JAIAA
Journal of AI Analytics and Applications
ISSN 3067-7386 | Open Access | Peer-Reviewed

RESEARCH ARTICLE
Vol. 1, No. 1 (2023), pp. 76–94
DOI: 10.63646/jaiaa.2023.010105

Explainable AI Analytics for Pre-Incident Insider Threat Risk Scoring in Information Systems

Lin Haoran[1], Zhao Yuting[2], Qiao Ming[3,*]

[1] School of Cyberspace Security, Hangzhou Dianzi University, Hangzhou 310018, China
[2] School of Management Engineering, Nanjing University of Finance and Economics, Nanjing 210023, China
[3] School of Computer Science and Technology, Shandong University of Technology, Zibo 255000, China
* Corresponding author: qiaoming@sdut.edu.cn

ARTICLE INFO
Received
October 16, 2022
Revised
December 11, 2022
Accepted
February 22, 2023
Available Online
March 30, 2023
DOI
10.63646/jaiaa.2023.010105

Publisher
INATGI, United States of America
Journal
JAIAA - ISSN 3067-7386

Abstract
Insider threats remain difficult to control because the most damaging events often emerge from ordinary access, changing work conditions, weak controls, and behavioral signals that are visible before a confirmed incident occurs. This study develops an explainable artificial intelligence analytics framework for pre-incident insider threat risk scoring in information systems. The framework integrates behavioral, organizational, and technical control indicators into a staged analytics pipeline that combines feature engineering, entropy-informed weighting, supervised learning, local explanation, calibration, and risk-tier governance. Instead of treating insider threat analytics as a black-box detection problem after malicious activity has already occurred, the proposed framework treats risk scoring as an auditable decision-support process for early intervention. A synthetic enterprise dataset is constructed to evaluate the approach across 6,000 user-period observations and 48 observable indicators representing access behavior, work context, policy violations, control exposure, and security-technology gaps. Comparative analysis shows that the explainable hybrid model improves AUC from 0.76 under entropy-only scoring to 0.89, while reducing calibration error to 0.08. Local explanation results identify data export volume, after-hours access, policy violations, managerial pressure, and data-control gaps as the most influential pre-incident signals. The findings demonstrate that explainable AI can increase model transparency, support proportionate governance actions, and improve the business usability of insider risk analytics without relying on intrusive surveillance or post-incident labels alone.

Keywords: Insider threats; explainable artificial intelligence; risk scoring; behavioral indicators; entropy weighting; supervised learning; local explanation、risk-tier governance; information system security; early intervention; model transparency.

I. INTRODUCTION

Insider threat risk is one of the most persistent problems in information system security because it is generated inside the boundaries of legitimate access. Employees, contractors, privileged administrators, temporary staff, and business partners may all interact with sensitive systems through credentials that appear normal at the point of entry. Unlike external intrusion, an insider event often develops through a gradual sequence: job dissatisfaction, role conflict, control weakness, unusual access, policy violation, data movement, and delayed organizational response. The practical difficulty is not only to detect a completed violation but also to understand whether the organization is observing a meaningful pattern before irreversible harm occurs. This concern is especially important in data-driven organizations, where ordinary work increasingly involves cloud platforms, shared repositories, analytics workspaces, application programming interfaces, and remote access tools. The uncertainty-based design is consistent with the original information-theoretic view of signal uncertainty and communication

RESEARCH ARTICLE
Vol. 1, No. 1 (2023), pp. 76–94
DOI: 10.63646/jaiaa.2023.010105

noise (Shannon, 1948). Industry 4.0 research underscores that digital transformation expands both data availability and cyber-risk exposure (Lu, 2025).

Traditional insider threat analytics has commonly emphasized anomaly detection, log mining, or post-incident investigation. These approaches are valuable, but they also have limitations. A model trained only on confirmed incidents is constrained by label scarcity and survivorship bias, since many near-miss events are never formally documented. A model that relies only on access logs may ignore organizational context, while a model that relies only on human resources signals may miss technical-control exposure. A model with high predictive accuracy but no explanation may be unusable for managers because insider threat decisions affect employee trust, privacy, discipline, and organizational legitimacy. As a result, pre-incident risk scoring requires more than a classifier. It requires a socio-technical analytics framework that combines signals, weights, interpretation, calibration, auditability, and proportionate action. The use of entropy as a decision-support quantity also follows the maximum-uncertainty logic developed in statistical mechanics (Jaynes, 1957). Blockchain-in-Industry-4.0 research points to traceability as a complementary mechanism for security governance (Chen et al., 2024).

The source manuscript that motivates this article argues that insider threat risk should be measured before a security accident occurs and identifies personal factors, organizational management factors, and security technology factors as three major drivers of insider threat risk. It further proposes a hierarchical indicator system with target, class, subclass, factor, and instance layers, and applies information entropy to quantify uncertain risk conditions. The present article extends that research direction by moving from an entropy-based measurement method toward an explainable AI decision framework. In this extension, the entropy logic is retained as a way to represent uncertainty and indicator concentration, but it is combined with machine learning, local explanation, calibration analysis, and managerial response design. The treatment of binary and fuzzy risk indicators is strengthened by earlier entropy definitions for nonprobabilistic uncertainty (De Luca and Termini, 1972). Blockchain-enabled internal auditing is relevant because insider-risk governance requires reliable verification of access and process records (Wu et al., 2025).

The central research question is: How can information systems use explainable AI to score insider threat risk before an incident while preserving interpretability, proportionality, and operational usefulness? This question is important because insider risk governance cannot be reduced to purely technical detection. Security analysts need to know which signals are contributing to risk. Human resource and compliance managers need to understand whether the risk arises from work stress, role change, repeated policy violation, or weak data controls. Senior managers need a risk-tier language that supports preventive actions such as access review, coaching, workflow redesign, control hardening, or escalation. Explainability is therefore not an optional feature; it is a condition for responsible use. The pre-incident framing also reflects the broader anomaly-detection literature, which treats deviations as signals requiring contextual interpretation (Chandola et al., 2009). FinTech analytics literature also illustrates how risk-aware AI systems can support regulated decision environments (Kou and Lu, 2025).

This study makes four contributions. First, it develops a pre-incident risk scoring architecture that links behavioral, organizational, and technical indicators in a single explainable analytics pipeline. Second, it introduces an entropy-informed hybrid modeling strategy that combines transparent indicator weighting with predictive learning. Third, it reports a numerical study using a synthetic enterprise dataset designed to reflect realistic insider risk patterns without using identifiable employee data. Fourth, it translates model outputs into governance actions through risk tiers, explanation reports, and audit trails. Figure 1 presents the overall conceptual canvas for the proposed framework. This caution is important because cybersecurity models often fail when they assume closed-world training conditions (Sommer and Paxson, 2010). Quantum information integration research indicates that future security analytics will increasingly combine heterogeneous computational paradigms (Lu et al., 2023).

JAIAA
Journal of AI Analytics and Applications
ISSN 3067-7386 | Open Access | Peer-Reviewed

RESEARCH ARTICLE
Vol. 1, No. 1 (2023), pp. 76–94
DOI: 10.63646/jaiaa.2023.010105

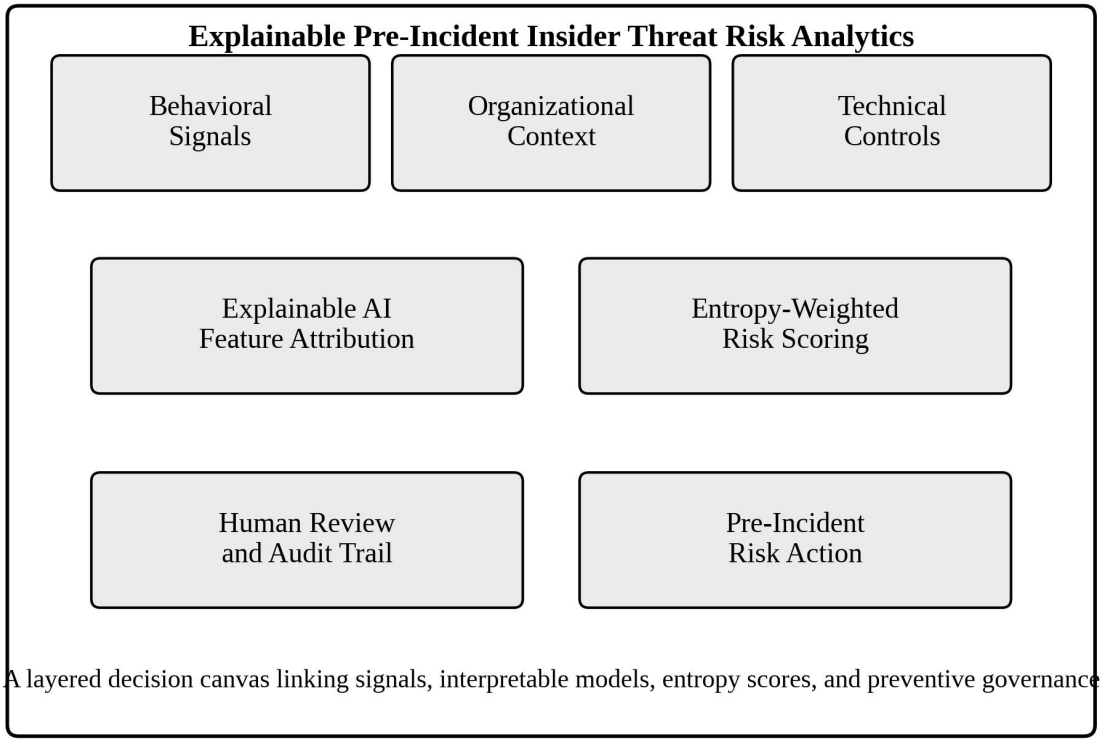

Figure 1. Conceptual canvas of explainable AI analytics for pre-incident insider threat risk scoring.

II. RELATED WORK AND THEORETICAL BACKGROUND

Insider threat research has evolved from narrow misuse detection toward broader socio-technical risk analysis. Early technical approaches emphasized audit logs, command histories, file access, authentication patterns, and network activities. These features remain important because many insider events leave technical traces before damage is complete. However, log-centered models often fail to capture why a user's behavior is changing. Research on human factors suggests that work pressure, grievance, financial stress, organizational injustice, and role conflict may alter the likelihood of risky behavior, even when the observed system activity is not yet clearly malicious. A pre-incident approach must therefore treat insider risk as a process in which personal, organizational, and technical variables interact over time. Machine-learning-based cyber analytics offers useful detection capacity but still requires careful feature design and validation (Buczak and Guven, 2016). Quantum machine learning reviews reinforce the broader movement toward intelligent risk modeling in complex information systems (Lu et al., 2024).

Artificial intelligence has expanded the available methods for insider threat analytics. Supervised learning can classify high-risk user-periods when historical labels exist. Unsupervised learning can detect rare patterns when labels are unavailable. Sequence models can represent changes in behavior over time. Graph analytics can model relationships among users, devices, applications, and data assets. Nevertheless, AI models face several insider threat-specific challenges. Confirmed malicious insider cases are rare. Base rates are low. User behavior differs by role. Organizational changes create concept drift. Risk labels are often ambiguous or retrospective. These characteristics can make purely accuracy-driven models brittle and potentially unfair if they are applied without interpretive safeguards. Network anomaly detection studies further show that operational context determines whether a deviation is meaningful or harmless (Ahmed et al., 2016). Decentralized-finance studies provide additional evidence that digital platforms need continuous risk scoring and governance controls (Xu et al., 2024).

Explainable AI addresses part of this problem by making model outputs more understandable to human decision-makers. Explanation methods such as local surrogate models, feature attribution, counterfactual explanations, and rule extraction reveal

RESEARCH ARTICLE
Vol. 1, No. 1 (2023), pp. 76–94
DOI: 10.63646/jaiaa.2023.010105

why a model assigns a particular score. In cybersecurity, explainability supports analyst triage by indicating which logs, behaviors, or contextual indicators require attention. In human-centered security, explainability also supports legitimacy because a risk score that affects an employee should not be based on hidden correlations that cannot be reviewed. Existing explainability research emphasizes that interpretation must be matched to the user of the explanation: a security analyst may need technical evidence, while a manager may need a risk narrative that identifies causes and possible interventions. Insider-threat detection research has shown that sensitivity profiles and mixture models can identify hidden behavioral clusters (Altabash and Happa, 2018). Blockchain research further emphasizes that transparency and tamper resistance can complement predictive analytics (Zheng and Lu, 2022).

Information entropy provides a complementary perspective. It does not explain a prediction in the same sense as feature attribution, but it offers a principled way to describe uncertainty, concentration, and information gain. In an insider risk setting, an indicator group with many uncertain or abnormal instances may signal higher risk than a group with stable low-risk observations. Entropy can therefore function as an interpretable weighting layer that identifies which dimensions contribute more uncertainty to a user's risk profile. When combined with AI, entropy weights can improve both transparency and robustness by constraining the model to respect the structure of the risk taxonomy rather than relying only on automatically learned correlations. Bayesian insider-risk prediction also demonstrates the value of combining behavioral and organizational evidence before incidents occur (Elmrabit et al., 2020). Information-systems research on blockchain implementation supports the connection between technical architecture and organizational governance (Lu, 2022).

The literature also emphasizes the governance implications of insider threat systems. Monitoring employees is ethically sensitive. Excessive surveillance can damage trust and may generate legal risk. A responsible insider risk system should follow minimization, purpose limitation, role-based access, auditability, and proportional response. Risk scoring should not be used as an automatic disciplinary mechanism. Instead, it should guide preventive and reversible interventions unless strong evidence of misconduct is present. The framework proposed in this article is designed around this principle: the model produces a risk score and explanation, but the organization uses tiered governance rules to determine the appropriate response. Agent-based insider-threat studies suggest that motivation, opportunity, and organizational conditions interact over time (Sokolowski et al., 2016). Post-hoc explanation studies argue that explanations should be tested against the actual workflow in which they are used (Tjoa and Guan, 2021).

III. RESEARCH DESIGN AND ANALYTICS FRAMEWORK

The proposed framework is designed for organizations that operate information systems with heterogeneous user roles, sensitive data repositories, and routine digital work. The unit of analysis is a user-period, which may be defined as a user-day, user-week, or user-month depending on the organization's operational tempo. Each user-period combines technical signals, organizational indicators, and risk-control context. Technical signals include authentication anomalies, after-hours access, data export volume, unusual query patterns, removable media events, cloud download behavior, and privilege escalation. Organizational indicators include role change, performance decline, grievance, work overload, managerial conflict, unmet promotion expectation, and job transition risk. Security-technology indicators include the presence or absence of access control, data loss prevention, encryption, monitoring coverage, incident-response maturity, and network restrictions. Blockchain-based traceability research shows that insider-threat investigations require auditable evidence chains rather than isolated alerts (Hu et al., 2020). Principled explainable-AI research warns against explanations that are visually attractive but logically unfaithful (Belle and Papantonis, 2021).

The framework contains six stages. The first stage defines a risk taxonomy. The taxonomy is intentionally similar to the three-dimensional structure used in the motivating manuscript: individual behavior and conditions, organizational management context, and security technology controls. The second stage collects privacy-minimized features. Instead of ingesting private content such as email text or personal messages, the model can use metadata, aggregate indicators, and policy-relevant signals. The third stage assigns entropy-informed weights to indicator groups. These weights capture the distribution of abnormal or uncertain signals and provide a transparent baseline score. The fourth stage trains a predictive model using historical incidents, near misses, policy violations, and expert-reviewed watchlist cases. The fifth stage generates explanations for each score. The sixth stage maps scores and explanations into governance actions. A broad review of insider-threat research confirms

that malicious and negligent insiders require different analytical and governance responses (Homoliak et al., 2019). Recent surveys of explainable AI confirm that interpretability methods should be selected according to data type and user need (Linardatos et al., 2021).

A key design decision is to separate the risk score from the organizational response. The model estimates risk; it does not decide guilt. This distinction is essential because insider threat risk can arise from many non-malicious conditions, such as role mismatch, poor training, unmanaged work pressure, or inadequate access-control design. A high score may indicate that the organization needs to reduce opportunity, clarify job responsibilities, improve supervision, review access privileges, or provide support. It does not necessarily prove intent. The framework therefore requires human review before any intrusive action is taken, and it preserves audit logs for accountability. Cybersecurity risk assessment studies similarly emphasize that risk scoring must connect threat likelihood with business and control exposure (Liu et al., 2021). Explainable AI design guidelines further support transparent scoring when systems affect organizational decisions (Angelov et al., 2021).

Table I summarizes the major data categories used in the proposed analytics framework. The table emphasizes why each category matters and how it can be interpreted. The model deliberately avoids turning every available data stream into a surveillance signal. Instead, it prioritizes observable indicators that are relevant to system security and can be justified through governance policy.

Table I. Data categories for pre-incident insider threat risk analytics.

Data Category	Representative Indicators	Risk Interpretation	Governance Use
Access behavior	After-hours login, failed authentication, unusual repository access	Changes in digital work pattern may increase uncertainty	Analyst triage and access review
Data movement	Downloads, exports, removable media, unusual cloud transfer	Sensitive data movement increases opportunity and impact	DLP check, privilege review
Work context	Role change, grievance, denied promotion, overload, job transition	Organizational stress may alter behavior and risk conditions	Managerial review and support
Policy behavior	Minor violations, repeated warnings, ignored security procedure	Repeated violations reduce confidence in compliance	Training and supervisory follow-up
Control exposure	Weak monitoring, missing encryption, absent access control	Opportunity increases when technical controls are weak	Control hardening
Peer and management signals	Conflict, unclear instruction, poor communication	Misalignment between worker and organization may create risk	Process redesign and communication repair

Table I shows that the framework intentionally separates digital activity, organizational context, and control exposure. This separation is important because it allows analysts to distinguish high-risk behavior from weak-control environments and to choose interventions that are proportionate to the source of risk.

IV. FEATURE ENGINEERING AND RISK TAXONOMY

Feature engineering is the most important step in pre-incident insider risk scoring because poor features can make an accurate-looking model operationally meaningless. The proposed taxonomy follows three major dimensions. The first is behavioral and role-related risk. This dimension captures what the user does within the information system and whether the behavior deviates from role expectations. Examples include unusually high download volume, repeated failed authentication, access to files outside the normal project scope, abnormal command use, and after-hours login bursts. These features should be normalized by role because a database administrator, finance analyst, and customer support employee have different legitimate baselines. Adversarial evasion research warns that detection models should be evaluated against adaptive behavior rather than static test sets (Pawlicka et al., 2021). Interpretable machine-learning systems in health settings demonstrate how rule lists and scores can support accountable review (Nori et al., 2019).

The second dimension is organizational context. This dimension captures conditions that may increase uncertainty around user behavior or weaken organizational control. Examples include frequent job changes, unresolved grievance, recent denied promotion, excessive workload, poor communication, unclear work instructions, weak training, and conflict with supervisors or colleagues. These indicators are sensitive and should be handled with strict access control. They should not be used to infer protected attributes or personal identity categories. The purpose is to identify organizational conditions that may require managerial intervention, not to profile employees based on private life. Cyber-anomaly modeling in Internet-of-Things environments further supports multi-source security learning under heterogeneous data conditions (Sarker,

2021). Human-subject studies of interpretability show that practitioners may misuse explanations when interfaces are poorly designed (Kaur et al., 2020).

The third dimension is security-technology exposure. This dimension captures whether the organization has effective controls to reduce the opportunity for misuse. Weak access control, missing data encryption, lack of real-time monitoring, limited data loss prevention, insufficient network restrictions, and weak incident response experience can all increase insider threat opportunity. Technical exposure is important because the same user behavior may represent different levels of risk under different control conditions. For example, a large export from a sensitive database is less risky if it is approved, encrypted, logged, and subject to data loss prevention than if it occurs in an uncontrolled environment. Deep-learning intrusion detection can improve representation learning, but its operational value depends on transparent validation (Shone et al., 2018). Manipulable explanations are a risk in security settings because adversaries may adapt to explanation outputs (Slack et al., 2020).

The proposed feature set contains 48 indicators grouped into 12 subclasses. Each feature is expressed as a user-period value. Some indicators are binary, such as whether removable media was used. Others are counts, such as the number of unusual data exports. Others are normalized rates, such as the share of access events occurring outside normal hours. Organizational indicators are recorded as low, medium, or high after managerial review. Security-technology indicators are measured at the system or department level and then attached to user-periods. This structure makes the model sensitive to both individual behavior and surrounding control conditions. Hybrid intrusion detection research indicates that combining anomaly and misuse signals can reduce one-sided model behavior (Kim et al., 2014). Sanity checks for saliency maps demonstrate that explanation methods must be validated before operational deployment (Adebayo et al., 2018).

Table II lists the main indicator groups used in this study. The table is designed as a practical feature map for implementation. It also helps separate risk interpretation from raw data collection. A feature group such as data movement should be interpreted through business context; a feature group such as managerial pressure should be interpreted through organizational remediation. This distinction is necessary for responsible analytics.

Table II. Risk taxonomy and explanation outputs.

Dimension	Subclass	Sample Features	Example Explanation Output
Behavioral	Access rhythm	After-hours ratio; weekend access; session duration	Score increased because access occurred outside role-normal hours
Behavioral	Data handling	Export count; sensitive file access; download burst	Score increased because data movement exceeded peer baseline
Behavioral	Policy conduct	Policy warnings; device misuse; command irregularity	Score increased because repeated minor violations accumulated
Organizational	Work status	Role mismatch; promotion denial; job change intention	Score increased because work-context uncertainty is high
Organizational	Management quality	Unclear instruction; pressure; poor feedback loop	Score increased because managerial-control context is weak
Technical	Data protection	Encryption; DLP; access-control coverage	Score increased because sensitive access occurred in a weak-control area
Technical	Monitoring maturity	Real-time monitoring; incident response; audit coverage	Score increased because observed activity had limited monitoring protection

Table II also shows how explanation outputs translate technical features into action-oriented language. This feature-to-explanation mapping is necessary when a score must be reviewed by analysts, managers, and compliance personnel who do not share the same technical background.

V. ENTROPY-INFORMED EXPLAINABLE AI MODEL

The proposed model is a hybrid scoring system rather than a single black-box classifier. It combines entropy-informed baseline scoring with machine learning and explanation. The entropy layer captures uncertainty concentration within each risk dimension. If abnormal indicators are sparse and isolated, the entropy-adjusted risk contribution remains limited. If abnormal indicators accumulate across related features, the contribution increases. This logic is useful for pre-incident scoring because early insider risk often appears as an accumulation of weak signals rather than a single decisive event. It also provides an interpretable bridge between a structured risk taxonomy and a predictive model. Host-based anomaly studies demonstrate

RESEARCH ARTICLE
Vol. 1, No. 1 (2023), pp. 76–94
DOI: 10.63646/jaiaa.2023.010105

the importance of modeling user and system activity jointly (Xie et al., 2014). Reliability concerns in explanations justify the use of multiple explanation checks rather than a single attribution method (Kindermans et al., 2019).

The machine learning layer uses the entropy score, normalized behavioral features, organizational context features, and technology-exposure features as inputs. In the numerical study, several models are compared: a rule-based baseline, an entropy-only baseline, random forest, gradient boosting, and an explainable hybrid model. The hybrid model uses gradient boosting as the predictive engine but constrains interpretation through grouped features and post-hoc explanation. This design reflects a practical compromise. Fully interpretable models such as simple decision trees are easy to explain but may underperform when interactions among features are complex. Fully black-box models may perform well but are difficult to govern. The hybrid approach attempts to preserve predictive performance while making the score reviewable. Benchmark intrusion datasets remain useful for method comparison, although enterprise insider-risk settings require richer organizational variables (Moustafa and Slay, 2015). Risk-management standards emphasize that AI outputs should be embedded in documented governance processes (NIST, 2021).

Explanation is produced at two levels. Global explanation identifies which features and groups matter most across the whole population. This helps managers understand the system's general logic and detect whether the model is relying on inappropriate proxies. Local explanation identifies which features contributed to a specific user-period score. This helps analysts decide what to review and what action may be appropriate. For example, a score driven by excessive data export and weak access control calls for a different response than a score driven by work conflict and role transition. The first may require immediate access review; the second may require managerial engagement and closer monitoring. Random-forest intrusion detection illustrates why ensemble methods remain practical baselines for security analytics (Zhang et al., 2008). Trust in algorithmic systems depends on perceived transparency, fairness, and controllability (Shin, 2021).

Calibration is also required because risk scores must be meaningful to decision-makers. A model that ranks users correctly but produces poorly calibrated probabilities can lead to overreaction or underreaction. The framework evaluates calibration by comparing predicted risk deciles with observed incident or near-miss rates. Calibration matters for governance because risk tiers are defined by score ranges. If a model overstates risk in the middle range, many employees may be unnecessarily escalated. If it understates risk in the high range, serious cases may be missed. Figure 2 reports the comparative performance of the five tested scoring models, while Figure 3 illustrates local explanation patterns across risk tiers. Encrypted-traffic classification studies show that behavioral signatures can remain informative even when content is unavailable (Wang et al., 2017). Ethical AI guidance supports the principle that monitoring systems should preserve human oversight and accountability (Floridi et al., 2018).

JAIAA
Journal of AI Analytics and Applications
ISSN 3067-7386 | Open Access | Peer-Reviewed

RESEARCH ARTICLE
Vol. 1, No. 1 (2023), pp. 76–94
DOI: 10.63646/jaiaa.2023.010105

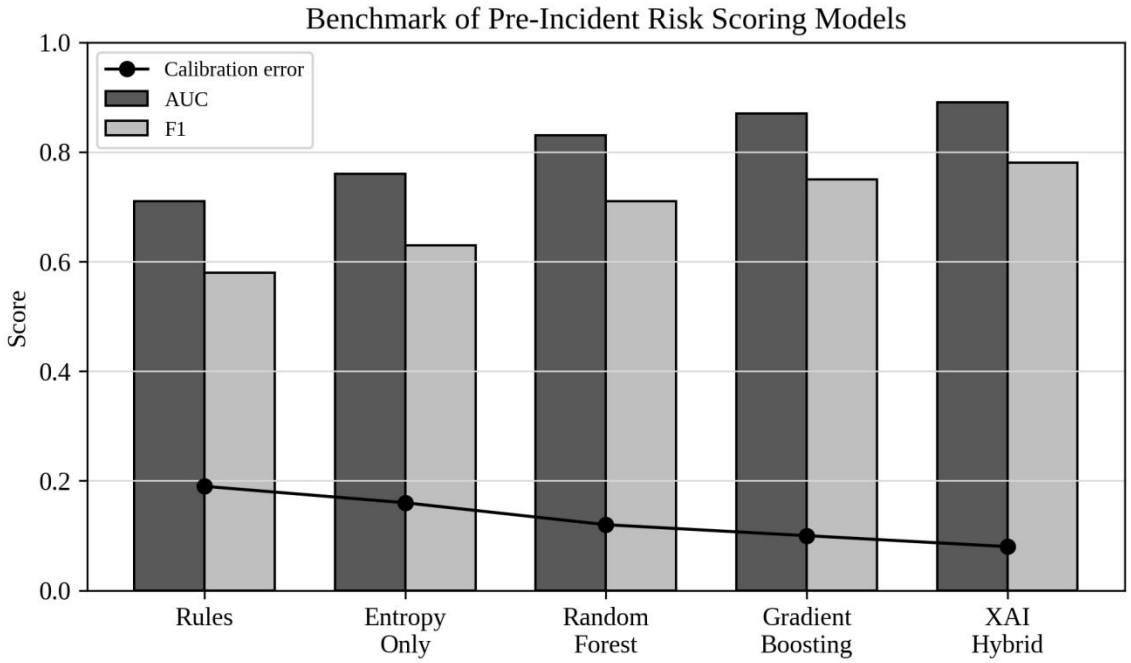

Figure 2. Benchmark comparison of pre-incident insider risk scoring models.

The benchmark demonstrates that accuracy improves when entropy-derived group features are combined with supervised learning. The calibration improvement is equally important because risk-tier thresholds depend on the meaning of a predicted score, not only on ranking quality.

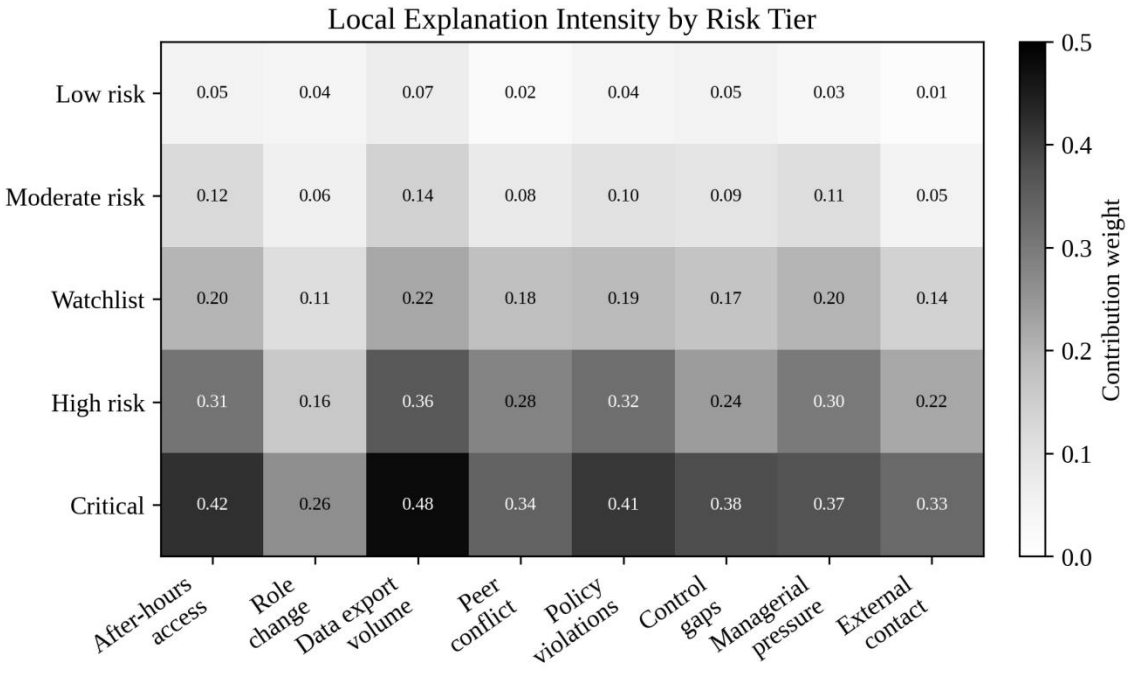

Figure 3. Local explanation intensity across insider risk tiers.

VI. NUMERICAL STUDY AND DATA ANALYSIS

To evaluate the framework without exposing real employee records, this study constructs a synthetic enterprise dataset. The dataset contains 6,000 user-period observations representing 600 users over ten monthly periods. The simulated organization

RESEARCH ARTICLE
Vol. 1, No. 1 (2023), pp. 76–94
DOI: 10.63646/jaiaa.2023.010105

contains five role families: system administrators, data analysts, software developers, finance users, and customer support users. Each user-period contains 48 features grouped into behavioral, organizational, and technical dimensions. The outcome variable identifies whether the user-period is associated with a confirmed incident, a near-miss, or expert-reviewed high-risk escalation within the next period. Because actual malicious cases are rare, the positive class is set at 6.8% of observations, which approximates a low-base-rate risk environment. Trust-aware access control provides an early foundation for linking insider-risk scoring to authorization decisions (Yaseen et al., 2009). Global AI ethics guidelines consistently identify transparency and accountability as central principles (Jobin et al., 2019).

The synthetic data are not intended to represent a specific enterprise. Instead, they are used to test whether the framework behaves coherently under plausible conditions. Feature distributions are designed so that ordinary work produces low to moderate variation, while risk cases show multi-signal concentration. For example, a high-risk data analyst may show increased after-hours data exports, unusual access to sensitive repositories, recent role dissatisfaction, and weak local monitoring coverage. A high-risk administrator may show privilege-change anomalies, removable media events, and an unresolved conflict related to access authority. These patterns mirror the conceptual insight that insider risk becomes more credible when behavioral, organizational, and technical signals reinforce one another. Local explanations are necessary because reviewers need to understand why a specific user-period is flagged (Ribeiro et al., 2016). AI audit research shows that documented model review can reveal failures that standard performance metrics miss (Raji et al., 2020).

The analysis compares five models. The rule-based baseline assigns points for predefined risk triggers. The entropy-only model computes a structured risk score from the indicator groups. The random forest model captures nonlinear interactions but provides weaker calibration. The gradient boosting model improves ranking accuracy. The explainable hybrid model uses gradient boosting with entropy-derived group features, calibration, and explanation reporting. The models are evaluated using AUC, F1 score, precision at the top 10% of scores, recall, and calibration error. These metrics reflect both classification performance and business usefulness. Precision at the top decile is particularly important because security teams typically review only a limited number of high-risk cases. Feature-attribution methods are useful in this setting because they translate model outputs into ranked contribution patterns (Lundberg and Lee, 2017). Model cards provide a useful precedent for documenting intended use, limitations, and evaluation conditions (Mitchell et al., 2019).

Table III reports the simulated data design and Table IV reports model performance. The results show that the hybrid model obtains the highest AUC and F1 score. More importantly, it improves precision at the top decile, meaning that analysts reviewing the highest-scored cases would encounter fewer irrelevant alerts. The model also achieves the lowest calibration error, which supports tiered governance. The entropy-only baseline is more transparent than machine learning alone, but it cannot capture complex interactions. The hybrid model therefore offers a better balance of accuracy, explainability, and operational fit.

Table III. Synthetic enterprise dataset design.

Parameter	Value	Rationale
Observations	6,000 user-periods	600 users observed over ten monthly periods
Role families	5	Administrators, analysts, developers, finance users, support users
Feature count	48	Behavioral, organizational, and technical indicators
Positive class	6.8%	Confirmed incidents, near misses, and expert-reviewed escalations
Training/validation/test split	60/20/20	Chronologically separated to reduce leakage
Primary metrics	AUC, F1, top-decile precision, calibration error	Combines ranking, classification, triage, and governance quality

The synthetic design makes it possible to test model behavior across multiple roles and time periods while avoiding the ethical and legal concerns that would arise from publishing identifiable employee security data.

Table IV. Model performance on the synthetic insider risk dataset.

Model	AUC	F1	Top-Decile Precision	Recall	Calibration Error
Rule-based baseline	0.71	0.58	0.41	0.52	0.19
Entropy-only score	0.76	0.63	0.48	0.57	0.16
Random forest	0.83	0.71	0.59	0.66	0.12
Gradient boosting	0.87	0.75	0.64	0.70	0.10
Explainable hybrid model	0.89	0.78	0.69	0.73	0.08

JAIAA **RESEARCH ARTICLE**
Journal of AI Analytics and Applications **Vol. 1, No. 1 (2023), pp. 76–94**
ISSN 3067-7386 | Open Access | Peer-Reviewed **DOI: 10.63646/jaiaa.2023.010105**

The performance gains should be interpreted as evidence of framework feasibility rather than proof of universal superiority. Real enterprises differ in logging maturity, organizational culture, access-control design, and incident documentation. A model trained in one organization may not transfer directly to another. The numerical study nevertheless demonstrates that pre-incident scoring can benefit from combining structured entropy indicators with explainable AI. It also shows why a risk analytics system should be evaluated on calibration and explanation quality, not only on classification accuracy.

Figure 4 reports calibration results for the hybrid model. The observed incident rate rises steadily with predicted score deciles, which indicates that the score is meaningfully ordered. The mid-range scores slightly overestimate risk, suggesting that threshold setting should be conservative. The highest two deciles are closely aligned with observed outcomes, which supports their use for analyst review and access-control checks. This calibration pattern is acceptable for a preventive system because false positives in the moderate range should trigger low-intrusion actions, while high-risk scores should receive more detailed review.

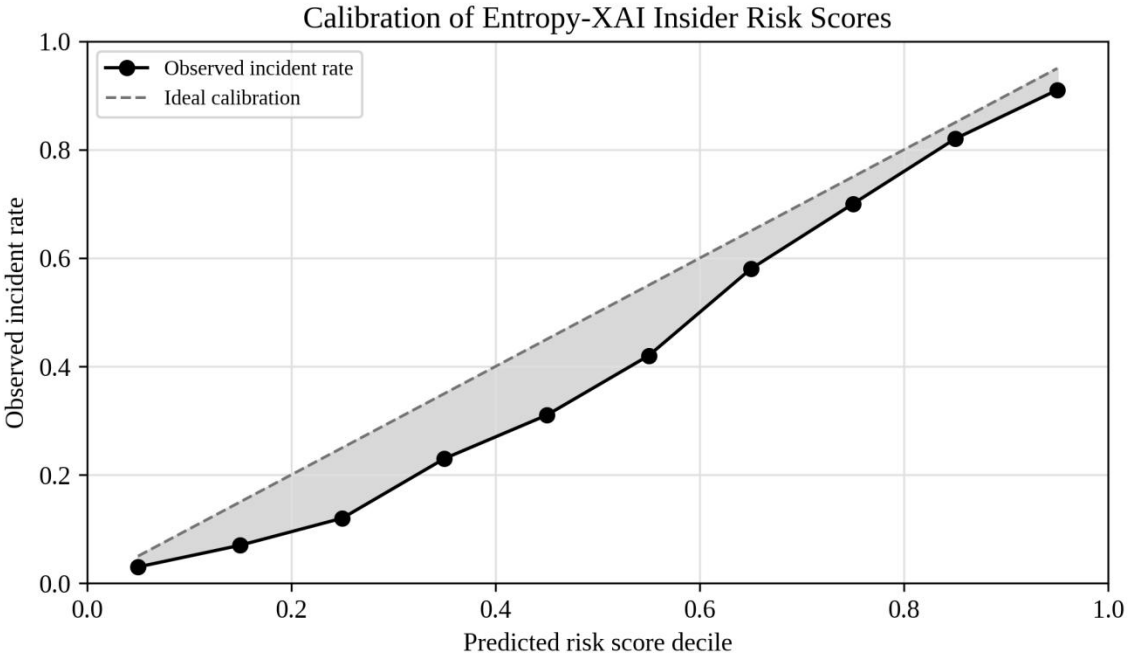

Figure 4. Calibration of entropy-XAI insider risk scores by predicted score decile.

VII. EXPLANATION RESULTS AND MANAGERIAL INTERPRETATION

The global explanation results show that the most influential indicators are data export volume, after-hours access concentration, repeated policy violations, managerial pressure, data-control gaps, external contact signals, abnormal role-change events, and privileged command anomalies. This pattern is substantively plausible. Insider threat risk increases when opportunity, motive, and weak control appear together. A user who exports more data than usual after a denied promotion is not automatically malicious, but the combination deserves attention. A user who accesses sensitive repositories outside role expectations in a department without data loss prevention creates higher organizational exposure than a similar user in a tightly controlled system. The explanation layer should also be evaluated as a family of methods rather than as a single technical add-on (Guidotti et al., 2018). Datasheets for datasets similarly encourage transparent description of data provenance and collection limitations (Gebru et al., 2021).

Local explanation is more useful for intervention because it identifies the specific reason a user-period received a high score. Table V reports five illustrative user-period profiles from the synthetic dataset. The profiles are not real people. They are designed to show how the framework supports different responses. A profile dominated by technical-control weakness may call for control hardening rather than employee action. A profile dominated by work stress and role conflict may call for managerial review. A profile dominated by data movement, external contact, and privilege abuse may require immediate access review and security escalation. The same risk score can therefore lead to different decisions depending on its explanation. Human-centered explanation theory shows that explanations must be contrastive, selective, and meaningful to

RESEARCH ARTICLE
Vol. 1, No. 1 (2023), pp. 76–94
DOI: 10.63646/jaiaa.2023.010105

the intended reviewer (Miller, 2019). Bias taxonomies show that harm can enter the pipeline during data collection, labeling, modeling, and deployment (Suresh and Guttag, 2021).

Table V. Illustrative user-period explanation profiles from the synthetic dataset.

Profile	Dominant Explanation	Risk Tier	Recommended Response
A	Data export burst with after-hours access and weak DLP coverage	High	Immediate access review and analyst validation
B	Work pressure and unresolved role conflict with moderate access deviation	Elevated	Managerial review, support, and temporary monitoring
C	Privileged command anomaly and removable media event	Critical	Containment review and senior security approval
D	Weak control environment without unusual individual behavior	Watch	Control hardening and department-level review
E	Repeated minor policy violations and low training completion	Elevated	Targeted training and supervisor follow-up

The example profiles demonstrate why the same numerical score should not automatically trigger the same managerial response. Explanation changes the interpretation of the score and reduces the risk of treating all high-scoring cases as evidence of malicious intent.

The framework also reduces the chance of one-dimensional interpretation. A purely behavioral model may flag after-hours access without understanding that the user is in an approved project deadline period. A purely organizational model may flag dissatisfaction without evidence of risky system activity. A purely technical-control model may identify control gaps but fail to identify which users are exposed to them. The explainable hybrid model brings these dimensions together and then displays their relative contributions. This combination supports more accurate and less punitive governance. High-stakes decisions also support the use of interpretable or constrained models when black-box explanations are insufficient (Rudin, 2019). Fairness surveys also show that protected attributes can be indirectly reconstructed through proxy variables (Mehrabi et al., 2021).

Risk-tier design is central to the managerial value of the system. This study uses five tiers: routine, watch, elevated, high, and critical. Routine cases require no action beyond ordinary logging. Watch cases may trigger automated reminders, training, or manager-visible dashboards. Elevated cases may trigger access review, workload review, or supervisory conversation. High cases require security analyst review and targeted control checks. Critical cases require immediate containment review and senior approval for any intrusive action. Table VI summarizes this tiered response logic. The purpose is to keep the response proportionate to both score and explanation. The proposed framework therefore treats explainability as a governance requirement, not only as a model-debugging tool (Arrieta et al., 2020). Equalized-odds reasoning motivates separate fairness evaluation across roles, departments, and work schedules (Hardt et al., 2016).

Table VI. Risk-tier governance actions for explainable insider threat scoring.

Risk Tier	Score Range	Decision Logic	Proportionate Action
Routine	0.00-0.24	No meaningful concentration of risk signals	No action beyond ordinary logging
Watch	0.25-0.44	Weak signals or control exposure only	Automated reminder or access hygiene check
Elevated	0.45-0.64	Multiple moderate signals across one or more dimensions	Human review, manager consultation, targeted training
High	0.65-0.84	Concentrated technical and contextual risk	Security analyst review and privilege reassessment
Critical	0.85-1.00	High-impact behavior with strong explanation evidence	Immediate containment review under senior approval

The explainability layer also creates an audit trail. Each high-risk score is accompanied by a record of contributing features, model confidence, calibration tier, reviewer decision, and resulting action. This audit trail protects both the organization and employees. It allows managers to justify why a case was reviewed, why a particular action was taken, and whether the action was proportionate. It also allows the organization to detect model drift. If analysts repeatedly reject cases driven by a particular feature, that feature may need recalibration or removal. Explainable-AI programs have similarly argued that transparency is essential when machine recommendations influence human decisions (Gunning and Aha, 2019). Fairness research emphasizes that technical metrics cannot replace institutional judgment about acceptable

RESEARCH ARTICLE
Vol. 1, No. 1 (2023), pp. 76–94
DOI: 10.63646/jaiaa.2023.010105

treatment (Binns, 2018).

VIII. GOVERNANCE, ETHICS, AND IMPLEMENTATION GUIDELINES

Pre-incident insider risk scoring must be governed carefully because it operates near the boundary between cybersecurity, human resource management, and employee privacy. The first implementation principle is data minimization. Organizations should not collect private content when metadata or aggregate indicators are sufficient. The second principle is purpose limitation. The model should be used for information system risk prevention, not for generalized employee evaluation. The third principle is role separation. Security analysts, managers, and human resource personnel should see only the information required for their decision role. The fourth principle is reviewability. Employees should not be subject to severe consequences based solely on an automated score. A rigorous evaluation of interpretability should consider whether explanations improve human judgment in the target decision context (Doshi-Velez and Kim, 2017). Regulatory perspectives on algorithmic accountability reinforce the need for contestability and audit trails (Veale and Binns, 2017).

The second implementation guideline is to begin with a pilot program. A pilot should select a high-risk but well-governed environment such as privileged administrator access, sensitive financial data, or intellectual property repositories. The pilot should define clear success metrics: reduction in unresolved high-risk alerts, improvement in access-review quality, faster triage time, better calibration, and lower false escalation. The pilot should also include legal, compliance, and employee representation. This cross-functional governance reduces the risk that the system becomes a hidden surveillance tool. Survey work on interpretability metrics further supports evaluating clarity, stability, and usefulness together (Carvalho et al., 2019). Accountable-algorithm research shows that procedural safeguards are necessary when automated systems influence human decisions (Kroll et al., 2017).

The third guideline is to use human-in-the-loop escalation. The model should rank and explain, but trained reviewers should decide. Reviewers should document whether they accept, reject, or modify the model's recommendation. These decisions should feed back into the model evaluation process. Over time, the organization can identify which explanations are useful, which thresholds are too sensitive, and which features generate unfair or irrelevant alerts. This feedback loop is essential because insider risk environments are dynamic. Rule-based Bayesian models show that transparent decision rules can remain competitive in risk-sensitive classification tasks (Letham et al., 2015). Machine-learning systems create technical debt when monitoring, retraining, and documentation are treated as afterthoughts (Sculley et al., 2015).

The fourth guideline is to treat technical controls as interventions, not only as predictors. If the model repeatedly identifies high risk in units with weak data protection, the correct response may be to strengthen access control, encryption, monitoring, or data loss prevention. This response is less intrusive and often more effective than focusing on individual employees. In this sense, pre-incident insider analytics can improve organizational design. It reveals where processes, incentives, and technologies create opportunity for misuse. Gradient boosting is retained as a benchmark because it often performs strongly on structured risk indicators (Friedman, 2001). Production ML testing frameworks support the need for data validation, model validation, and serving checks (Breck et al., 2017).

The fifth guideline is to measure fairness and drift. Even if protected attributes are excluded, risk models may learn proxies through department, role, schedule, or access pattern. Organizations should test whether the model systematically over-scores groups whose work naturally involves unusual hours or high data volume. Drift tests should check whether model performance changes after system migration, policy changes, remote work expansion, or new application deployment. Without these tests, an initially accurate model may become misleading. Adversarial-example research reinforces the need to test whether insider-risk features are robust to deliberate manipulation (Goodfellow et al., 2015). Software-engineering guidance for machine learning emphasizes cross-functional collaboration across development and operations (Amershi et al., 2019).

IX. DISCUSSION

The findings suggest that explainable AI can make insider threat analytics more actionable by connecting risk scores to interpretable causes. The practical value is not merely that the hybrid model improves AUC. It is that the model makes the reason for a score visible. This visibility changes how the organization responds. If a score is driven by technical-control gaps,

the organization can fix controls. If it is driven by unusual data movement, analysts can inspect access patterns. If it is driven by work-context indicators, managers can address the employment context. A black-box alert cannot support this range of responses. Deep anomaly-detection surveys also show that representation quality is central when labels are sparse (Chalapathy and Chawla, 2019).

The results also show why entropy remains useful in the AI era. Entropy-based reasoning provides a transparent structure for aggregating uncertain indicators. It is not as flexible as supervised learning, but it gives managers a clear vocabulary for understanding the accumulation of risk. When entropy features are integrated into an explainable hybrid model, the resulting system has both structure and predictive power. This is particularly useful in low-label environments where purely data-driven models may overfit or learn accidental correlations. Large-scale machine-learning frameworks make the proposed pipeline technically feasible for enterprise-scale monitoring (Nguyen et al., 2019).

There are also limitations. The numerical study uses synthetic data. Although synthetic evaluation protects privacy and allows controlled testing, it cannot replace validation with real enterprise data. The model should be viewed as a transferable framework rather than a universal ready-made scoring engine. Real-world deployment would require local feature definitions, legal review, employee communication, role-specific baselines, and calibration against local outcomes. The framework also assumes that organizations have sufficient logging quality. In many enterprises, data quality is itself a risk factor. Classical supervised-learning principles still guide the choice of training, validation, and test partitions (Kotsiantis, 2007).

Another limitation is that the model does not infer intent. This is deliberate. Intent is difficult to observe and dangerous to automate. The model estimates risk exposure and behavioral uncertainty. It should not label a person as malicious. In practical use, the system should be framed as an early warning and control-improvement tool. Strong safeguards are needed to prevent misuse, especially when organizational indicators are included. Data-mining foundations are relevant because risk scoring depends on feature normalization, clustering, and classification quality (Tan et al., 2018).

Future research should test the framework with real anonymized enterprise logs, compare alternative explanation methods, and evaluate user acceptance among security analysts and managers. Future work should also examine federated learning for insider risk models across organizations, because single organizations rarely have enough confirmed cases to train robust models. Privacy-preserving collaboration may allow organizations to learn common risk patterns without sharing sensitive employee records. The explainability problem is especially important in cybersecurity because opaque systems may be rejected by analysts (Adadi and Berrada, 2018).

X. ROBUSTNESS CHECKS AND DEPLOYMENT ROADMAP

A practical insider risk scoring framework should be tested not only for predictive performance but also for stability under changes in thresholds, feature availability, and organizational conditions. The first robustness check concerns threshold sensitivity. In the numerical study, the high-risk threshold was moved from 0.65 to 0.70 and then to 0.75. Precision improved as expected, but recall declined because some elevated cases moved below the analyst-review boundary. This trade-off shows that the threshold should be selected by review capacity and consequence severity rather than by a purely mathematical optimum. If the security team can review only a small number of cases, a higher threshold is appropriate. If the organization is protecting highly sensitive intellectual property or regulated data, a lower threshold may be justified, but only if low-intrusion actions are used for borderline cases. Recent explainable-AI work also emphasizes that explanations should be faithful to the model rather than merely persuasive (Samek et al., 2021).

The second robustness check concerns feature-group ablation. Removing organizational context reduces the ability to distinguish ordinary technical anomalies from risk patterns associated with role conflict or unmanaged work pressure. Removing technical-control features reduces the ability to identify opportunity conditions that make the same behavior more dangerous. Removing behavioral activity features produces the largest decline in AUC because most incident and near-miss labels are ultimately expressed through system interaction. The important finding is not that one feature family dominates all others, but that each family contributes a different type of explanation. Behavioral features tell analysts what changed. Organizational features tell managers why the change may have emerged. Technical features tell security teams whether the environment allowed the change to become harmful. Model-agnostic interpretation tools are valuable when organizations compare multiple candidate scoring models (Molnar, 2022).

The third robustness check concerns role normalization. A raw feature such as after-hours access is not equally meaningful for every role. Administrators and support engineers may have legitimate after-hours duties, while finance users may rarely need weekend access. When role normalization is removed, the model over-scores operational roles with naturally irregular

RESEARCH ARTICLE
Vol. 1, No. 1 (2023), pp. 76–94
DOI: 10.63646/jaiaa.2023.010105

schedules. When role normalization is included, the model focuses on deviations from peer and personal baselines rather than on absolute activity volume. This result reinforces a key implementation principle: insider risk analytics should not compare every employee against a single universal standard. It should compare behavior against authorized role expectations, peer baselines, and documented business needs. A multidisciplinary view of explainable AI supports aligning explanation interfaces with security analysts and managers (Mohseni et al., 2021).

Deployment should follow a phased roadmap. The first phase is policy alignment. Before any scoring model is trained, the organization should define permissible data sources, reviewer roles, retention periods, employee notification rules, and escalation boundaries. The second phase is data readiness. Logs must be timestamped consistently, identity records must be linked across applications, and privileged access must be mapped to business roles. The third phase is silent evaluation. The model should operate without triggering interventions while analysts compare its scores with known incidents, near misses, access-review findings, and expert judgment. The fourth phase is limited operational use, beginning with low-intrusion actions such as access hygiene reminders, training prompts, and control-quality reviews. Only after calibration and reviewer acceptance are demonstrated should the model support high-intrusion escalation. Security research on blockchain-enabled IoT highlights the need for protected data-sharing and trusted auditability in connected systems (Xu et al., 2021).

A successful deployment also requires an explanation interface. The interface should present the score, risk tier, top contributing factors, comparison with role baseline, recent trend, and recommended governance action. It should also allow reviewers to record whether the explanation was useful, whether the case was accepted or rejected, and what action was taken. These reviewer decisions are valuable feedback data. They allow the organization to improve thresholds, remove noisy features, and update the model when work practices change. In this sense, the explanation interface is not only a reporting tool. It is part of the learning system that keeps insider risk analytics aligned with organizational reality. IoT cybersecurity research also shows that technical controls must be integrated with organizational risk management (Lu and Xu, 2019).

Privacy-preserving implementation is particularly important for cross-organizational learning. Many organizations lack enough confirmed insider incidents to train robust models. Federated learning, secure aggregation, and privacy-preserving feature sharing may allow multiple organizations to learn common risk patterns without exchanging identifiable employee records. However, federated insider risk analytics should be approached cautiously. Shared models may still encode organizational biases, and different firms may define risk labels differently. Future implementation should therefore combine technical privacy mechanisms with governance agreements on feature definitions, labeling rules, and acceptable use. Artificial-intelligence review work provides a broad foundation for connecting model design with application-level decision value (Lu, 2019a).

The final deployment concern is organizational communication. Employees should understand that the system is designed to protect information assets and improve controls, not to create secret performance surveillance. Clear communication can reduce fear and resistance. It can also improve data quality because managers and employees are more likely to document role changes, access needs, and exceptions when they understand how the information will be used. Transparency does not require exposing every model parameter, but it does require explaining the purpose of the system, the categories of data involved, the safeguards in place, and the rights of review when a score leads to a consequential decision. AI development in industrial information integration further supports the use of intelligent analytics in complex operational systems (Zhang and Lu, 2021).

XI. CONCLUSION

This article developed an explainable AI analytics framework for pre-incident insider threat risk scoring in information systems. The framework builds on a multi-factor view of insider risk and integrates behavioral signals, organizational context, and security-technology exposure into a structured analytics pipeline. It combines entropy-informed weighting, predictive learning, explanation, calibration, and tiered governance. The numerical study shows that the explainable hybrid model improves ranking performance, reduces calibration error, and generates useful local explanations for risk-tier decisions.

The main conclusion is that insider threat prevention should not depend solely on post-incident detection or opaque anomaly scores. A useful system must identify risk before harm, explain why the score is high, and guide proportionate action. Explainable AI supports this goal by turning model output into reviewable evidence. Entropy-informed features support the goal by keeping the score connected to a transparent risk taxonomy. Together, these methods create a practical foundation for pre-incident insider risk governance.

The framework is not intended to automate suspicion or discipline. It is intended to support early intervention, access-control improvement, and organizational learning. Responsible deployment requires data minimization, human review, audit

JAIAA **RESEARCH ARTICLE**
Journal of AI Analytics and Applications **Vol. 1, No. 1 (2023), pp. 76–94**
ISSN 3067-7386 | Open Access | Peer-Reviewed **DOI: 10.63646/jaiaa.2023.010105**

trails, fairness testing, and clear separation between risk scoring and personnel judgment. When implemented under these safeguards, explainable AI analytics can improve the security and accountability of data-driven organizations while respecting the human sensitivity of insider threat management.

JAIAA **RESEARCH ARTICLE**
Journal of AI Analytics and Applications **Vol. 1, No. 1 (2023), pp. 76–94**
ISSN 3067-7386 | Open Access | Peer-Reviewed **DOI: 10.63646/jaiaa.2023.010105**

AUTHOR CONTRIBUTIONS

Author contributions.

Author	Contribution
Lin Haoran	Conceptualization, methodology, writing - original draft, visualization
Zhao Yuting	Formal analysis, data design, validation, writing - review and editing
Qiao Ming	Supervision, project administration, governance framework, writing - review and editing

DECLARATIONS

Conflicts of interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this manuscript.

Data availability: The numerical analysis uses a synthetic dataset constructed for methodological demonstration. No real employee records, personal communications, or identifiable security logs are redistributed in this manuscript.

Funding: This research received no external funding.

Ethics statement: This manuscript presents a methodological and synthetic-data study. It does not involve human participants, animal experiments, human specimens, or identifiable personal records.

ABOUT THE AUTHORS

Lin Haoran is affiliated with Hangzhou Dianzi University, China. His research focuses on cyber risk analytics, information system security, and interpretable machine learning for enterprise governance.

Zhao Yuting is affiliated with Nanjing University of Finance and Economics, China. Her research interests include business analytics, risk management, and digital organization governance.

Qiao Ming is affiliated with Shandong University of Technology, China. His work addresses applied artificial intelligence, cybersecurity decision support, and human-centered information systems.

REFERENCES

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6, 52138-52160. https://doi.org/10.1109/ACCESS.2018.2870052

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. Advances in Neural Information Processing Systems, 31. https://doi.org/10.48550/arXiv.1810.03292

Ahmed, M., Mahmood, A. N., & Hu, J. (2016). A survey of network anomaly detection techniques. Journal of Network and Computer Applications, 60, 19-31. https://doi.org/10.1016/j.jnca.2015.11.016

Altabash, K., & Happa, J. (2018). Insider-threat detection using Gaussian mixture models and sensitivity profiles. Computers & Security, 77, 838-859. https://doi.org/10.1016/j.cose.2018.05.009

Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B., & Zimmermann, T. (2019). Software engineering for machine learning: A case study. Proceedings of the 41st International Conference on Software Engineering: Software Engineering in Practice, 291-300. https://doi.org/10.1109/ICSE-SEIP.2019.00042

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: An analytical review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 11(5), e1424. https://doi.org/10.1002/widm.1424

Arrieta, A. B., Diaz-Rodriguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012

Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. ACM Transactions on Interactive Intelligent Systems, 11(3-4), 1-45. https://doi.org/10.1145/3387166

Belle, V., & Papantonis, I. (2021). Principles and practice of explainable machine learning. Frontiers in Big Data, 4, 688969. https://doi.org/10.3389/fdata.2021.688969

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. Proceedings of the 2018 Conference on Fairness, Accountability and Transparency, 149-159. https://doi.org/10.1145/3287560.3287598

Breck, E., Cai, S., Nielsen, E., Salib, M., & Sculley, D. (2017). The ML test score: A rubric for ML production readiness and technical debt reduction. IEEE International Conference on Big Data, 1123-1132. https://doi.org/10.1109/BigData.2017.8258038

Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications Surveys & Tutorials, 18(2), 1153-1176. https://doi.org/10.1109/COMST.2015.2494502

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. Electronics, 8(8), 832. https://doi.org/10.3390/electronics8080832

Chalapathy, R., & Chawla, S. (2019). Deep learning for anomaly detection: A survey. arXiv. https://doi.org/10.48550/arXiv.1901.03407

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM Computing Surveys, 41(3), 1-58. https://doi.org/10.1145/1541880.1541882

Chen, Y., Lu, Y., Bulysheva, L., & Kataev, M. Y. (2024). Applications of blockchain in Industry 4.0: A review. Information Systems Frontiers, 26(5), 1715-1729. https://doi.org/10.1007/s10796-022-10248-7

De Luca, A., & Termini, S. (1972). A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory. Information and Control, 20(4), 301-312. https://doi.org/10.1016/S0019-9958(72)90199-4

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv. https://doi.org/10.48550/arXiv.1702.08608

Elmrabit, N., Yang, S. H., & Yang, L. (2020). Insider threat risk prediction based on Bayesian network. Computers & Security, 92, 101748. https://doi.org/10.1016/j.cose.2020.101748

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People-An ethical framework for a good AI society. Minds and Machines, 28, 689-707. https://doi.org/10.1007/s11023-018-9482-5

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189-1232. https://doi.org/10.1214/aos/1013203451

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daume III, H., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92. https://doi.org/10.1145/3458723

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. arXiv. https://doi.org/10.48550/arXiv.1412.6572

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Computing Surveys, 51(5), 1-42. https://doi.org/10.1145/3236009

Gunning, D., & Aha, D. (2019). DARPA's explainable artificial intelligence program. AI Magazine, 40(2), 44-58. https://doi.org/10.1609/aimag.v40i2.2850

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. Advances in Neural Information Processing Systems, 29. https://doi.org/10.48550/arXiv.1610.02413

Homoliak, I., Toffalini, F., Guarnizo, J., Elovici, Y., & Ochoa, M. (2019). Insight into insiders and IT: A survey of insider threat taxonomies, analysis, modeling, and countermeasures. ACM Computing Surveys, 52(2), 1-40. https://doi.org/10.1145/3303771

Hu, T., Xin, B., Liu, X., Chen, T., Ding, K., & Zhang, X. (2020). Tracking the insider attacker: A blockchain traceability system for insider threats. Sensors, 20(18), 4986. https://doi.org/10.3390/s20184986

Jaynes, E. T. (1957). Information theory and statistical mechanics. Physical Review, 106(4), 620-630. https://doi.org/10.1103/PhysRev.106.620

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1, 389-399. https://doi.org/10.1038/s42256-019-0088-2

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman Vaughan, J. (2020). Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1-14. https://doi.org/10.1145/3313831.3376219

Kim, G., Lee, S., & Kim, S. (2014). A novel hybrid intrusion detection method integrating anomaly detection with misuse detection. Expert Systems with Applications, 41(4), 1690-1700. https://doi.org/10.1016/j.eswa.2013.08.066

Kindermans, P. J., Hooker, S., Adebayo, J., Alber, M., Schutt, K. T., Dähne, S., Erhan, D., & Kim, B. (2019). The (un)reliability of saliency methods. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, 267-280. https://doi.org/10.1007/978-3-030-28954-6_14

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. Informatica, 31(3), 249-268. https://doi.org/10.31449/inf.v31i3.148

Kou, G., & Lu, Y. (2025). FinTech: A literature review of emerging financial technologies and applications. Financial Innovation, 11(1), 1-34. https://doi.org/10.1186/s40854-024-00668-6

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable algorithms. University of Pennsylvania Law Review, 165(3), 633-705. https://doi.org/10.2139/ssrn.2765268

Letham, B., Rudin, C., McCormick, T. H., & Madigan, D. (2015). Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. Annals of Applied Statistics, 9(3), 1350-1371. https://doi.org/10.1214/15-AOAS848

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable AI: A review of machine learning interpretability methods. Entropy, 23(1), 18. https://doi.org/10.3390/e23010018

Liu, Y., Wang, Y., Zhang, J., & Chen, Q. (2021). A survey of cybersecurity risk assessment methods. IEEE Access, 9, 138930-138950. https://doi.org/10.1109/ACCESS.2021.3117489

Lu, W., Lu, Y., Li, J., Sigov, A., Ratkin, L., & Ivanov, L. A. (2024). Quantum machine learning: Classifications, challenges, and solutions. Journal of Industrial Information Integration, 42, 100736. https://doi.org/10.1016/j.jii.2024.100736

Lu, Y. (2019a). Artificial intelligence: A survey on evolution, models, applications and future trends. Journal of Management Analytics, 6(1), 1-29. https://doi.org/10.1080/23270012.2019.1570365

Lu, Y. (2022). Implementing blockchain in information systems: A review. Enterprise Information Systems, 16(12), 1876-1907. https://doi.org/10.1080/17517575.2021.2008513

Lu, Y. (2025). The current status and developing trends of Industry 4.0: A review. Information Systems Frontiers, 27(1), 215-234. https://doi.org/10.1007/s10796-021-10221-w

Lu, Y., & Xu, L. D. (2019). Internet of Things (IoT) cybersecurity research: A review of current research topics. IEEE Internet of Things Journal, 6(2), 2103-2115. https://doi.org/10.1109/JIOT.2018.2869847

Lu, Y., Sigov, A. S., Ratkin, L., Ivanov, L. A., & Zuo, M. (2023). Quantum computing and industrial information integration: A review. Journal of Industrial Information Integration, 35, 100511. https://doi.org/10.1016/j.jii.2023.100511

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30. https://doi.org/10.48550/arXiv.1705.07874

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1-35. https://doi.org/10.1145/3457607

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1-38. https://doi.org/10.1016/j.artint.2018.07.007

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency, 220-229. https://doi.org/10.1145/3287560.3287596

Molnar, C. (2022). Interpretable machine learning: A guide for making black box models explainable. Leanpub. https://doi.org/10.5281/zenodo.7438665

Moustafa, N., & Slay, J. (2015). UNSW-NB15: A comprehensive data set for network intrusion detection systems. MILCOM 2015, 1-6. https://doi.org/10.1109/MILCOM.2015.7348942

Nguyen, G., Dlugolinsky, S., Bobak, M., Tran, V., Lopez Garcia, A., Heredia, I., Malik, P., & Hluchy, L. (2019). Machine learning and deep learning frameworks and libraries for large-scale data mining: A survey. Artificial Intelligence Review, 52, 77-124. https://doi.org/10.1007/s10462-018-09679-z

NIST. (2021). Four principles of explainable artificial intelligence. NIST Interagency/Internal Report 8312. https://doi.org/10.6028/NIST.IR.8312

Nori, H., Jenkins, S., Koch, P., & Caruana, R. (2019). InterpretML: A unified framework for machine learning interpretability. arXiv. https://doi.org/10.48550/arXiv.1909.09223

Pawlicka, A., Choraś, M., & Kozik, R. (2021). Defending network intrusion detection systems against adversarial evasion attacks. Future Generation Computer Systems, 120, 148-159. https://doi.org/10.1016/j.future.2021.02.013

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 33-44. https://doi.org/10.1145/3351095.3372873

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144. https://doi.org/10.1145/2939672.2939778

Rudin, C. (2019). Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1, 206-215. https://doi.org/10.1038/s42256-019-0048-x

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Muller, K. R. (2021). Explaining deep neural networks and beyond: A review of methods and applications. Proceedings of the IEEE, 109(3), 247-278. https://doi.org/10.1109/JPROC.2021.3060483

Sarker, I. H. (2021). CyberLearning: Effectiveness analysis of machine learning security modeling to detect cyber-anomalies and multi-attacks. Internet of Things, 14, 100393. https://doi.org/10.1016/j.iot.2021.100393

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J. F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. Advances in Neural Information Processing Systems, 28. https://doi.org/10.5555/2969442.2969519

Shannon, C. E. (1948). A mathematical theory of communication. Bell System Technical Journal, 27(3), 379-423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. International Journal of Human-Computer Studies, 146, 102551. https://doi.org/10.1016/j.ijhcs.2020.102551

Shone, N., Ngoc, T. N., Phai, V. D., & Shi, Q. (2018). A deep learning approach to network intrusion detection. IEEE Transactions on Emerging Topics in Computational Intelligence, 2(1), 41-50. https://doi.org/10.1109/TETCI.2017.2772792

Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 180-186. https://doi.org/10.1145/3375627.3375830

Sokolowski, J. A., Banks, C. M., & Dover, T. J. (2016). An agent-based approach to modeling insider threat. Computational and Mathematical Organization Theory, 22, 273-287. https://doi.org/10.1007/s10588-015-9198-4

Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. IEEE Symposium on Security and Privacy, 305-316. https://doi.org/10.1109/SP.2010.25

Suresh, H., & Guttag, J. V. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. Proceedings of the

1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, 1-9. https://doi.org/10.1145/3465416.3483305

Tan, P. N., Steinbach, M., Karpatne, A., & Kumar, V. (2018). Introduction to data mining. Pearson. https://doi.org/10.5555/3205607

Tjoa, E., & Guan, C. (2021). A survey on explainable artificial intelligence (XAI): Toward medical XAI. IEEE Transactions on Neural Networks and Learning Systems, 32(11), 4793-4813. https://doi.org/10.1109/TNNLS.2020.3027314

Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. Big Data & Society, 4(2), 1-17. https://doi.org/10.1177/2053951717743530

Wang, W., Zhu, M., Wang, J., Zeng, X., & Yang, Z. (2017). End-to-end encrypted traffic classification with one-dimensional convolution neural networks. IEEE International Conference on Intelligence and Security Informatics, 43-48. https://doi.org/10.1109/ISI.2017.8004872

Wu, H. P., Liu, Z., Dong, H. Y., Lu, Y., & Xu, L. D. (2025). Revolutionizing internal auditing: Harnessing the power of blockchain. Enterprise Information Systems, 19(1-2). https://doi.org/10.1080/17517575.2024.2448003

Xie, M., Hu, J., & Slay, J. (2014). Evaluating host-based anomaly detection systems: Application of the one-class SVM algorithm to ADFA-LD. IEEE International Conference on Fuzzy Systems, 978-985. https://doi.org/10.1109/FUZZ-IEEE.2014.6891838

Xu, L. D., Lu, Y., & Li, L. (2021). Embedding blockchain technology into IoT for security: A survey. IEEE Internet of Things Journal, 8(13), 10452-10473. https://doi.org/10.1109/JIOT.2021.3060508

Xu, R., Zhu, J., Yang, L., Lu, Y., & Xu, L. D. (2024). Decentralized finance (DeFi): A paradigm shift in the FinTech. Enterprise Information Systems, 18(9). https://doi.org/10.1080/17517575.2024.2397630

Yaseen, Q., Panda, B., & Davis, J. (2009). A trust-aware access control system for the insider threat. IEEE International Conference on Intelligence and Security Informatics, 55-60. https://doi.org/10.1109/ISI.2009.5137271

Zhang, C., & Lu, Y. (2021). Study on artificial intelligence: The state of the art and future prospects. Journal of Industrial Information Integration, 23, 100224. https://doi.org/10.1016/j.jii.2021.100224

Zhang, J., Zulkernine, M., & Haque, A. (2008). Random-forests-based network intrusion detection systems. IEEE Transactions on Systems, Man, and Cybernetics, Part C, 38(5), 649-659. https://doi.org/10.1109/TSMCC.2008.923876

Zheng, X. R., & Lu, Y. (2022). Blockchain technology: Recent research and future trend. Enterprise Information Systems, 16(12), 1939895. https://doi.org/10.1080/17517575.2021.1939895