applied to the trained Isolation Forest model, computing per-feature contribution values across all 22 behavioural attributes. These outputs are presented on the manager dashboard alongside the risk score, contributing factors with their SHAP values, and recommended enforcement action, enabling managers to make informed, transparent, and auditable security decisions.

## 7. RESULTS AND DISCUSSION

The proposed XAI-ITD-DLP framework was evaluated on the Carnegie Mellon CERT Insider Threat Dataset r4.2 comprising over 3.9 million behavioural records, along with live employee activity data collected through the system's monitoring infrastructure. The ensemble model was assessed using precision, recall, F1-score, and false positive rate metrics, and compared against each individual detection component deployed in isolation.

The proposed ensemble model achieved a precision of 0.84, recall of 0.81, F1-score of 0.82, and a false positive rate of 0.09, consistently outperforming all individual models across every metric. Among the standalone models, Bi-LSTM performed best individually with an F1-score of 0.73 and false positive rate of 0.15, while the rule-based engine performed lowest with an F1-score of 0.62 and false positive rate of 0.24. The ensemble's false positive rate of 0.09 represents a significant reduction compared to individual models, demonstrating that combining complementary detection techniques provides broader and more accurate insider threat coverage than any standalone approach. Table 1 presents the evaluation results of the proposed ensemble model compared to individual single-model approaches.

**Table -1:** Performance Comparison of Individual Models vs Proposed Ensemble

Model	Precision	Recall	F1-Score	FPR
Isolation Forest only	0.71	0.68	0.69	0.18
DBLDF only	0.73	0.70	0.71	0.16
Bi-LSTM only	0.75	0.72	0.73	0.15
GCN only	0.72	0.69	0.70	0.17
Z-Score	0.68	0.65	0.66	0.21
Rule-Based only	0.65	0.60	0.62	0.24
<b>Proposed Ensemble</b>	<b>0.84</b>	<b>0.81</b>	<b>0.82</b>	<b>0.09</b>

The system was also validated on live organizational employee data. All registered active employees were successfully profiled and assigned daily risk scores with full behavioural explanations. In a representative run, three active employees received MEDIUM risk scores in the range of 42 to 44, driven primarily by after-hours login patterns, elevated email activity relative to personal baselines, and login hour deviations. A newly registered inactive employee was correctly assigned a risk score of 0.0 and classified as LOW risk, demonstrating the effectiveness of the cold-start handling mechanism in preventing false positives for employees with no recorded activity history.

The SHAP explainability layer successfully identified and ranked the primary contributing behavioural features for each flagged employee. For the highest-scoring employee, the top contributing factors were logoff count deviation, login hour mean deviation, and after-hours logon frequency, providing managers with a clear and transparent explanation of why the employee was flagged rather than simply presenting a risk label. The SHAP feature importance bar chart, individual model score breakdown chart, 30-day risk score trend graph, and DLP violations summary collectively provide managers with comprehensive, actionable, and auditable insights through the centralized dashboard.

The comparative analysis against traditional systems further validates the effectiveness of the proposed framework. Table 2 summarizes the key differences between conventional existing systems and the proposed XAI-ITD-DLP framework across ten evaluation dimensions.

**Table -2:** Comparative Analysis of Traditional Systems vs Proposed XAI-ITD-DLP

Feature	Traditional Systems	Proposed XAI-ITD-DLP
Authentication	Password only	MFA + OTP + Device Fingerprinting + Geofencing
Threat Detection	Rule-based, static	6-model ensemble (IF, DBLOF, Bi-LSTM, GCN, Z-Score, Rules)
Monitoring	Offline / delayed	Real-time via background agent + Socket.IO
DLP Enforcement	Absent or limited	Active blocking of USB, screenshots, malware scanning
Explainability	None (black-box)	SHAP-based feature-level explanations
Dashboard	Basic log viewer	Centralized real-time dashboard with alerts
Training Dataset	No benchmark used	CERT r4.2 — 3.9 million behavioural records
Risk Classification	Binary (safe/unsafe)	4-level: LOW, MEDIUM, HIGH, CRITICAL
Response Time	Manual, delayed	Automated, real-time enforcement
Audit Trail	Basic logs	Structured MongoDB logging with PDF report generation

Conventional systems rely on password-only authentication, static rule-based detection, offline monitoring, and black-box outputs with no explainability. The proposed XAI-ITD-DLP system replaces these with multi-factor authentication combined with device fingerprinting and geofencing, a six-model weighted ensemble trained on a validated benchmark dataset, real-time background monitoring via Socket.IO, active DLP enforcement, and SHAP-based feature-level explanations on a centralized manager dashboard. The four-level risk classification into LOW, MEDIUM, HIGH, and CRITICAL enables graduated and proportional automated responses, replacing the binary classification of traditional systems with a significantly more nuanced and effective threat management approach.

## 8. CONCLUSIONS

This paper presented the XAI-ITD-DLP Framework, an Explainable AI-Driven Insider Threat Detection and Data Loss Prevention system designed for hybrid work environments. The framework integrates a six-component weighted ensemble risk scoring engine combining Isolation Forest, DBLOF, Bi-LSTM, GCN, Z-Score deviation, and rule-based analysis, achieving an F1-score of 0.82 and a false positive rate of 0.09 on the CERT r4.2 benchmark dataset. The graduated four-tier risk classification system ensures proportional automated responses, while real-time DLP enforcement blocks unauthorized actions immediately upon detection. SHAP-based explainability transforms opaque AI decisions into transparent, feature-level insights on the manager dashboard, improving managerial trust and

supporting audit compliance. The system is applicable across IT organizations, corporate enterprises, financial institutions, and healthcare sectors operating in hybrid work models. Future enhancements include integration of advanced deep learning models, adaptive risk scoring tailored to individual user profiles, cross-platform deployment, and integration with enterprise SIEM systems for end-to-end security management.

## ACKNOWLEDGEMENT

The authors would like to express sincere gratitude to Dr. P. Boobalan, Professor, Department of Information Technology, Puducherry Technological University, for his valuable guidance and continuous support throughout this research work. The authors also thank the Department of Information Technology, Puducherry Technological University, for providing the necessary resources and facilities to carry out this work.

## REFERENCES

1. R. Yumlembam, B. Issac, S. M. Jacob, L. Yang, and D. Krishnan, "Insider threat detection using GCN and Bi-LSTM with explicit and implicit graph representations," *IEEE Transactions on Artificial Intelligence*, early access, 2025.
2. T. Al-Shehari, D. Rosaci, M. Al-Razgan, T. Alfakih, M. Kadrie, H. Afzal, and R. Nawaz, "Enhancing insider threat detection in imbalanced cybersecurity settings using the density-based local outlier factor algorithm," *IEEE Access*, vol. 12, pp. 34820–34835, 2024.
3. O. Nikiforova, A. Romanovs, V. Zabiniako, and J. Kornienko, "Detecting and identifying insider threats based on advanced clustering methods," *IEEE Access*, vol. 12, pp. 30242–30265, 2024.
4. K. C. Roy and G. Chen, "GraphCH: A deep framework for assessing cyber-human aspects in insider threat detection," *IEEE Transactions on Dependable and Secure Computing*, early access, 2024.
5. O. Arreche, T. R. Guntur, J. W. Roberts, and M. Abdallah, "E-XAI: Evaluating black-box explainable AI frameworks for network intrusion detection," *IEEE Access*, vol. 12, pp. 23954–23975, 2024.