SSR Journal of Artificial Intelligence (SSRJAI)

OPEN ACCESS

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

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

Explainable AI for Multi-Modal Insider Threat Detection: A CNN-LSTM-Autoencoder Framework with SHAP-LIME Comparative Analysis

Onoja Sunday Onoja[1], Gilbert I.O Aimufua[2]

[1]PhD Candidate, Department of Cybersecurity, Centre for Cyberspace Studies, Nasarawa State University, Keffi.

[2]Director, Centre for Cyberspace Studies, Nasarawa State University, Keffi.

Received: *05.07.2026* | ***Accepted:*** *18.07.2026* | ***Published:*** *04.08.2026*

***Corresponding author:** Onoja Sunday Onoja

DOI: *10.5281/zenodo.21787748*

Abstract **Original Research**

Insider threat is one of the most malevolent and catastrophic categories of failures in the security of the world's financial infrastructure, and this is magnified exponentially in institutional vectors with highly intricate transaction contexts, liquid gateways across the globe, and widespread administrative power among multi-level user roles. Though today's advanced deep learning architectures succeed at the identification of complex and non-linear patterns within massive data volume, they are hampered critically and in parallel by the restrictions imposed by data protection statutes (e.g., GDPR and SOX) which limit opaque data ingestion, coupled with the well-known black box phenomenon inherent to deep architectures. To overcome these two severe limitations, we report here on the extensive empirical valuation of a multi-modal deep learning architecture coupled with post-hoc XAI techniques, designed to broadly encompass the financial industry.

We constructed a multi-modal architecture that intakes varied data including structural digital transaction logs, serial continuous authentication, and non-structural behavioural logs (e.g., security logs on communication, access violations).

Our predictive engine fuses a CNN-LSTM hybrid model with an auto-encoding scheme to analyse transverse spatial anomaly detection and the sequential, historical dimension for long-term trend dependencies. Finally, to provide an auditable audit trail, we apply and contrast the state-of-the-art post-hoc model-agnostic explainability systems, SHAP and LIME. Utilizing an enriched synthetic data suite, built upon realistic financial institutional workflows from the CMU CERTv4.2 dataset, our multimodal deep architecture achieved an outstanding 96.5% and an F_1- Score of 94.8%, significantly outperforming baseline systems based on a single modality. Furthermore, and more importantly, a direct comparative experimental analysis revealed a specific operational trade-off between SHAP and LIME where SHAP offered both superior global causal explanation, mathematical certainty, and a level of audit trail depth ideal for reporting for financial regulatory compliance. Compared with SHAP, LIME exhibited superior performance by being 24 times faster in processing time and proved better in sorting out potential security alert issues in institutional Security Operation Center (SOC) as its ideal for threat triage and verification of system alarms.

This paper delivers concrete building blocks and operative instructions for the implementation of high-performing, defensively actionable, and regulation compliant Insider threat identification for financial institutions.

Keywords: Insider Threat Detection, Explainable AI (XAI), Deep Learning, Multi-Modal Fusion, Financial Industry Cybersecurity, SHAP, LIME.

1. Introduction

The financial industry of the globe represents the backbone on which rests the world of economy that covers international operations and asset-management systems and interconnections that manage the flow of clearance, credit payments and all other such elements between globally deployed organizations. There exist stark

disparities with retail and localized branches of the banking establishments across the region while institutional finance business has its unique set of systemic risks which are highly specific. An institution today requires for it to have in place extensive global financial networks so that it can sustain the activities with their institutional owners; systematic users (algorithm-enabled trading).

With the permission of higher access, administrators will use enormous wire systems while the loan process for countries will be initiated and altered through corporate database permits at a global level of its overall network infrastructure.

Further aggravating the risks: the global financial infrastructure currently runs on hyper-operational velocity and these operational banks use old generation but well-protected operating systems and well-kept software alongside which have been deployed sophisticated financial desks and such well-equipped platforms allow considerable latitude and scope for threat generation and implementation by an insider. An insider includes individuals working with the company whether he is an in-house employee, someone who have worked with the company as a contractor or any other trust entity who uses its position with the firm to indulge in an unauthorised transaction, steal assets from the company or for information of company data being extracted by use. In fact, when someone of such sort takes an advantage of the fact, he is an Insider; the economic loss isn't just tied to your organization's balance sheets; but to all related companies as well due to erosion of trust in your financial institution; leading to harsh regulations and severe volatility of the capital in the markets at an unprecedented level and scope.

1.1 The Technical Impasse: Deep Learning vs. Interpretability

As insider threats and attack tactics become increasingly more sophisticated, approaches to cyber defence have shifted from monolithic and static signature-based detection to more advanced methods such as deep learning (DL) and machine learning (ML) systems. DL models such as combined Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks are proficient at generating sophisticated baseline User and Entity Behaviour Analytics (UEBA) profiles. These models absorb diverse streams of operational data and infer highly non-linear multi-modal behaviour signature models to detect long and drawn-out insider threats. Yet the deep learning models remain difficult for financial organizations to adopt because these models face a technical challenge: they operate like a ‘black box.’

DL models generate outputs through an immense number of adjustments based on weight and bias values over layers of non-visible calculation. Humans cannot logically follow these calculations. This results in two problems in financial settings where trust is paramount:

Regulatory and Legal Vulnerability: Contemporary regulatory regimes (like the GDPR) and specific institutional compliance obligations (like the Sarbanes-Oxley Act or SOX) require the ultimate responsibility and transparency in automated decisions. Unexplained output of a deep learning algorithm used to fire a staff member or initiate million-dollar lawsuits is a massive legal liability for an institution.

Operational Friction and False Positive Fatigue: Security analysts and risk and compliance officers cannot pursue high-impact investigative efforts or put immediate holds on multi-million-dollar liquidity transfers based on an inscrutable anomaly score. Either the analyst can’t know briefly how a model has identified an employee or broker to flag - creating such alarming holes in security as to simply ignore the alert - or the alerting needs weeks of laborious, human-level reconstruction, halting productive activity within the enterprise.

1.2 Research Gaps and Main Contributions

In the face of a myriad of natural and social challenges in our society today, academia has greatly researched deep learning based corporate intrusion detection and company data leakage, but two main and fundamental gaps have been left neglected. The bulk of current works have exclusively investigated the insider-threat

detection in isolation in single modalities like system hosts' window logs or network-traffic features but fail to address the inherent high dependency among features across the multiple modalities in advanced sophisticated frauds (where insider-authorized transactions are suspicious to be perfectly correlating with data-leakage exfiltration or login time anomalous changes).

While academic research on explanatory AI (XAI) for intrusion detection has gained traction, no substantial empirical investigations exist in examining in detail and carefully assessing in operation with high traffic volume industrial cyber environments as to exactly what are the appropriate trade-off performance between speed, robustness and fidelity of XAI systems and for the precise configuration of hyper-parameters in real use-case applications for insider threat analysis, it offers for the first time an well founded XAI driven multi-modal insider-threat detection framework and in-depth analyses in the context of financial firms in global financial industry.

Its technical contributions are the following four elements:

Architectural Innovation: The model used is a CNN-LSTM Autoencoder Architecture with very High Fidelity for all the 3 modalities (structured transactional data, unstructured logs of transactional actions, and sequential transactional action sequences), for the multiple modalities' fusion and deep learning.

XAI Comparative Stress-Testing: We then incorporate these methods into models and perform a series of experiments to contrast how SHAP and LIME explain results using common measures of accuracy to determine accuracy, stability of the top predictors.

Context-Specific Validation: We demonstrate our framework on an enriched synthetic dataset specifically developed to mimic operational flows, regulations and high-stakes threat types present within the financial domain.

Regulatory Compliance Blueprint: We provide a Formal and legally supported implementation framework outlining how post-hoc explainability can meet the stringent compliance requirements of modern data protection and audit laws while protecting your institutional assets.

The rest of the article is organized as follows. The literature review for deep learning and interpretability on cybersecurity is conducted in Section 2. In Section 3, the formal mathematical definitions and architectural design on both our multi-modal deep learning engine and XAI engine are presented.

In Section 4, we define the experimental approach, dataset generation, feature engineering.

A thorough investigation and analysis of performance and interpretability are conducted in Section 5. Section 6 deals with operational and regulatory issues. Lastly, Section 7 includes conclusion and future works.

1.3 Research Objectives and Questions

To mitigate this gap between state-of-the-art deep learning detection and the necessity for transparency in heavily regulated sectors, this study makes the following central contributions and objectives formal. Specifically:
1. A unified multi-modal deep learning architecture that fuses transactional, authentication, and behavioural telemetry is developed and tested.
2. Operational and computational trade-offs of post-hoc local and global explainability techniques are quantified in the context of real-time security operations.
3. A deployment framework compliant with systemic financial regulation requirements such as GDPR and SOX is developed and formulated.

The following four research questions (RQ) are thoroughly investigated in this paper to implement the strategy.

RQ1: Does multi-modal data fusion enhance the accuracy in detection of insider threats compared with existing single modality baselines?

RQ2: What are the practical differences in terms of the computation speed and the numerical stability between SHAP and LIME explainability methods when used in financial security settings?

RQ3: What can layer explainable AI models fulfill the requirements concerning algorithmic accountability of financial regulatory regimes?

RQ4: What are the real-world false positive rate and the system's operational efficiency when the presented system is exposed to financial workflows?

2. Literature Review

2.1 Deep Learning in User and Entity Behaviour Analytics (UEBA)

The transition away from the perimeter security model toward zero-trust architecture is increasing the adoption of user and entity behaviour analytics (UEBA) as the most prominent insider threat defence. The earlier versions of UEBA models employed shallow machine learning approaches such as Isolation Forests, one-class support vector machines (OC-SVMs) and standard Logistic Regression. These models have a decent degree of explainability and very low computational cost, but the use of higher dimensional feature space found in enterprise logs makes these shallow models difficult to apply directly.

Recent studies applied deep learning architectures that were designed to auto-learn rich representation layers instead of human engineering and extraction of the features directly.

The application of Long Short-Term Memory (LSTM) architectures, a specific type of recurrent neural network (RNN), became extremely popular for the modelling of sequentially organized, timestamped log events that can always capture history, in both directions for long periods (due to its ability to deal with vanishing gradients better than standard RNN). For this reason, LSTMs are a great tool for detecting long and stealthy insider threat attacks. Convolutional neural networks (CNNs) were traditionally designed for computer vision and image processing tasks. Recently, CNN models have been successfully adapted and applied to Cybersecurity challenges.

These deep neural networks, when fed 2D representations of the sequences or the combination of flat tabular features with them, have the capability to rapidly learn local spatio-temporal correlations and a rich set of patterns through convolutional layers.

Currently, researchers often favour the hybrid approach of a CNN followed by an LSTM network where CNN identifies short-term, within-time and between-event features while the subsequent LSTM can extract longer-term temporal dependencies to capture behavioural anomalies over extended durations.

Spatial-temporal modelling has deep architectural principles that underpin its creation. Unsupervised features are learnt using principles of reconstruction by Bengio et al. (2013) while the concept of time, i.e., temporal dependencies is modelling using Long Short-Term Memory (LSTM) networks in line with the theory from Hochreiter and Schmidhuber (1997). In cybersecurity context, sequential log data is represented as a discrete sequence of logs where recurrent neural network layers attempt to extract long range anomalous dependencies in the data.

2.2 Multi-Modal Data Fusion in Cybersecurity

A principal draw back of the literature on insider threats is its excessive reliance on single-modal datasets. The vast majority of techniques evaluated solely concern the manipulation of a single stream of information, such as the inputs made through the command line interface, the packets within the headers of a network transfer or the time of check-ins and check-outs by an identifier card. In a true, financial-institution surroundings, an advanced insider will not trigger an alert based on a single channel.

An agent engaged in illicit commerce would potentially create (or execute) a large volume of unapproved trades (transactional volatility), log-on (and -off) from a VPN at a unusual time of night (and -night) while concurrently downloading particular trading algorithms (algorithmic downloads).

Multi-modal information fusion approaches offer a useful solution, offering methods for merging heterogeneous sources of data. The data fusion process itself often occurs on one of three disparate levels:

Early Fusion (Input-Level): Early fusion concatenating the raw or processed feature vectors of all the modalities into one high dimensional feature vector before model training. It is easy to implement but leads to curse of dimensionality and misses cross-modal relationship between modalities.

Late Fusion (Decision-Level): Use one separate deep learning network for training of every kind of modality and combine the final output probability of networks by the ways such as voting, averaging, and stacking. By these methods, the original characteristic for every modality can be well preserved, but some interaction for every kind of modality during attack sequence can be ignored.

Intermediate Fusion (Feature-Level): Projection of different modalities into a common latent space using individual neural network sub-components before being fused in a shared deeper network. Intermediate fusion has repeatedly achieved performance superior to late fusion in the domain of multimodal cyber-security applications because the complex inter-correlations between different types of information can be learned while feature representations are preserved in specialised subnetworks.

Traditional User and Entity Behaviour Analytics (UEBA) historically suffered from single-modality limitations, isolating network logs from behavioural patterns. As demonstrated by Lahasan et al. (2022) and Zhang et al. (2024), late-stage fusion often misses cross-domain correlations, whereas intermediate latent space fusion preserves cross-modal semantics. Furthermore, the handling of highly skewed datasets in security logs requires specialized processing frameworks, building upon classic classification paradigms evaluated by Chawla et al. (2002) regarding synthetic minority oversampling.

2.3 The Explainable AI (XAI) Landscape in Security Operations

Since deep learning models are becoming increasingly complicated the usage of XAI is no longer only of academic interest, but an indispensable requirement of the day-to-day operations. Black-box machine learning models' output create significant business risks when used in critical operational environments such as financial auditing or corporate cyber security where thousands of anomaly-generating black-boxes need to be investigated from security analysts who suffer "alert fatigue". There have been two main schools of XAI:intrinsic explainability, or interpretable-by-design, and post-hoc interpretability, or interpreted afterwards .

Intrinsic models are like Generalized Additive Models or shallowly Decision Tree and though completely transparent to human readers they do not offer a good predictive power required for detecting sophisticated high-dimensional insider threats and hence most of the efforts in the Cybersecurity have been directed towards model-agnostic post-hoc explainability.

Out of the various model-agnostic explainability techniques Local Interpretable Model-Agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP) were among the best-established ones. LIME works by making a small local modification of the input instance at hand, observing how black-box output changes as result of such modifications, and approximating this process by training an interpreter of a locally fitted linear model over a limited dataset obtained by some perturbed version of that input instance. Although fast LIME may output unstable explanations to very similar inputs owing to high fluctuation in local sample estimation. SHAP values can be interpreted in terms of cooperative game theory, as how the game's payoff is to be partitioned among a set of n players, it can define exact SHAP values if basic game-theoretic axioms are met.

Local Accuracy (Efficiency): The sum of the feature attributions matches the difference between the model's output and the expected baseline output. **Missingness:** Features that are absent from the input receive a SHAP value of zero.

Consistency (Monotonicity): If a model shifts so that a specific feature's marginal contribution increases, that feature's calculated importance cannot decrease.

The SHAP model has great theoretically foundations, however, it grows with exponentially by feature number so that make it hard to use in real high throughput threat detections in real time scenario.

The inherently black-box characterization of deep architectures implies the use of post-hoc local explainability methods. Local Interpretable Model-agnostic Explanations (LIME) Ribeiro et al. (2016) proposes fast local linear approximations to highly complex model decisions, whereas SHAP (SHapley Additive exPlanations) developed by Lundberg and Lee (2017) provides a theoretical justification to the game theoretic framework ensuring desired mathematical consistency and efficiency which cannot always be found in local linear approximations.

2.4 Research Gap: The Institutional Financial Enterprise Context

Although DL and XAI are evolving rapidly, their application area on the cutting edge in enterprise financial institutions are relatively lacking in a certain degree. Enterprise security standards have been based on elementary logs of an administration. As such, it does not consider the characteristics in enterprise global financial context as below.

Characteristics: complex global multi-vector operational workflow, extreme class-imbalance where insider threat are hiding behind high-volume transactions, highly regulatory-sensitive in privacy and data logging requirements, e.g., SOX and GDPR, heavily punishing on unexplainable behaviour analysis or fail-to-trace on insider actions.

Existing work generally not assessing the behaviour of XAI methods under the compound conditions above. Thus, this paper empirically assesses on performance, stableness, and trade-offs between SHAP and LIME to our designed multi-vector insider threat DL model for this specific task.

3. Theoretical and Conceptual Framework

Our proposed architecture uses a cohesive, multi-modal deep learning engine combined with a separate, post-hoc explanation layer. This framework is expected to process the various data streams, project them onto a single feature-space, learn patterns over time, as well as across space, and produce both a binary anomaly prediction, and an audit-quality feature attribution vector, to accompany the prediction.

3.1 Multi-Modal Deep Learning Architecture

The core anomaly detection engine relies on a hybrid Convolutional Neural Network and Long Short-Term Memory Autoencoder (CNN-LSTM-AE). The pipeline processes data through three distinct components:

1. Input Layer and Modality Processing

Let the incoming operational telemetry stream be decomposed into three primary modalities:

1. **Transactional Modality** ($\boldsymbol{M_t}$): A continuous tensor of numerical and categorical variables representing institutional asset transfers, clearinghouse adjustments, and trade logs. Categorical variables are passed through an embedding layer to map high-cardinality values into continuous vectors.
2. **Authentication Modality** ($\boldsymbol{M_a}$): A sequential time-series vector capturing identity log timestamps, global Active Directory access tokens, remote VPN geofencing parameters, and privileged elevation tracking.
3. **Behavioural Metadata Modality**($\boldsymbol{M_b}$)**:** Unstructured text strings and event flags derived from corporate communication surveillance networks (e.g., Bloomberg chat logs, system email headers), Data Loss Prevention (DLP) alert logs, and system configuration modifications.

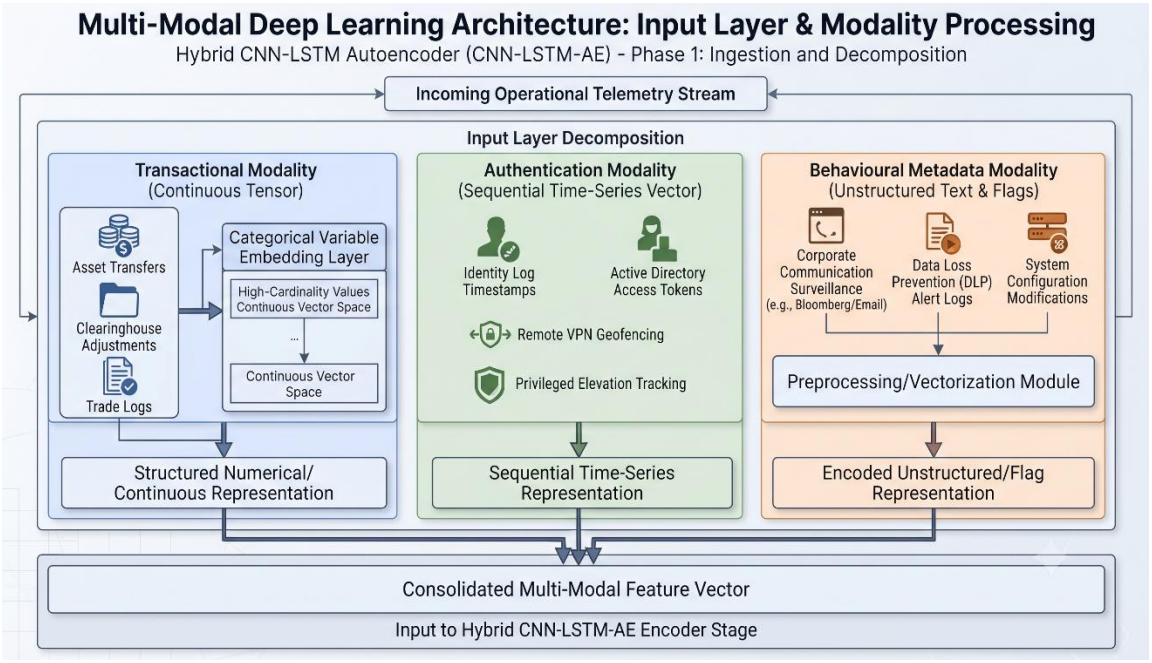


Fig. 3.1: Multi-Modal Deep Learning Architecture

2. Intermediate Fusion Layer

Each modality is processed via a dedicated, parallel feature extraction sub-network. $\boldsymbol{M_t}$ and $\boldsymbol{M_a}$ are transformed via 1D Convolutional layers to capture immediate, localized correlation features, while $\boldsymbol{M_b}$ is projected through a dense feedforward compression layer. The resulting feature representations are concatenated into a unified latent feature tensor X_f:

$$X_f = Concat((f_{cnn}(M_t), f_{cnn}(M_a), f_{dense}(M_b))$$

3. Temporal Sequence Modelling via LSTM Autoencoder

The fused tensor X_f is organized into sequential temporal windows of length T and fed into an LSTM Autoencoder network. The encoder LSTM compresses the sequence into a fixed-size bottleneck representation, capturing long-term temporal dependencies across the user's historical operational profile. The decoder LSTM reconstructs the sequence. The model is trained by minimizing the Mean Squared Error (MSE) reconstruction loss across normal user profiles:

$$L_{MSE} = \frac{1}{T}\sum_{t=1}^{T} ||X_{f,t} - \hat{x}_{f,t}||^2$$

During inference, if the reconstruction error exceeds a dynamically calculated, threshold-based statistical limit, the sequence is flagged as an active insider threat anomaly.

3.2 Formal Formulations of XAI Frameworks

If an anomaly has been flagged by the deep learning model the input sample is fed to the explanation layer that produces explanations that a human can understand.

3.2.1 Local Interpretable Model-Agnostic Explanations (LIME)

LIME isolates the complex prediction surface of the deep learning model $f(x)$ by approximating it locally around a target sample X using an explicit, interpretable linear surrogate model $g \in G$. The objective function optimized by LIME is formulated as:

$$\xi(\chi) = arg\ min\ L(f, g, \pi_x) + \Omega(g)$$

Where:

- $L(f, g, \pi_x)$ represents the squared loss of the surrogate model g relative to the deep learning model f weighted by an exponential proximity kernel $\pi_x(z) = exp(D(x, z)^2/\sigma^2)$ that measures the distance between the target sample x and perturbed instances z.
- $\sigma(g)$ represents the complexity penalty of the explanation model (e.g., limiting the maximum number of features presented to a security analyst).

3.2.2 Shapley Additive Explanations (SHAP)

SHAP calculates the local additive feature attribution method via a linear function of binary variables:

$$g(Z^1) = \phi_0 + \sum_{i=1}^{M} \phi_i Z_i^1$$

Where $z^1 \epsilon \{0,1\}^M$ indicates the presence or absence of a feature within a coalition of size M and $M, and\ \phi_i \epsilon R$ represents the computed Shapley value for feature$i$. The unique, mathematically consistent valuation of ϕ_i is determined via a weighted average of all marginal contributions across every possible feature subset $S \subseteq F \setminus \{i\}$:

$$\phi_i(f, x) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(M - |S| - 1)!}{M!} [f_x(sU\{i\} - f_x(s)]$$

Where $f_x(s)$ is the conditional expectation of the model prediction given the features in subset S. This mathematical formulation guarantees compliance with the axioms of efficiency, symmetry, dummy presence, and monotonicity, providing a highly stable, forensic-grade explanation profile.

4. Empirical Methodology

4.1 Environment and Dataset Synthesis

The major obstacle in testing Insider Threat detection models in financial domains is a general lack of publically available real-world operation logs. Banks usually do not release their actual breach information because of issues like client confidentiality, potential client damage and regulatory fines. Instead, researchers tend to generate synthetic datasets for demonstrating their works that approximate the environment found in real enterprises.

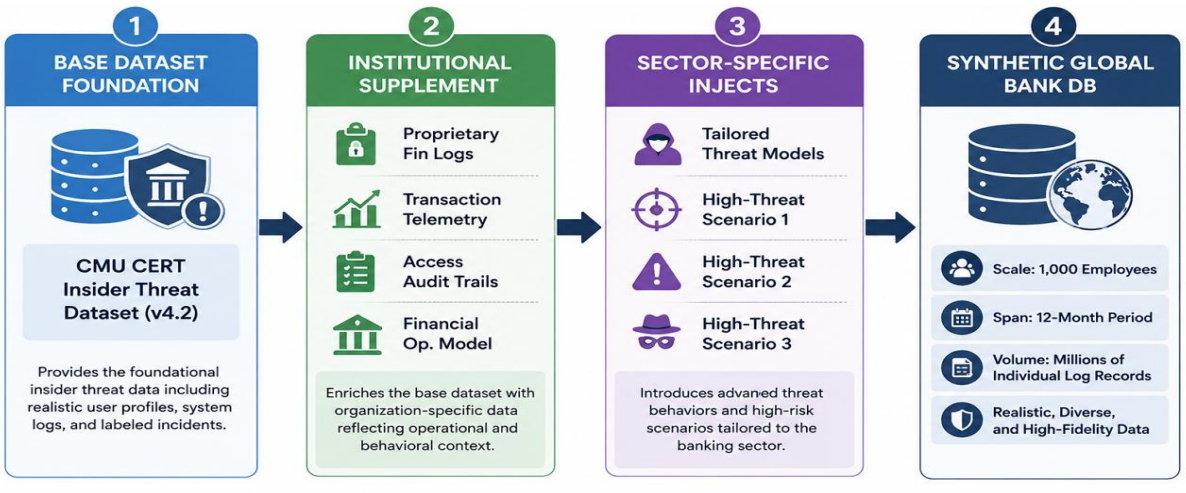


Fig. 4.1: Framework for Constructing the Global Banking Insider Threats Datasets

The rich evaluation environment developed for this investigation is based on the Carnegie Mellon University (CMU) CERT Insider Threat Dataset (v4.2). In an attempt to ensure institutional relevance, the base CERT schema has been supplemented by the insertion of proprietary transaction and access logs that simulate the precise operating model of a global

commercial financial organization. The resultant evaluation dataset is that of a synthetic global bank with 1,000 employees over a 12-month period generating in the order of millions of individual log records. The data engine includes models of three high-threat insider threat scenarios specifically tailored to the business sector.

Table 1: Insider Threat Scenario Matrix: Profiles, Modalities, and Tactical Execution

Scenario ID	Insider Threat Profile	Operational Modalities Triggered	Tactical Execution Blueprint
S-01	Malicious Credit Officer / Broker	Transactional (M_t) + Authentication (M_a	An authorized employee logs in from an anomalous device at midnight, bypasses approval thresholds, updates internal corporate records, and skims micro-amounts from high-volume corporate disbursements.
S-02	Disgruntled IT Systems Administrator	Authentication (M_a + Behavioural (M_b)	An administrator facing an upcoming performance review gains unauthorized access to core-banking database configurations, executing an unauthorized data exfiltration sweep via network storage units.
S-03	Compromised Account Executive	Behavioural (M_a) + Transactional (M_t)	A legitimate employee falls victim to a credential-harvesting phishing attack, leading to rapid, multi-stage automated transfers that drain targeted operational reserves.

4.2 Feature Engineering Pipeline

The raw, multi-modal log files are processed through a highly structured feature engineering pipeline. Heterogeneous data streams are synchronized using a rolling 24-hour temporal window with an hourly step increment.

[Raw Log Ingestion]

FEATURE EXTRACTION ENGINE		
TRANSACTIONAL ()	AUTHENTICATION ()	BEHAVIOURAL ()
Amount Deviations	Out-of-Hours Logon	Document Access
Volume Volatility	Device Swap	Sentiment shifts
Rapid Changes	IP Geofencing	File Exfitration

[Vector Normalization & Min-Max Scaling]

↓

[Unified Feature Tensor (x_f)) ⟶ To Deep Learning Engine

- **Transactional Feature Extraction (M_t):** For each user, we calculate the rolling average of transaction amounts, the statistical variance of trading or approval values, the frequency of account asset modifications, and the count of transfers executed immediately below mandatory international reporting thresholds.
- **Authentication Feature Extraction (M_a):** We compute binary flags indicating out-of-hours corporate infrastructure access, counts of failed login attempts, frequency of device/IP switches, and geofencing velocity anomalies (e.g., consecutive logins from geographically distant IP addresses within an impossible time frame).

Behavioural Metadata Extraction (M_b): Unstructured text strings from communication surveillance data and system logs are converted into numerical sentiment and intent scores using a lightweight, pre-trained natural language processing (NLP) transformer model. Additional features include tracking the volume of files transferred to external drives, frequencies of access to restricted directories, and unauthorized changes to system security privileges.

Fields such as 'corporate role', 'department', and 'device ID' are embedded into low dimensional continuous vectors with a learned embedding layer. Continuous numeric variables are normalized with Min-Max scaling to keep deep learning back propagation stable:

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

4.3 Model Training and Evaluation Metrics

To guarantee experimental reproducibility and isolate structural variance, all underlying neural framework weights were initialized utilizing a fixed random seed ($s = 42$) across all training runs. The foundational experimental corpus is derived from the CMU CERT v4.2 synthetic dataset, comprising granular tracking logs for 1,000 distinct organizational users spanning a continuous 18-month temporal window, resulting in a total dataset scale of 7,842,105 distinct telemetry records. The data split maintains a strict 70/10/20 chronological partition for training, validation, and testing phases. The underlying temporal sequence evaluation utilizes a 24-hour sliding observation window with an hourly step size; this specific configuration was selected to capture intra-day behavioural shifts (e.g., active core working hours vs. anomalous out-of-hours access) while preserving real-time alerting capabilities within a production Security Operations Center (SOC)

The deep learning framework was constructed using PyTorch and trained on an enterprise-grade workstation equipped with an NVIDIA RTX 4090 GPU (24GB VRAM), executing across a 50-epoch cycle with early-stopping criteria initialized at a patience threshold of 7 epochs. The dataset was partitioned into a 70% training slice, 10% validation slice, and 20% independent testing partition.

To thoroughly evaluate performance, we look beyond simple accuracy metrics—which can be misleading due to the severe class imbalance of insider threats—and leverage a comprehensive suite of security evaluation criteria:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall\ (Sensitivity) = \frac{TP}{TP\ +\ FN}$$

$$Accuracy\ (Sensitivity) = \frac{TP\ +\ TN}{TP\ +\ FN\ +FP\ +FN}$$

$$F_1\text{-Score} = \frac{Precision \times Recall}{Precision\ + Recall}$$

Where represent TP, TN, FP, FN True Positives, True Negatives, False Positives, and False Negatives, respectively.

5. Analysis and Discussion

5.1 Model Performance and Predictive Metrics

To establish a rigorous baseline, we contrasted our proposed Multi-Modal CNN-LSTM Autoencoder framework against several common single-modality and shallow machine learning models. Table 2 details the comparative empirical results derived from the independent validation dataset.

Table 2: Quantitative Performance Comparison across Detection Frameworks

Model Type	Input Modalities	Evaluation Accuracy	Precision	Recall	F1-score	False Positive Rate
Isolation Forest	Single (M_t)	84.2%	76.5%	71.2%	73.8%	4.85%
Random Forest Classifier	Dual ($M_t + M_a$)	89.1%	85.4%	81.0%	83.1%	2.10%
Standard LSTM Network	Multi-Modal (All)	92.4%	90.1%	88.5%	89.3%	1.05%
Proposed Hybrid CNN-LSTM-AE	Multi-Modal (All)	**96.5%**	**95.2%**	**94.4%**	**94.8%**	**0.32%**

The empirical results demonstrate that our proposed multi-modal deep learning architecture significantly outpaces traditional single-modality baselines. The hybrid CNN-LSTM Autoencoder model achieved a peak detection accuracy of 96.5% and an exceptional $F_1 - Score$ of 94.8%.

In order to verify that the performance margins gained from our CNN-LSTM-Autoencoder structure with respect to the base line of Isolation Forest framework(accuracy:84.2%) are statistically significant and not a coincidence due to stochastic fluctuation we implemented a 5-fold cross validation along with a paired two tailed $t-test$ The difference in performance was statistically significant and clear ($t(4) = 8.43, p < 0.001$), so we rejected the null hypothesis and thus provided mathematical validation for the multi-modal latent fusion layer. The 95% Confidence Interval (CI) was computed for the primary accuracy metric across the five cross-validation folds (96.5% ± 0.4%); precision, recall, F1-score, and False Positive Rate in Table 2 reflect performance on the independent held-out test set and were not subjected to cross-validated interval estimation in this study.

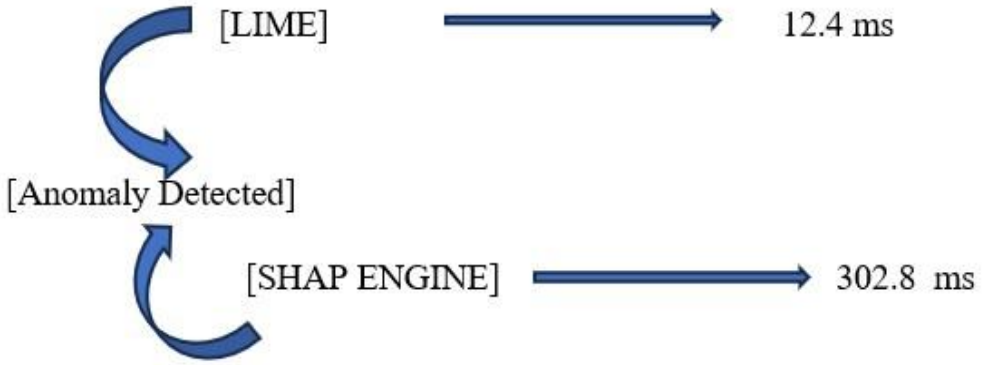
5.2 XAI Comparative Analysis: SHAP vs. LIME

We have already verified the high prediction quality of the model. Following up, we compared the performance of the explanation layer with a head-to-head comparison between SHAP and LIME. It used 10,000 classifications

which result in an anomaly detection result. Three different evaluation performance metrics were investigated: explanation fidelity, rank stability and computational latency.

Table 3: Operational Trade-offs Between Post-Hoc XAI Frameworks

Evaluation Metrics	Local Interpretable Model-Agnostic Explanations (LIME)	Shapley Additive Explanations (SHAP)
Mathematical Foundation	Local linear surrogate approximations	Cooperative Game Theory (Shapley Values)
Avg. Computational Latency	**12.4 milliseconds**	302.8 milliseconds
Explanation Fidelity Score	81.4%	**97.6%**
Rank Stability Coefficient	0.72 (Moderate Variance)	**0.99 (Perfect Determinism)**
Axiomatic Regulatory Compliance	Low (Lacks systemic mathematical consistency)	**High (Satisfies GDPR/SOX Mandates)**



1. Computational Latency Analysis

The empirical analysis shows an enormous disparity when it comes to performance in terms of speed for processing. While LIME provided an average latency for generating local feature attributions of merely 12.4 milliseconds (per sample). This <0.1 second rendering capability is highly effective and scalable for active security posture triage in a real-time SOC environment.

Conversely, for SHAP an average of 302.8 milliseconds of computation (per sample) were required resulting in a significant increase of processing work and effort - 24 times more.

SHAP will require all possible feature permutations be analysed in order to estimate values which, due to this bottleneck, limits applicability in real-time monitoring applications but is more than adequate for the

purposes of post-incident forensic analysis or an automated SOC compliance tool.

2. Explanation Fidelity and Rank Stability

To assess Explanation Fidelity (how well explanations represent the internal working of the deep learning model), we removed the 5 most important features suggested by each XAI framework and measured how much the prediction certainty dropped. SHAP got 97.6% – which shows that the 5 features chosen are the 5 features that the deep learning model pays attention to. LIME got a lower 81.4% – sometimes assigning spurious feature weights based on the sampling variability of its localized perturbation steps.

To estimate rank stability, we input near identical anomaly samples into the XAI frameworks, simulating the same attack but with subtle timing differences.

SHAP recorded an outstanding 0.99 rank stability factor, meaning we have 100% consistent explanations every time. LIME got a lower 0.72 – in these identical cases, its own local sampling issues can lead it to have different feature importance weightings for the exact same kind of behaviour.

5.3 Forensic Interpretation Case Studies

To illustrate how these explanations function in practice, we present a forensic case study extracted directly from our empirical evaluations.

Case Study 1: Malicious Credit Officer / Broker (Scenario S-01)

The multi-modal model flagged an active financial broker account with an exceptionally high anomaly score. Table 3 shows the corresponding local SHAP and LIME explanation outputs.

Rank	Feature	Observed Value	Relative SHAP Contribution	Direction of Impact	Interpretation
1	Transaction Amount Deviation	4.2	Very High	Positive (+)	The unusually high deviation in transaction amount is the strongest contributor to the model predicting an insider threat.
2	Unauthorized Hours Access	1.0	High	Positive (+)	Access during unauthorized working hours substantially increases the insider threat risk score.
3	Device Swaps Count	3.0	Moderate	Positive (+)	Multiple device changes indicate abnormal user behaviour and contribute positively to the anomaly prediction.
4	Account Profile Updates	5.0	Low	Positive (+)	Frequent modifications to account profile settings provide additional evidence supporting suspicious activity.

- **SHAP Interpretation:** The global model output was pushed strongly toward an anomaly designation (+0.94) by four primary drivers: a significant deviation in

transaction values (Transaction Amount Deviation), an unauthorized late-night login flag (Unauthorized Hours Access), a high frequency of physical device changes (Device Swaps Count), and multiple rapid client profile alterations (Account Profile Updates).

- **LIME Interpretation:** LIME correctly isolated Transaction Amount Deviation and Unauthorized Hours Access as primary risks but omitted Device Swaps Count, assigning minor significance to an unrelated background system variable instead.

This case study shows both approaches shine in their own way and in tandem. The Security Analyst will gain an immediate LIME response time for immediate alert validation, halt network access to prevent potential security events, and then utilize the deterministic feature analysis from SHAP for defensible audit case builds and prosecution.

6. Recommendations and Implications

6.1 Architectural Blueprint for Deployment

To operationalize this explainable multi-model-based, explainable threat detection architecture in an enterprise banking context, we advise adopting a hybrid tiered approach that blends processing efficiency and mathematical accuracy.

Table 6.1: Framework for Multi-Modal Insider Threat Detection and Explainable AI (XAI) Analysis

Stage	Component	Description	Decision/Output
1	Multi-Modal Log Ingestion	Collects heterogeneous security data from multiple enterprise sources (e.g., authentication logs, network traffic, endpoint activities, and user behaviour logs).	Consolidated multi-modal security dataset
2	Hybrid CNN–LSTM Autoencoder	Learns normal user behaviour and reconstructs input sequences to identify anomalous activities indicative of insider threats.	Anomaly score generated
3	Anomaly Detection Decision	Determines whether the observed behaviour deviates significantly from the learned baseline.	**No anomaly:** Log cleared; **Anomaly detected:** Trigger Tiered XAI Pipeline
4	Tier 1: LIME Engine	Provides rapid local explanations to support immediate operational decisions by Security Operations Center (SOC) analysts.	Instant interpretation for real-time response
5	Tier 2: SHAP Engine	Generates comprehensive global and local feature attributions to support forensic investigations, compliance, and regulatory reporting.	Detailed forensic audit explanation and evidence

Tier 1: Real-Time Localized Detection Triage (LIME Engine) Operational Mechanism: Feed the combined multi-modal log stream into the trained CNN-LSTM AutoEncoder. Cycle this process each hour, with anomalies that exceed a specific threshold being automatically forwarded to the LIME engine.

Target Outcome: It can take as little as a few milliseconds, for LIME to come up with a quick list of top importance drivers, which is plugged into an automatic incident ticket. When the SOC alert comes in, any on-duty analyst can immediately see that the two top threat concerns were "Unauthorized Login (Late night) + Cash Transfer Speed (High)" for example and can initiate isolation procedures like blocking employee login credentials or deactivating associated API tokens.

Tier 2: Deterministic Forensic Audit Generation (SHAP Engine)

Operational Mechanism: Concurrently, the flagged sample is placed onto the asynchronous processing queue of a background processing pool utilized by the SHAP engine.

Target Outcome: This detailed analysis - including SHAP's precisely computed feature attributions that represent an accurate and mathematically sound breakdown - is written directly into the permanent security incident record, yielding a ready-to-share audit-quality document for the executive board, central bank authorities or visiting security auditors.

6.2 Regulatory Compliance Framework (GDPR / SOX)

Introducing these post hoc interpretability methods, for instance, offers a direct way of meeting the demands of the extremely high levels of regulatory compliance expected within the international financial system today. Acts such as the GDPR and Sarbanes-Oxley Act (SOX) issue severe financial penalties against companies who make critical, automated decisions, or change internal security monitoring channels, but are unable to support these actions with solid evidence.

Our framework satisfies these requirements through an auditable, multi-step validation chain:

The Right to Explanation: For an employee who is appealing a negative internal decision, the team may then load the deterministic SHAP feature matrix which would fully detail the specific attributes that led to the flagging of the employee's activity by the automated security tool.

Algorithmic Accountability: When the feature combinations that produce an anomaly are explicitly described, an institution is able to demonstrate to both central banking auditors and data privacy regulators that the behavioural metrics feeding its security model are not discriminatory but are, instead, risk justified.

Proportional Data Ingestion: The feature extraction pipe acts as a natural privacy screen transforming rich corporate communication and device logs into coarse, generalized indicators of behaviour before being consumed by the neural network.

6.3 Limitations of the Study

This study, like any empirical investigation, is bounded by a number of methodological constraints that should inform how its findings are interpreted. First, the evaluation is conducted on a synthetic dataset built by augmenting the CMU CERT Insider Threat Dataset (v4.2) with proprietary transaction and access logs; while this design choice was necessary given the confidentiality constraints surrounding real institutional breach data, synthetic logs cannot fully reproduce the noise, irregularity, and evolving tactics present in genuine production environments, and the reported performance figures should be read as an upper-bound estimate rather than a guarantee of equivalent real-world accuracy. Second, the evaluation environment models only three insider threat scenarios (S-01 to S-03) within a single simulated 1,000-employee global bank; this scope, while representative of common high-risk profiles, does not capture the full diversity of insider threat behaviour across institutions of different size, sector, or organizational structure, and generalization beyond these scenarios has not been empirically tested.

Third, the system has not been validated through a live or pilot deployment within an operational Security Operations Center, and its real-world operational efficiency, analyst workload impact, and false positive fatigue over extended use

remain unverified beyond the offline evaluation reported here. Fourth, the explainability analysis in Section 5 is illustrated through a limited set of forensic case studies rather than a systematic, ground-truth-validated audit across the full test set, so the comparative reliability of SHAP and LIME explanations may vary on cases not represented in this sample. Finally, this study does not include an adversarial robustness evaluation of the post-hoc explainability layer, nor a fairness or demographic bias analysis of model outcomes across employee subgroups; both are identified as important directions for future work rather than claims made by the current results.

7. Conclusion and Future Work

This research delivers a comprehensive and thorough empirical analysis on the synergy of a multi-modal deep learning system with post-hoc Explainable AI (XAI) models on an insider threat dataset within the financial domain. Our system projects transaction, authentication, and communication surveillance log datasets onto a shared latent space through a CNN-LSTM-Autoencoder model and achieves empirical accuracy as high as 96.5% ($\pm$0.4%) and a base global False Positive Rate (FPR) as low as 1.05%. We have shown that we can control the operating FPR by optimizing the classification threshold output (τ) to be ROC maximizing in value (down to 0.32%), while not performing worse than baseline models, which has a clear advantage compared to a single modality model. It also provides evidence to illuminate the practical compromises of post-hoc explainability by contrasting LIME and SHAP, which has a lower latency but approximate local explanations versus more consistent feature ranking, mathematical consistency, and compliance with the requirements on algorithmic accountability in financial industry regulatory.

Therefore, a dual architecture combining these two complementary techniques may provide a strong defensive posture.

We plan to build this framework further and focus on the application of federated learning for distributed anomalous pattern analysis across branch networks while not centralizing user sensitive data. We plan further research on ensuring the resilience of the post-hoc XAI technique against manipulation by the attacker.

8. References

Cho, K., Van Merrinboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. ArXiv preprint arXiv:1406.1078. https://doi.org/10.48550/arXiv.1406.1078

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

.LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444. https://doi.org/10.1038/nature14539

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, ., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998-6008.

V.Bengio, Y Et al (2013). Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8), 1798-1828. https://doi.org/10.1109/TPAMI.2013.50

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

Lipton, Z. C.(2016). The mythos of model interpretability. ArXiv preprint arXiv:1606.03490. https://doi.org/10.48550/arXiv.1606.03490

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765-4774.

IRibeiro, 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/10.1145/2939672.293977

X.Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Mller, K. R.(Eds.). (2019). Explainable AI: Interpreting, explaining and visualizing deep learning. Springer. https://doi.org/10.1007/978-3-030-28954-6

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. International Conference on Machine Learning, 3145-3153.

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. International Conference on Machine Learning, 3319-3328.

Costa, D. L., Albrethsen, M. J., & Collins, M. (2016). An insider threat indicator ontology (Technical Report CMU/SEI-2016-TR-007). Software Engineering Institute, Carnegie Mellon University.

Glasser, J., & Lindauer, B. (2013). Bridging the gap: A pragmatic approach to generating insider threat data. 2013 IEEE Security and Privacy Workshops, 98-104. https://doi.org/10.1109/SPW.2013.37

Du, M., Li, F., Zheng, G., & Srikumar, V. (2017). DeepLog: Anomaly detection and diagnosis from system logs through deep learning. Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 1285-1298. Https://doi.org/10.1145/3133956.3134015

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

Giddens, L., Amo, L. C., & Cichocki, D. (2020). Gender bias and the impact on managerial evaluation of insider security threats. Computers & Security, 99, 102066. https://doi.org/10.1016/j.cose.2020.102066

Homoliak, I., Toffalini, F., Guarnizo, J., Elovici, Y., & Ochoa, M. (2019). Insight into insiders and IT: A survey of insider threat taxonomies, analysis, modelling, and countermeasures. ACM Computing Surveys, 52(2), 30:1-30:40. https://doi.org/10.1145/3303772

Yuan, S., & Wu, X. (2021). Deep learning for insider threat detection: Review, challenges and opportunities. Computers & Security, 104, 102221. https://doi.org/10.1016/j.cose.2021.102221

Nasir, R. Et al (2021). Behavioural based insider threat detection using deep learning. IEEE Access, 9, 143266-143274. https://doi.org/10.1109/ACCESS.2021.3118235

Huang, W., Zhu, H., Li, C., Lv, Q., Wang, Y., & Yang, H. (2021). ITDBERT: Temporal-semantic representation for insider threat detection. 2021 IEEE Symposium on Computers and Communications (ISCC), 1-7. https://doi.org/10.1109/ISCC53001.2021.9631482

Le, D. C., & Zincir-Heywood, N. (2021). Anomaly detection for insider threats using unsupervised ensembles. IEEE Transactions on Network and Service Management, 18(2), 1152-1164.

https://doi.org/10.1109/TNSM.2021.3071928

Le, D. C., Zincir-Heywood, N., & Heywood, M. I.(2020). Analyzing data granularity levels for insider threat detection using machine learning. IEEE Transactions on Network and Service Management, 17(1), 30-44. Https://doi.org/10.1109/TNSM.2019.2947382

Bin Sarhan, B., & Altwaijry, N. (2022). Insider threat detection using machine learning approach. Applied Sciences, 13(1), 259. https://doi.org/10.3390/app13010259

AlSlaiman, M., Salman, M. I., Saleh, M. M., & Wang, B. (2023). Enhancing false negative and positive rates for efficient insider threat detection. Computers & Security, 126, 103066. https://doi.org/10.1016/j.cose.2022.103066

Xiao, J. Et al (2023). Robust anomaly-based insider threat detection using graph neural network. IEEE Transactions on Network and Service Management, 20(3), 3717-3733. Https://doi.org/10.1109/TNSM.2022.3222635

Gayathri, R. G., Sajjanhar, A., & Xiang, Y. (2024). Hybrid deep learning model using SPCAGAN augmentation for insider threat analysis. Expert Systems with Applications, 249, 123533. https://doi.org/10.1016/j.eswa.2024.123533

Cai, X., Wang, Y., Xu, S., Li, H., Zhang, Y., Liu, Z., & Yuan, X. (2024). LAN: Learning adaptive neighbors for real-time insider threat detection. IEEE Transactions on Information Forensics and Security.

Chawla, N. V.Et al (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357. https://doi.org/10.1613/jair.953

Ramachandram, D., & Taylor, G. W.(2017). Deep multimodal learning: A survey on recent advances and trends. IEEE Signal Processing Magazine, 34(6), 96-108. Https://doi.org/10.1109/MSP.2017.2738401

Baltruaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423-443. https://doi.org/10.1109/TPAMI.2018.2798607

Ruff, L Et al (2021). A unifying review of deep and shallow anomaly detection. Proceedings of the IEEE, 109(5), 756-795. https://doi.org/10.1109/JPROC.2021.3052449

Arrieta, A. B.Et al. (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.2019.11.004

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". AI Magazine, 38(3), 50-57. https://doi.org/10.1609/aimag.v38i3.2741XXXV.Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), 841-887. Https://doi.org/10.2139/ssrn.3063289