SSR Journal of Artificial Intelligence (SSRJAI)

OPEN ACCESS

Homepage: https://ssrpublisher.com/ssrjai/

Volume 3, Issue 8, 2026 ISSN: 3049-0413

An Explainable Adaptive Hybrid Artificial Intelligence Framework for Insider Threat Detection in Financial Institutions

Gilbert I.O Aimufua, Ohagwam Chidinma Maureen

Centre for Cyberspace Studies, Department of Cybersecurity, Nasarawa State University, Keffi

Received: *05.07.2026* | ***Accepted:*** *24.07.2026* | ***Published:*** *01.08.2026*

***Corresponding author:** Ohagwam Chidinma Maureen

DOI: *10.5281/zenodo.21737036*

Abstract **Original Research**

Financial institutions depend on trusted employees, contractors and service accounts, yet this trust creates an attack surface that conventional perimeter controls cannot observe adequately. This paper develops an Explainable Adaptive Hybrid Artificial Intelligence (EAHAI) framework for insider threat detection and for assessing whether security awareness training is reducing measurable insider-risk behaviour. The framework combines Isolation Forest filtering, bidirectional long short-term memory sequence modelling, Shapley Additive explanations, adaptive behavioural risk scoring and Zero Trust policy enforcement. A socio-technical assessment layer is added to link training inputs to observable outcomes, including knowledge gain, phishing susceptibility, policy-violation rates, reporting delay, behavioural-risk reduction and analyst-confirmed events. The paper defines the measurement scales, evaluation criteria, validation procedures and analytical techniques required for institutional replication. Because production banking telemetry and labelled insider incidents are rarely available for publication, the empirical component is presented as a transparent synthetic proof-of-concept based on CERT-style behavioural variables rather than as evidence from a real bank. In a deterministic simulation of 17,280 user-day records and 2,880 test windows, the proposed hybrid score achieved an F1-score of 0.944, ROC-AUC of 0.993 and false-alarm rate of 0.017, while producing interpretable feature attributions and training-effectiveness estimates. The study contributes a scalable, explainable and ethically governed design for insider-risk analytics, and identifies the conditions under which it should be validated before operational deployment.

Keywords: insider threat detection; explainable artificial intelligence; adaptive risk scoring; security awareness training; Zero Trust; financial cybersecurity.

1. Introduction

Digital finance is now a densely connected socio-technical system. Banks and other regulated financial institutions operate mobile banking, cloud analytics, algorithmic risk platforms, open-banking interfaces and third-party technology services. These arrangements improve efficiency but make the boundary of the organisation porous. Insider threat is therefore not limited to a malicious employee stealing data; it also includes negligent behaviour, compromised credentials, privileged misuse, contractor abuse and inadvertent violations of

policy. The problem is difficult precisely because insiders usually possess legitimate credentials and can perform harmful actions through apparently authorised workflows [1,2].

The empirical signature of insider threat is also unlike many external intrusions. Malicious insiders often progress through low-intensity preparatory activity: unusual login times, repeated access to sensitive repositories, new removable-media use, gradual escalation of file downloads, or communications with destinations that are unusual for their role. Rule-based systems and signature-driven intrusion detection tools struggle with this pattern because each event may be defensible in isolation. Financial institutions therefore require behavioural detection that can learn baselines, model temporal context, explain decisions to analysts, and trigger proportionate access-control responses instead of producing opaque risk scores [3,4].

Artificial Intelligence (AI) offers useful capabilities for this setting. Isolation Forest can isolate unusual observations efficiently in high-dimensional and imbalanced data [5,6]. Recurrent architectures such as long short-term memory (LSTM) networks can represent time-dependent behaviour [7], and bidirectional variants exploit both preceding and subsequent context within an observation window [8]. Explainable Artificial Intelligence (XAI), particularly Shapley Additive explanations (SHAP), can translate complex predictions into feature-level evidence that analysts can assess [10,12]. However, these methods are often treated as separate tools. Their integration with Zero Trust enforcement and with measurable security awareness outcomes remains under-developed.

This paper upgrades the existing EAHAI idea into a complete research design suitable for a short Elsevier or Springer-style article. The key novelty is not a new classifier in isolation. It is a replicable, adaptive and explainable assessment framework in which behavioural telemetry, temporal AI, analyst feedback and security awareness indicators are fused into a single decision-support architecture. The framework is deliberately conservative: it does not claim that synthetic results prove effectiveness in live finance environments. Instead, it provides a mathematically specified and ethically bounded model that can be tested with secondary datasets, synthetic case studies and, later, institutionally approved operational telemetry.

The contributions are fourfold. First, the paper proposes an integrated insider-risk architecture linking Isolation Forest, Bi-LSTM, SHAP, adaptive behavioural scoring and Zero Trust action. Second, it develops an assessment framework that defines how security awareness training effectiveness is measured through knowledge, behaviour and risk outcomes rather than attendance alone. Third, it specifies instruments, reliability procedures, validation checks and evaluation metrics needed for replication. Fourth, it reports a transparent synthetic proof-of-concept to demonstrate how analytical outputs would be derived without exposing confidential banking logs.

2. Literature review and research gap

Early insider-threat research emphasised the organisational and behavioural roots of misuse. CERT studies of financial-sector fraud highlighted that malicious actions can occur during normal working conditions and may be preceded by observable workplace and system indicators [1,2]. Greitzer and Frincke argued that technical audit logs become more valuable when combined with contextual and behavioural evidence rather than interpreted as isolated system events [3]. This socio-technical stance remains important for modern financial cybersecurity because excessive surveillance without explanation can undermine trust, whereas purely voluntary training cannot detect covert misuse.

From a machine-learning perspective, insider detection is usually treated as anomaly detection. Chandola et al. described anomaly detection as the identification of rare observations that deviate from the majority pattern [4]. Isolation Forest is attractive in this context because it isolates anomalies using random partitions rather than modelling all normal density; this gives favourable time and memory behaviour for large telemetry streams [5,6]. Its weakness is that it is largely point-wise. A single unusual file

download may not be malicious; a sequence of abnormal downloads, after-hours logins and external communications may be.

Deep learning addresses this temporal limitation. LSTM networks were developed to capture long-range dependencies [7], and bidirectional LSTM extends sequence learning by considering both forward and backward dependencies [8]. Tuor et al. demonstrated the value of deep and recurrent models for unsupervised insider-threat detection on structured cybersecurity data streams [9]. Yet deep sequence models are more expensive to run than simpler anomaly filters and are commonly criticised for limited interpretability. In a security operations centre, a model that cannot explain why an employee is considered high-risk is difficult to defend operationally or ethically.

XAI responds to this problem. SHAP provides additive feature attributions derived from game-theoretic Shapley values [10,25], while LIME offers local surrogate explanations for black-box models [11]. Surveys of XAI emphasise that explanations should support trust, debugging and accountability, not merely decorate model output [12]. For financial institutions, explanation is a governance requirement as much as a usability requirement, because high-risk AI-supported decisions may affect access privileges, investigations and employee rights.

Zero Trust offers the policy environment in which adaptive insider-risk scores can be used. NIST SP 800-207 defines Zero Trust as a shift from static perimeter assumptions to continuous, resource-focused access decisions [13]. NIST's AI Risk Management Framework further emphasises validity, reliability, transparency, accountability and human oversight for trustworthy AI systems [14]. Security awareness training is also a formal requirement in recognised guidance such as NIST SP 800-50 and is consistent with ISO/IEC 27001 and ISO/IEC 27002 controls for information security management and people-related controls [15-17]. However, most training programmes still measure completion rather than behavioural risk reduction.

The resulting gap is clear. Current studies offer strong components: anomaly filters, sequence models, XAI tools, awareness training and Zero Trust. What is missing is a compact framework that (i) filters telemetry efficiently, (ii) models behavioural sequences, (iii) explains decisions, (iv) adapts to drift, (v) measures security awareness training effects on insider-risk behaviour, and (vi) links the risk score to auditable Zero Trust actions. The EAHAI framework is designed to address that gap.

Table 1. Gap analysis linking the evidence base to the EAHAI design response.

Evidence stream	Established strength	Observed limitation	EAHAI design response
Unsupervised anomaly detection	Efficient detection of rare deviations in imbalanced data [4-6]	Weak representation of long-running behavioural campaigns	Use Isolation Forest only as an initial filter, not as the final threat decision.
Sequential deep learning	Captures multi-step behavioural evolution [7-9]	Computational cost and low transparency when applied to all events	Route only anomalous windows into Bi-LSTM and explain outputs through SHAP.
Explainable AI	Converts black-box predictions into feature-level evidence [10-12]	Explanations are often post hoc and disconnected from policy action	Embed SHAP into risk fusion and analyst review before Zero Trust enforcement.

Evidence stream	Established strength	Observed limitation	EAHAI design response
Security awareness training	Required by security governance and training guidance [15-17]	Often assessed through completion rather than risk outcomes	Measure knowledge, simulated phishing, policy violations, reporting delay and risk-score reduction.
Zero Trust	Supports continuous resource-centred access decisions [13]	Few insider models provide auditable behavioural evidence for policy changes	Map risk tiers to permit, verify, restrict and contain actions with human oversight.

3. Theoretical and conceptual framework

The proposed framework is grounded in three complementary assumptions. The first is behavioural: insider risk manifests as deviations from a user's expected role, context and history rather than as a single universal signature. The second is socio-technical: training, policy, workload, reporting culture and access-control design all shape risky behaviour. The third is governance-oriented: automated detection must be explainable, auditable and proportionate because insider analytics affects people as well as systems.

Figure 1 presents the conceptual architecture. Enterprise telemetry is transformed into normalised behavioural vectors. Isolation Forest provides a fast anomaly score, a(u,t), which determines whether a user-window requires deeper temporal analysis. Suspicious windows are passed to a Bi-LSTM that estimates p(u,t), the temporal probability of insider-risk behaviour. SHAP attributes the contribution of features such as privileged access, upload volume, off-hours activity and policy violations. These outputs are fused with context, resource sensitivity, device trust and training assimilation indicators to produce an adaptive risk score. The Zero Trust decision engine then chooses a proportionate response, while analyst feedback updates thresholds, feature weights and training priorities.

The security awareness component is not an appendix to the technical model. It is treated as an assessment layer that tests whether training reduces observable insider-risk indicators. This responds to the common weakness of awareness programmes: attendance can be high while risky behaviour remains unchanged. The framework therefore measures both learning outcomes and operational outcomes, then compares them with behavioural risk data generated by the detection pipeline.

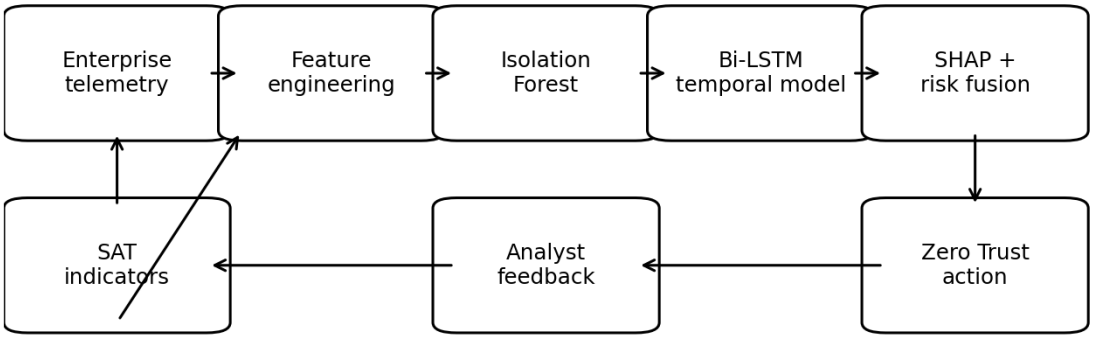

Fig. 1. Conceptual architecture of the EAHAI framework

4. Assessment framework development

The assessment framework is developed before data collection and analysis to avoid post hoc selection of favourable metrics. It defines constructs, indicators, scales and evaluation criteria for two questions: (RQ1) How accurately and explainably can EAHAI identify insider-risk windows? (RQ2) Does security awareness training reduce measurable insider-risk behaviour after controlling for baseline differences and behavioural drift?

The central construct is adaptive insider-risk reduction. It is measured by combining model-derived indicators with human and organisational indicators. Knowledge and attitude are measured through pre- and post-training tests, scenario-based judgement items and a short Likert-scale awareness instrument. Behaviour is measured through phishing-simulation clicks, reporting delay, policy violations, abnormal authentication, sensitive file access, removable-media use and external data movement. Model effectiveness is measured through precision, recall, F1-score, ROC-AUC, PR-AUC, Matthews correlation coefficient, false-alarm rate and latency. Explanation quality is measured using top-k stability of explanations, analyst agreement and the proportion of high-risk alerts with an intelligible rationale.

The framework explicitly distinguishes malicious from negligent risk. Security awareness training is expected to reduce negligent behaviours such as phishing susceptibility, policy violations and reporting delay. It should not be assumed to prevent a determined malicious insider. For this reason, training effectiveness is not defined as the elimination of insider threat but as a statistically and operationally meaningful reduction in preventable risk indicators. Figure 2 summarises the assessment logic.

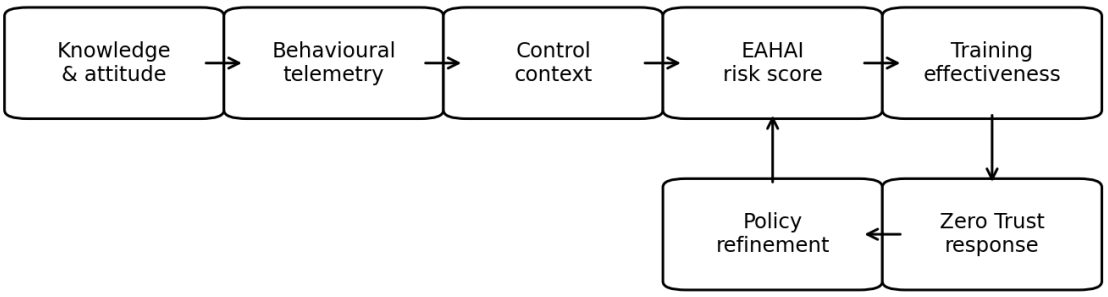

Fig. 2. Assessment model for linking security awareness training to behavioural-risk outcomes.

Table 2. Constructs, indicators, measurement scales and evaluation criteria.

Construct	Indicator	Scale/source	Criterion	Use in analysis
Training assimilation	Knowledge test; scenario judgement; self-reported awareness	0-100 score; 5-point Likert instrument; validated items	Reliability alpha >= 0.70; content validity index >= 0.80	Creates TAI(u,t), a training assimilation index.
Risk behaviour	Phishing click; policy violation; reporting delay; off-hours access; USB use	Binary, counts, hours and rates per user-window	Reduction after training compared with baseline/control	Direct behavioural outcome for awareness effectiveness.
Detection performance	Accuracy, precision, recall, F1, ROC-AUC, PR-AUC, MCC	Confusion matrix and ranking metrics	High recall with controlled false-alarm rate	Evaluates EAHAI threat identification.
Explainability	Top features, explanation stability, analyst agreement	Mean |SHAP|; Jaccard@k; review rubric	Stable top-k explanations and actionable analyst notes	Supports auditability and SOC adoption.
Operational response	Permit, verify, restrict, contain	Risk-tier decision rules	Proportionate action; human review for high-impact decisions	Connects risk score to Zero Trust enforcement.

5. Mathematical formulation

Let x(u,t) = [x1, x2, ..., xm] denote the behavioural feature vector for user u during observation window t. Features include authentication activity, access to sensitive resources, upload volume, removable-media events, device trust, privilege level, department sensitivity and training assimilation. Because telemetry variables have different units, each continuous feature is normalised as x'j = (xj - min(xj))/(max(xj) - min(xj) + epsilon). Temporal windows are represented as S(u,t) = {x'(u,t-w+1), ..., x'(u,t)}, where w is the sequence length.

The Isolation Forest score is denoted a(u,t), with larger values representing greater deviation

from the normal baseline. The Bi-LSTM maps S(u,t) to p(u,t) = sigmoid(W h(u,t) + b), where h(u,t) concatenates forward and backward hidden states. The training assimilation index is defined as TAI(u,t) = omega1 K(u,t) + omega2 J(u,t) + omega3 Q(u,t), where K is knowledge-test performance, J is scenario judgement and Q is policy-compliance behaviour. Weights are pre-specified by expert review and may be updated after reliability testing.

The adaptive risk score is: R(u,t) = clip[100 {0.72 p(u,t) + 0.20 a(u,t) + 0.08 C(u,t) - 0.05 TAI(u,t)}, 0, 100], where C(u,t) combines privilege, resource sensitivity and device-trust risk. This formulation intentionally gives greatest weight to temporal evidence while allowing context and training assimilation to influence prioritisation. Thresholds are not fixed permanently. For validated normal windows, theta(t+1) = lambda theta(t) + (1-lambda) Q(1-alpha)(R_t) + eta Delta(t), where Q(1-alpha) is a high quantile of recent normal risk scores and Delta(t) is a drift measure based on the distance between current and previous behavioural centroids.

Training effectiveness is assessed through pre-post and comparison-group changes. For outcome Y, the difference-in-differences estimate is DID = (mean Y_trained, post - mean Y_trained, pre) - (mean Y_control, post - mean Y_control, pre). A negative DID is favourable for harmful behaviours such as phishing clicks, policy violations, reporting delay and risk-event rate. For knowledge, normalised gain is G = (post - pre)/(100 - pre). These formulae make assessment outcomes reproducible and prevent substituting anecdotal satisfaction scores for behavioural evidence.

***Algorithm 1**. EAHAI adaptive insider-risk assessment and Zero Trust response.*

Input: Enterprise telemetry L; training records T; analyst feedback F; thresholds theta_low, theta_high.

1. Clean, synchronise and normalise logs from identity, endpoint, file, network, email and cloud sources.
2. Construct behavioural vectors x(u,t) and temporal windows S(u,t) for each user and window.
3. Estimate Isolation Forest anomaly score a(u,t) against the recent normal baseline.
4. If a(u,t) < theta_filter, classify as routine and retain for baseline drift estimation; else send S(u,t) to Bi-LSTM.
5. Compute temporal probability p(u,t) and SHAP feature attributions phi_j(u,t).
6. Compute training assimilation TAI(u,t) and contextual sensitivity C(u,t).
7. Fuse p(u,t), a(u,t), C(u,t) and TAI(u,t) to produce R(u,t).
8. Apply Zero Trust policy: low risk = permit; medium risk = verify or step-up MFA; high risk = restrict, contain and alert SOC.
9. Record analyst judgement and investigation outcome; update thresholds, feature weights, drift estimates and training priorities.

Output: Risk tier, explanation report, action decision and training-effectiveness indicators.

6. Research methodology

A design-science research strategy is appropriate because the study builds and evaluates an artefact intended to solve a practical organisational problem [26]. The methodology has four phases: conceptual design, assessment-framework specification, computational proof-of-concept and validation planning. The present article completes the first three phases and

specifies the fourth for future institutional deployment.

The target setting is a financial institution with identity-management logs, endpoint telemetry, file-access records, cloud audit trails, email metadata, training records and SOC investigation outcomes. In a live study, these data would be collected only after data protection assessment, ethics approval, role-based access control, pseudonymisation and staff consultation. Personally identifiable information would not be needed for model development; users would be represented by pseudonymous identifiers and role attributes. High-impact automated decisions would remain subject to human review.

The proposed data collection instruments are: (i) a log-extraction protocol specifying fields, retention period and pseudonymisation method; (ii) a security awareness knowledge test aligned with institutional policy and NIST SP 800-50 [15]; (iii) scenario-based judgement items and Likert-scale awareness questions adapted from established human-aspects measurement practice [18,21]; (iv) phishing-simulation and reporting logs; (v) an analyst validation form capturing true incident, benign anomaly, policy breach and insufficient-evidence outcomes; and (vi) an explanation-review rubric to assess whether top SHAP features are operationally meaningful.

Validation is conducted at three levels. Content validity is assessed by a panel of cybersecurity, compliance and financial-risk experts. Construct reliability is assessed using Cronbach's alpha for multi-item awareness scales, with 0.70 used as a minimum acceptable value for early research instruments [27]. Model validation uses time-based train/test separation, cross-validation within the training period, threshold selection on validation data only and final reporting on an untouched test period. Calibration, drift and false-alarm behaviour are examined because a high AUC is insufficient if the system overwhelms analysts with low-value alerts.

Because real financial-sector telemetry could not be disclosed in this article, a synthetic proof-of-concept was implemented. The generator created CERT-style user-day records, not real banking data. It used role, privilege, department sensitivity, authentication counts, sensitive file access, upload and download volumes, external communications, removable-media events, device trust, policy violations, reporting delay and training assimilation. Malicious campaigns were injected as multi-day sequences, whereas training mainly affected negligent behaviours. The deterministic run used seed 42, 360 pseudonymous users, 48 days, 17,280 user-day records and 2,880 test windows.

7. Data analysis and results

The synthetic evaluation should be read as a reproducible demonstration of the framework rather than as empirical proof of performance in a real bank. The test set contained 2,880 temporal windows with a positive insider-risk window rate of 18.2%. Thresholds were selected on training data and then applied to the held-out period. Table 3 reports model performance. The hybrid score retained high recall while keeping the false-alarm rate low. Bi-LSTM alone produced slightly higher F1 in this synthetic setting, but it does not provide the same adaptive risk fusion, training-effectiveness linkage and policy mapping. The Isolation Forest-only baseline was useful as a high-recall filter but was unsuitable as a standalone decision model because of its high false-alarm rate.

Table 3*. Synthetic proof-of-concept detection results on the held-out test period.*

Model	Threshold	Accuracy	Precision	Recall	F1	ROC-AUC	False-alarm rate
Isolation Forest only	0.050	0.212	0.183	0.958	0.307	0.631	0.954

Model	Threshold	Accuracy	Precision	Recall	F1	ROC-AUC	False-alarm rate
Bi-LSTM only	0.845	0.981	0.942	0.952	0.947	0.997	0.013
EAHAI hybrid risk	0.640	0.979	0.928	0.960	0.944	0.993	0.017

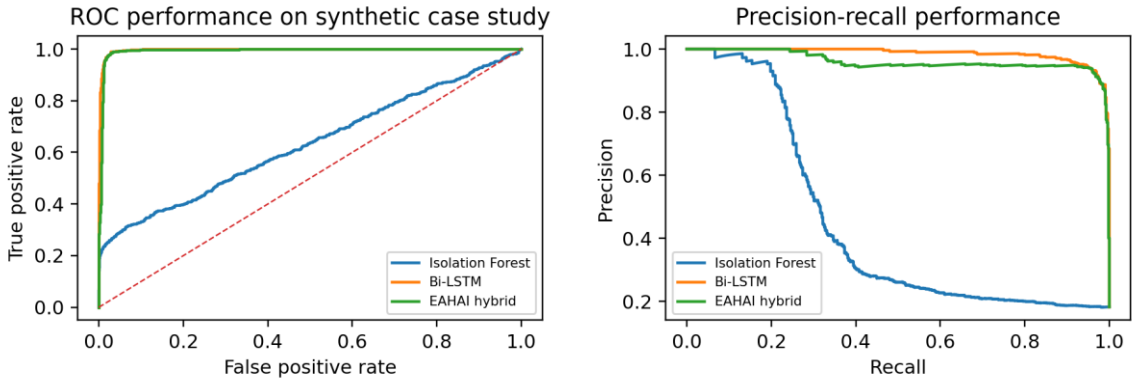

Fig. 3*. ROC and precision-recall curves for the synthetic proof-of-concept.*

The SHAP explanation layer was demonstrated using a tree-based surrogate model fitted to the same feature space. The highest mean absolute attributions were policy violation, reporting delay, device trust, failed logins, removable-media events and external email volume. This ranking is plausible for the synthetic generator and helps illustrate how an analyst would receive behavioural evidence rather than a bare probability. The top-five attribution sets were stable across two bootstrap surrogate models in the deterministic run (Jaccard@5 = 1.00), although this value should not be generalised beyond the synthetic setting.

Training-effectiveness analysis used the pre-post and comparison-group logic defined earlier. Mean training assimilation in the training cohort increased from 0.586 to 0.762 after the simulated intervention. The comparison cohort remained stable at approximately 0.572. Harmful behavioural indicators decreased more sharply for the training cohort: phishing-click rate, policy-violation rate, reporting delay and risk-event rate all showed favourable difference-in-differences estimates. This is not evidence that a particular training package works in practice; it demonstrates how the framework would quantify whether training is associated with reduced behavioural risk after baseline adjustment.

Table 4. *Synthetic security awareness training assessment outputs.*

Group	Period	Mean TAI	Phishing click	Policy violation	Reporting delay (h)	Risk events /100 user-days
Comparison cohort	pre	0.572	0.193	0.102	21.07	6.63
Comparison cohort	post	0.572	0.206	0.111	21.04	6.78
Training cohort	pre	0.586	0.194	0.101	20.84	6.17

Group	Period	Mean TAI	Phishing click	Policy violation	Reporting delay (h)	Risk events /100 user-days
Training cohort	post	0.762	0.157	0.058	17.82	3.33

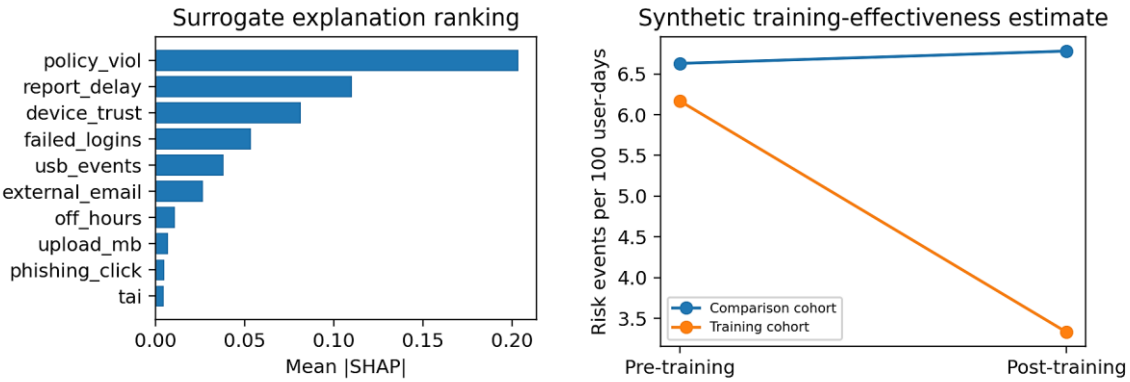

Fig. 4. *Surrogate explanation ranking and synthetic training-effectiveness estimate.*

8. Discussion

The findings illustrate the value of separating detection, explanation and action. Isolation Forest is useful for scalable filtering, but its standalone threshold produces too many false positives in the synthetic experiment. The Bi-LSTM captures temporal dependencies well, but a deep model alone does not solve governance questions about explanation, drift or response proportionality. The hybrid score therefore operates as a decision-support layer rather than as a claim that one algorithm is universally superior.

The most important methodological contribution is the explicit connection between security awareness training and insider-risk measurement. Training is often justified by regulatory expectation or completion dashboards. This study treats training as an intervention that must be evaluated through behavioural outcomes. A financial institution adopting the framework could ask whether a training cohort shows lower phishing susceptibility, faster reporting and fewer policy breaches after controlling for role, privilege and baseline risk. This is more defensible than reporting that 98% of staff completed a module.

Explainability is also treated pragmatically. SHAP values cannot prove intent and should not be used as a disciplinary shortcut. Their value lies in triage: they show why an alert is high-risk, which logs to review first and whether the model is focusing on meaningful behaviours. In regulated environments, this evidence should be stored with the alert, analyst decision and final outcome to support audit and model improvement.

The framework is compatible with Zero Trust because it produces dynamic, evidence-based access decisions. Low-risk behaviour can proceed without friction. Medium risk can trigger step-up multi-factor authentication or closer monitoring. High risk can restrict privileged operations, isolate sessions or escalate to the SOC. This tiered approach avoids treating every anomaly as malicious while still reducing the window of opportunity for damaging insider activity.

9. Implications for practice and research

For practice, EAHAI suggests a phased adoption path. Institutions should first map telemetry sources and data-protection constraints, then pilot the model on pseudonymised historical data, then evaluate analyst workload and explanation quality before

any live enforcement. Security awareness teams should be connected to SOC analytics so that training priorities are informed by actual behavioural patterns rather than generic annual content. Model outputs should be integrated with SIEM, SOAR and identity governance platforms only after calibration and policy approval.

For research, the framework creates testable hypotheses. Future studies can compare fixed thresholds with drift-adaptive thresholds, assess whether SHAP explanations reduce analyst decision time, examine whether awareness interventions lower specific behavioural indicators, and test whether risk-tiered Zero Trust responses reduce harm without producing excessive friction. Multi-institutional studies would be especially valuable, but they require privacy-preserving methods, common feature schemas and clear governance agreements.

The paper also has implications for responsible AI. Insider-risk analytics can affect employees' rights, reputation and working conditions. It should therefore be designed with transparency, proportionality, access controls, human review and appeal pathways. NIST AI RMF principles of validity, reliability, accountability and transparency are not optional additions; they are conditions for legitimate deployment [14].

10. Limitations and future work

The main limitation is that the empirical evaluation is synthetic. Synthetic telemetry is useful for reproducibility and confidentiality, but it cannot represent all organisational realities of a financial institution. The reported performance values should therefore be treated as proof-of-concept outputs, not as expected production performance. A real deployment may encounter missing logs, role changes, adversarial adaptation, incomplete labels, seasonal activity, union or works-council concerns, and regulatory restrictions on employee monitoring.

A second limitation is that the training-effectiveness component is simulated. The framework defines instruments and analytical methods, but it does not replace a controlled field study. Future research should validate the awareness scale, measure reliability on real participants, and assess whether behavioural changes persist beyond the immediate post-training period. Difference-in-differences and interrupted time-series designs are preferable to simple before-and-after comparisons when operationally feasible.

A third limitation concerns explainability. SHAP provides useful feature attributions, but explanations can be unstable when features are correlated or when background data are poorly chosen. Future work should compare SHAP with counterfactual explanations, evaluate analyst comprehension experimentally and test whether explanation review improves incident outcomes. Finally, adversarial robustness should be examined because insiders may learn to mimic normal behaviour or manipulate low-weight features.

11. Conclusion

This paper has presented EAHAI, an explainable adaptive hybrid AI framework for insider threat detection in financial institutions. The framework integrates efficient anomaly filtering, temporal behavioural learning, SHAP-based explanations, adaptive risk scoring, security awareness training assessment and Zero Trust decision-making. Its scientific novelty lies in the integration of detection, explanation, training-effectiveness measurement and policy response within a single replicable model. The synthetic case study shows how results can be generated transparently without exposing confidential banking data, while the methodology specifies the instruments, validation procedures and metrics required for institutional replication. Before live adoption, the framework should be validated on ethically approved, pseudonymised operational data and assessed for accuracy, fairness, analyst usability, employee rights and compliance. Used responsibly, EAHAI can help financial institutions move from reactive rule-based monitoring towards explainable, adaptive and evidence-led insider-risk management.

Declarations

Ethics and data protection: The proof-of-concept uses synthetic records only and does not

contain personal data. Any future institutional validation must obtain ethics approval, complete data-protection impact assessment, minimise personal data, apply pseudonymisation and preserve human review for high-impact actions.

Data availability: The results reported in this manuscript were produced from a deterministic synthetic generator using seed 42. No production financial-institution data or confidential CMU CERT raw files were used.

Funding: No funding information has been provided. Authors should update this statement before submission.

Conflict of interest: The author(s) declare no known competing interests. This statement should be confirmed before submission.

Declaration of generative AI and AI-assisted technologies: AI-assisted tools were used to support language refinement, formatting and internal consistency checks. The author(s) remain responsible for the scholarly content, interpretation, references, ethical compliance and final submission decisions.

References

[1] Cappelli, D.M., Moore, A.P., Trzeciak, R.F., 2012. The CERT Guide to Insider Threats: How to Prevent, Detect, and Respond to Information Technology Crimes. Addison-Wesley, Boston.

[2] Cummings, A., Lewellen, T., McIntire, D., Moore, A.P., Trzeciak, R.F., 2012. Insider Threat Study: Illicit Cyber Activity Involving Fraud in the U.S. Financial Services Sector. Software Engineering Institute, Carnegie Mellon University, CMU/SEI-2012-SR-004.

[3] Greitzer, F.L., Frincke, D.A., 2010. Combining traditional cyber security audit data with psychosocial data: towards predictive modeling for insider threat mitigation, in: Probst, C.W., Hunker, J., Gollmann, D., Bishop, M. (Eds.), Insider Threats in Cyber Security. Springer, Boston, pp. 85-113. https://doi.org/10.1007/978-1-4419-7133-3_5.

[4] Chandola, V., Banerjee, A., Kumar, V., 2009. Anomaly detection: A survey. ACM Computing Surveys 41(3), Article 15. https://doi.org/10.1145/1541880.1541882.

[5] Liu, F.T., Ting, K.M., Zhou, Z.-H., 2008. Isolation Forest, in: 2008 Eighth IEEE International Conference on Data Mining. IEEE, pp. 413-422. https://doi.org/10.1109/ICDM.2008.17.

[6] Liu, F.T., Ting, K.M., Zhou, Z.-H., 2012. Isolation-based anomaly detection. ACM Transactions on Knowledge Discovery from Data 6(1), Article 3. https://doi.org/10.1145/2133360.2133363.

[7] Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Computation 9(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735.

[8] Graves, A., Schmidhuber, J., 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks 18(5-6), 602-610. https://doi.org/10.1016/j.neunet.2005.06.042.

[9] Tuor, A., Kaplan, S., Hutchinson, B., Nichols, N., Robinson, S., 2017. Deep learning for unsupervised insider threat detection in structured cybersecurity data streams. arXiv:1710.00811.

[10] Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems 30.

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

[12] 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.

[13] Rose, S., Borchert, O., Mitchell, S., Connelly, S., 2020. Zero Trust Architecture. NIST Special Publication 800-207. https://doi.org/10.6028/NIST.SP.800-207.

[14] National Institute of Standards and

Technology, 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. https://doi.org/10.6028/NIST.AI.100-1.

[15] National Institute of Standards and Technology, 2003. Building an Information Technology Security Awareness and Training Program. NIST Special Publication 800-50. https://doi.org/10.6028/NIST.SP.800-50.

[16] ISO/IEC, 2022. ISO/IEC 27001:2022 Information security, cybersecurity and privacy protection - Information security management systems - Requirements. International Organization for Standardization, Geneva.

[17] ISO/IEC, 2022. ISO/IEC 27002:2022 Information security, cybersecurity and privacy protection - Information security controls. International Organization for Standardization, Geneva.

[18] Parsons, K., McCormac, A., Butavicius, M., Pattinson, M., Jerram, C., 2014. Determining employee awareness using the Human Aspects of Information Security Questionnaire (HAIS-Q). Computers & Security 42, 165-176. https://doi.org/10.1016/j.cose.2013.12.003.

[19] Safa, N.S., Von Solms, R., Furnell, S., 2016. Information security policy compliance model in organizations. Computers & Security 56, 70-82. https://doi.org/10.1016/j.cose.2015.10.006.

[20] Van Niekerk, J., Von Solms, R., 2010. Information security culture: A management perspective. Computers & Security 29(4), 476-486. https://doi.org/10.1016/j.cose.2009.10.005.

[21] Kruger, H.A., Kearney, W.D., 2006. A prototype for assessing information security awareness. Computers & Security 25(4), 289-296. https://doi.org/10.1016/j.cose.2006.02.008.

[22] 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.

[23] Chalapathy, R., Chawla, S., 2019. Deep learning for anomaly detection: A survey. arXiv:1901.03407.

[24] LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436-444. https://doi.org/10.1038/nature14539.

[25] Shapley, L.S., 1953. A value for n-person games, in: Kuhn, H.W., Tucker, A.W. (Eds.), Contributions to the Theory of Games II. Princeton University Press, Princeton, pp. 307-317.

[26] Hevner, A.R., March, S.T., Park, J., Ram, S., 2004. Design science in information systems research. MIS Quarterly 28(1), 75-105. https://doi.org/10.2307/25148625.

[27] Cronbach, L.J., 1951. Coefficient alpha and the internal structure of tests. Psychometrika 16, 297-334. https://doi.org/10.1007/BF02310555.