Explainable Alert Prioritisation: An Evaluation Protocol for SHAP-Based Insider Threat Detection in Critical National Infrastructure

Tayyeb Nadeem Somro and Bilal Arshad

School of Architecture, Built Environment, Computing and Engineering,
Birmingham City University, UK
tayyeb.nadeemsomro@mail.bcu.ac.uk, bilal.arshad2@mail.bcu.ac.uk

Abstract. We present an evaluation protocol for testing whether SHAP-based feature attribution can support alert prioritisation in insider threat detection, rather than serve only as a cosmetic explanation layer applied after detection. Insider threat detection in Critical National Infrastructure (CNI) environments increasingly relies on unsupervised anomaly scoring and supervised classification over user activity telemetry. These models flag rare, high-consequence behaviour, but their outputs are not always interpretable to the analysts who must justify escalation decisions under regulatory frameworks such as the NCSC Cyber Assessment Framework (CAF). We specify a hybrid Isolation Forest and Random Forest pipeline over the CERT Insider Threat Dataset, define two metrics for explanation quality, the Explanation Concordance Rate (ECR) and the Alert Prioritisation Fidelity Score (APFS), and articulate a falsifiable hypothesis, *Explainability Inversion*: that the alerts assigned the highest anomaly confidence may be systematically the least explainable, a prediction motivated by documented SHAP attribution instability under feature correlation. We situate the protocol against two directly comparable prior studies on SHAP-based insider threat explanation, against current evidence on SOC analyst trust and alert fatigue, against the UK's August 2025 Cyber Assessment Framework update (CAF 4.0), and against the current landscape of CERT-alternative insider threat datasets. A targeted literature search did not identify peer-reviewed work studying SHAP or LIME attribution instability specifically within tree-based anomaly detection models, as distinct from standard classifiers; whether such instability extends to this setting is treated here as an open empirical question rather than a closed gap. The protocol, the metrics, and the ground-truth indicator mapping are specified in full here; empirical execution is scoped as the next phase of this research rather than presented as a finding of this paper.

Keywords: Explainable AI · Insider Threat · SHAP · Critical National Infrastructure · Alert Prioritisation · Isolation Forest · NCSC CAF · Evaluation Protocol

1 Introduction

Insider threat programmes in Critical National Infrastructure (CNI) operators face a structural tension. The term itself, an authorised individual who uses legitimate access for unauthorised or malicious ends [13], already implies that detection cannot rely solely on the perimeter controls used against external attackers. Unsupervised anomaly detectors such as Isolation Forest are attractive precisely because malicious insider behaviour is rare and rule-based detection cannot enumerate it in advance [3], an approach that broader taxonomies of insider threat detection situate alongside supervised and graph-based alternatives [1]. Yet the same property that makes anomaly detectors useful, their sensitivity to statistically unusual behaviour, makes their outputs opaque: an anomaly score alone gives an analyst no account of why a given employee's activity was flagged.

Alert fatigue and automation bias, and the resulting erosion of analyst trust in opaque outputs, are documented problems across SOC environments generally [20], and practitioner-facing guidance on insider threat mitigation already assumes analysts need a defensible rationale before escalating, not just a score [14]. In a CNI context this is compounded by regulatory expectations of proportionate, justifiable security decision-making: the NCSC's Cyber Assessment Framework was updated to version 4.0 in August 2025 specifically to extend its coverage of AI-related cyber risk across the framework, although that update is outcome-based rather than a prescriptive AI explainability checklist [9, 21]. The practical question this raises for CNI SOC teams is not whether an anomaly detector can flag insider behaviour, but whether the explanation attached to that flag can withstand the scrutiny an escalation decision requires.

This paper makes the following contributions:

- A specification of a hybrid Isolation Forest and Random Forest pipeline for insider threat detection over the CERT Insider Threat Dataset [2, 17], with an explicit, sourced ground-truth indicator mapping for the scenarios used.
- Two evaluation metrics, ECR and APFS, designed specifically to test whether SHAP attribution is operationally useful for alert prioritisation rather than merely present, with a fully specified measurement protocol.
- A named, falsifiable hypothesis, *Explainability Inversion*, motivated by documented SHAP instability under feature correlation [12] and extended here to the tree-based anomaly detection setting, where a targeted search did not identify prior work testing it directly, leaving the question open rather than closed.
- Identification of two directly comparable prior works [18, 19] against which results from this protocol are designed to be benchmarked, and a positioning of the CERT dataset against the current landscape of CERT-alternative benchmarks [22, 23].

2 Related Work and Problem Formalisation

2.1 Insider Threat Detection

Glasser and Lindauer [2, 17] introduced and documented the CERT Insider Threat Dataset as a synthetic but behaviourally grounded benchmark spanning data exfiltration, sabotage, and fraud scenarios across simulated organisational logs. Liu et al. [3] proposed Isolation Forest, which isolates anomalies via random partitioning rather than density estimation, making it well suited to the heavily imbalanced class distribution typical of insider threat data. Breiman's Random Forest [4] remains a common supervised baseline once partial ground truth is available. Sanzgiri and Dasgupta's taxonomy of insider threat detection techniques [1] situates both approaches within a broader landscape spanning rule-based, statistical, and graph-based methods. Surveys of machine learning approaches to insider threat [6, 8] report detection metrics on CERT-derived benchmarks; few of the underlying studies evaluate whether the resulting alerts are interpretable enough to act on. This gap, between detection performance and explanation usefulness, is the gap this protocol is designed to test.

2.2 Explainable AI in Security Contexts

Lundberg and Lee's SHAP framework [5] unifies feature attribution under a game-theoretic Shapley value formulation and is widely used for tree-based security classifiers, within the broader landscape of post-hoc explanation methods surveyed by Vilone and Longo [7]. Rudin [15] argues that post-hoc explanation of black-box models is a questionable foundation for high-stakes decisions compared with inherently interpretable models. This paper does not adjudicate that debate, but it adopts Rudin's underlying premise: SHAP's explanatory value should be tested for faithfulness rather than assumed. Slack et al. [12] showed that post-hoc attribution methods including SHAP and LIME can be destabilised by feature correlation, meaning an explanation can be mathematically well-defined without being faithful or stable. A targeted search for follow-on work studying this instability within tree-based anomaly detection models (Isolation Forest, Random Forest used for outlier scoring), rather than standard supervised classifiers, returned general tutorials and applied demonstrations of combining SHAP with Isolation Forest, but did not identify a peer-reviewed study isolating attribution instability under feature correlation specifically for anomaly-scoring models. Whether this instability extends to anomaly-scoring models is, on the basis of this search, an open empirical question, which the Explainability Inversion hypothesis (Section 3) is designed to test.

2.3 Directly Comparable Prior Work

Two existing studies overlap substantially with the approach proposed here and define the baseline against which results from this protocol are designed to be evaluated. Grzeczkowicz et al. [18] use SHAP values produced by an anomaly

detection model to classify CERT insider threat scenarios after detection, reporting high scenario-classification accuracy. Their approach addresses justification at the scenario level rather than the alert-ranking level targeted here, but the underlying SHAP-over-CERT pipeline is closely related, and any implementation of the protocol in Section 4 should be benchmarked against it directly. T. S.P. [19] proposes a hybrid anomaly detection and XAI framework for insider threat in cloud environments, evaluated on CERT and ADFA-LD, and defines an Explainability Score and a Fidelity Score that are conceptually close to the ECR and APFS metrics proposed below. Benchmarking the protocol in this paper against both [18] and [19] is a required step before any claim of improvement, and is listed as such in Section 8.

2.4 Dataset Landscape

CERT is not claimed to model CNI telemetry directly; it is used here as a controlled insider-threat benchmark through which the proposed explanation-evaluation protocol can be specified before validation on operational CNI data. The dataset's limitations are acknowledged within the literature that continues to use it: a 2025 dataset paper introducing the SPEDIA dataset describes the field's continuing dependence on CERT as characterised by well-known limitations in realism, imbalance, and currency [22]. SPEDIA proposes a hybrid alternative combining real participant activity from a controlled cyber exercise, role-based simulated activity, and CERT-derived synthetic events, and explicitly extends rather than replaces the CERT format [22]. Separately, a 2026 benchmark, OrgForge-IT, addresses a different limitation of CERT, namely that it is static and offers no mechanism for generating new labelled instances with varied organisational profiles, by proposing a verifiable synthetic generation pipeline aimed at LLM-based insider threat detection [23]. Neither SPEDIA nor OrgForge-IT is used in the protocol proposed in Section 4; CERT is retained specifically so that results remain directly comparable to the named baselines in Section 2.3, both of which also use CERT. This is a deliberate comparability-versus-realism trade-off, noted further in Section 8.

2.5 Problem Formalisation

Let $\mathcal{X}$ denote the space of alert-level feature vectors derived from user activity logs (login timing, removable media usage, email attachment volume, file access counts), and $y_i \in \{0, 1\}$ the CERT-released ground-truth label for instance i. The Isolation Forest anomaly score is:

$$a(x) = 2^{-\frac{E(h(x))}{c(n)}} \tag{1}$$

where $E(h(x))$ is the expected path length over the forest and $c(n)$ the average path length normaliser for sample size n. The Random Forest classifier outputs $f_{\mathrm{RF}}(x) \in [0, 1]$. The composite risk score proposed for prioritisation is:

$$\rho(x) = w_a\, a(x) + w_r\, f_{\mathrm{RF}}(x), \quad w_a + w_r = 1 \tag{2}$$

SHAP attribution decomposes a model output additively over features $j \in \{1, \ldots, m\}$:

$$f(x) = \mathbb{E}[f(X)] + \sum_{j=1}^{m} \phi_j(x) \tag{3}$$

For each scenario category s, let $G_s \subset \{1, \ldots, m\}$ denote the scenario-specific ground-truth indicator feature set, defined explicitly in Table 1. The Explanation Concordance Rate over the top-k SHAP features $\mathrm{Top}_k(x)$ is:

$$\mathrm{ECR} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\big(\mathrm{Top}_k(x_i) \cap G_{s_i} \neq \emptyset\big) \times 100 \tag{4}$$

The Alert Prioritisation Fidelity Score is the Spearman rank correlation between a SHAP-weighted composite ranking R_{SHAP} and the ground-truth severity ranking R_{GT}:

$$\mathrm{APFS} = \rho_{\mathrm{Spearman}}(R_{\mathrm{SHAP}}, R_{\mathrm{GT}}) \tag{5}$$

2.6 Why ECR and APFS Are Needed

Detection accuracy alone does not answer the question this protocol is built around. An alert can be correctly detected, with a high anomaly score and a label that matches the ground truth, while its SHAP explanation points an analyst towards an operationally irrelevant feature. Standard classification metrics (precision, recall, F1) are blind to this failure mode, because they evaluate the label, not the explanation attached to it.

SOC alert handling is also fundamentally a prioritisation problem rather than a binary classification problem: analysts work down a queue under time pressure, and systematic surveys of alert prioritisation in SOC environments treat the ordering of alerts as a distinct criterion from the underlying detection task, separate from whether any individual alert is eventually correctly classified [25]. The order in which alerts are reviewed matters operationally as much as detection correctness, particularly in resource-constrained CNI SOC teams [20]. A metric that evaluates explanation quality therefore needs to assess both whether an individual explanation is meaningful and whether explanation-weighted prioritisation across many alerts tracks genuine severity. SHAP explanations cannot be left to visual inspection at scale; a SOC processing large alert volumes needs an automatable check against scenario-specific ground-truth indicators, not a human reading a force plot for each case.

ECR and APFS are designed to answer these two distinct questions separately. ECR measures, for a single alert, whether the SHAP explanation points to at least one feature that is operationally meaningful for that scenario type, using the indicator mapping in Table 1 as the operational definition of meaningful: this is local explanation concordance. APFS measures, across the alert population, whether ranking alerts by explanation-weighted risk produces an ordering consistent with ground-truth severity: this is ranking fidelity. The two can diverge

Table 1. Ground-Truth Indicator Mapping by CERT Scenario

Release	Scenario Summary	Ground-Truth Log Field	Source
r4.2, Sc. 1	User begins removable-media use atypical for their role and uploads data externally before leaving the organisation.	Removable-device connect/disconnect event count; outbound upload event to an external destination.	[2, 17], scenario description
r4.2, Sc. 2	User logs into a co-worker's machine after hours using their credentials and uploads documents.	Logon event timestamp outside the user's established hours, on a host not assigned to them.	[2, 17], scenario description
r4.2, Sc. 3	System administrator installs a keylogger, exfiltrates data via removable media, and emails data to a personal account before leaving.	Removable-device file-copy count; email attachment to an external/personal domain.	[2, 17], scenario description

by design. An explanation pipeline could reliably point to a relevant feature on each individual alert (high ECR) while still producing a ranking that does not track severity well (low APFS), for instance if it is similarly confident in its top feature regardless of how severe the underlying behaviour actually is. Reporting both metrics, rather than either alone, is therefore a designed property of the protocol, not a redundancy.

3 Indicator Mapping and Explanation Risk Hypothesis

A central weakness in much SHAP-over-insider-threat work is that the ground truth a SHAP explanation is checked against is rarely made explicit. Table 1 specifies, scenario by scenario, exactly which CERT log field is treated as the ground-truth indicator and the documentation it is drawn from.

The mapping in Table 1 is drawn from the published scenario descriptions and accompanying documentation [2, 17] rather than from direct inspection of

the raw per-user answer key. An implementation of the protocol in Section 4 should re-derive and verify G_s directly against the dataset's released answer files before computing ECR; this is a verification step required of any execution, not a flaw in the mapping as specified.

3.1 Stakes of Misexplanation

Four risks motivate treating explanation quality as a first-class evaluation target rather than a secondary check. First, repeated low-concordance explanations could erode analyst trust in the system generally, consistent with the automation-bias and trust-calibration literature on SOC environments [20]; a grey-literature thesis on explainable anomaly classification reports a related concern about trust and interpretability in security contexts, used here only as a weak supporting signal [24]. Second, escalation decisions resting on an unfaithful explanation are difficult to defend under CAF-style assurance [9, 21], or against the broader risk-governance expectations set out in the NIST AI Risk Management Framework [16], which similarly emphasises traceable justification without mandating a specific explainability mechanism. Third, an explanation pointing to the wrong feature could direct an investigation towards an unrelated behaviour pattern, a risk distinct from but related to the misattribution concerns documented for ATT&CK-based threat intelligence extraction more generally [10]. Fourth, attribution instability under model retraining could silently change what a model is attending to without a corresponding change in reported accuracy, a risk documented in general for SHAP/LIME [12].

3.2 Indicator Visibility Levels

Level 1 (Single-Channel): only one log source is available, limiting both detection recall and the diversity of features SHAP can draw on. **Level 2 (Cross-Channel):** email, removable media, and authentication logs are jointly available; this is the configuration the protocol in Section 4 targets. **Level 3 (Full Telemetry):** cross-channel logs supplemented with HR or psychometric indicators, which raise privacy and fairness concerns, consistent with UK guidance on AI and data protection [11], and are kept out of scope here.

3.3 The Explainability Inversion Hypothesis

Isolation Forest assigns its most extreme anomaly scores to points isolated earliest by random partitioning, which in practice often corresponds to unusual combinations of several weakly correlated features rather than an extreme value on a single dominant feature. Slack et al. [12] show that SHAP attribution becomes less stable, and attribution mass is more easily redistributed between features, when features are correlated, though their study examines this in standard classifiers rather than anomaly-scoring models. If the features driving the most extreme anomaly scores are also the most mutually correlated in the CERT feature

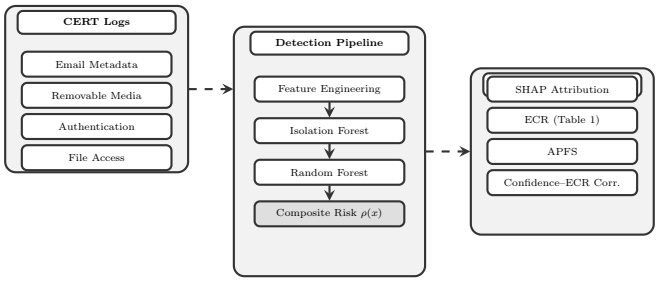

Fig. 1. Specified methodology: CERT log channels would feed the detection pipeline, whose composite risk score and SHAP attributions would be jointly evaluated against the indicator mapping in Table 1.

set, the points with the highest anomaly confidence may, by this mechanism, also be the points whose top-k SHAP attribution is least stable and least concordant with a scenario-specific ground-truth indicator. We label this the *Explainability Inversion hypothesis*: it is a theoretically motivated, falsifiable prediction, extending Slack et al.'s classifier-context finding to an anomaly-detection context where a targeted search did not identify direct prior evidence, to be tested under the protocol in Section 5.

4 Proposed Methodology

Figure 1 shows the pipeline specified for testing the hypothesis above.

4.1 Dataset Construction

The protocol specifies generating alert-level instances by aggregating daily user activity from the CERT Insider Threat Dataset into feature vectors: removable media file count, after-hours login count, external email attachment volume and size, and file access frequency across shared drives. Consistent with the scenarios in Table 1, the resulting instance set would be heavily imbalanced, since the genuine base rate of insider threat activity in CERT releases is low. This base rate should not be artificially corrected, since rebalancing would distort the anomaly detection setting under test.

4.2 Model Configurations

Three configurations are specified for comparison: **IF-only** (unsupervised Isolation Forest anomaly score), **RF-only** (supervised Random Forest trained on the labelled subset), and **Hybrid** ($\rho(x)$ combining both). The protocol specifies that SHAP values would be computed using TreeExplainer for all three configurations, with $k = 3$ for top-feature concordance checks, matching the indicator sets in Table 1.

5 Evaluation Protocol

The protocol evaluates three questions in sequence: whether explanation quality, measured by ECR and APFS, differs systematically between the three model configurations; whether the Explainability Inversion hypothesis holds in this setting; and how the resulting figures compare against the two named baselines. No execution of this protocol has been performed for this paper; the remainder of this section specifies what an execution should report.

Baseline explanation quality. For each configuration, ECR and APFS would be computed on a held-out test set, using the indicator sets in Table 1. The supervised configuration is expected to show higher raw ECR than the unsupervised configuration, since supervised training implicitly biases the model towards features known to discriminate the labelled scenarios. This expectation should be checked rather than assumed, since higher ECR does not necessarily imply higher rank fidelity (APFS); the two metrics are designed to be able to diverge (Section 2.6).

Explainability Inversion test. The primary output of this test would be the Pearson correlation r_{EI} between per-alert composite risk $\rho(x_i)$ and per-alert concordance, computed over the malicious-instance subset for each configuration. A negative r_{EI} across configurations would be consistent with the mechanism proposed in Section 3, though not on its own proof of it. A null or positive correlation would falsify the Explainability Inversion hypothesis as stated, and a future execution should report whichever result is obtained without adjusting the hypothesis post hoc.

Benchmarking against named prior work. A future execution should report ECR and APFS alongside a reimplementation, on the same CERT subset, of the scenario-classification accuracy reported by Grzeczkowicz et al. [18] and the Explainability and Fidelity Scores reported by T. S.P. [19], since this comparison is what establishes whether the protocol offers any improvement over the closest existing work.

Explanation Overhead Ratio. The protocol specifies measuring the added per-alert latency of SHAP computation relative to base model inference (EOR) for each configuration, as a practical deployability check alongside the explanation-quality metrics.

6 Operationalising the Metrics: The Lucid Design Pattern

The metrics in Section 2.6 are intended for use, not only for measurement. Lucid is a design pattern showing one way ECR and APFS could feed an analyst-facing decision rule; it is not a validated system, has not been implemented, and is presented as a secondary illustration of the protocol's practical relevance rather than as a separate contribution to be evaluated on its own terms.

Lucid would route an alert to mandatory analyst review whenever its composite risk $\rho(x)$ is high but its SHAP attribution is unstable or discordant with

Algorithm 1: Lucid Design Pattern: Justification Logic (Not an Evaluated Method)

Input: Alert features x, scenario indicator set G_s, tolerance ϵ_{stab}
Output: Justified Alert J (design pattern, not yet evaluated)
$\rho \leftarrow \text{CompositeRisk}(x)$;
$\phi \leftarrow \text{SHAP}(x)$; $\hat{\phi} \leftarrow \text{SHAP}(x + \delta)$ for small perturbation δ;
$\phi_{\text{stable}} \leftarrow \{\phi_j : |\text{rank}(\phi_j) - \text{rank}(\hat{\phi}_j)| \leq \epsilon_{\text{stab}}\}$;
if $Top_k(\phi_{stable}) \cap G_s = \emptyset$ **then**
  RouteToAnalystReview; **return**
end
if ρ *high* ***and*** *concordance low* **then**
  RouteToAnalystReview;
else
  $J \leftarrow \text{CAFTemplate}(\phi_{\text{stable}}, G_s)$;
end
return J;

the relevant indicator set, directly operationalising the Explainability Inversion hypothesis as a decision rule. Stable, concordant explanations would instead be rendered into a structured justification referencing the relevant CAF outcome under CAF 4.0's current structure [21], rather than presented as raw SHAP values. The algorithm below sets out this logic as a design pattern; it has not been evaluated, and no claim about its effect on ECR, APFS, or analyst behaviour is made.

7 Discussion

This paper's contribution is a specification: a hybrid detection pipeline, two explanation-quality metrics with a justified rationale (Section 2.6), a sourced ground-truth indicator mapping, a falsifiable hypothesis grounded in documented SHAP behaviour, and a worked design pattern (Lucid) showing how the metrics could be used operationally. It claims a CAF-aligned framing of AI-risk coverage [21], broadly consistent with the NIST AI RMF's risk-governance approach [16], without claiming that either framework mandates explainability for automated alert-ranking specifically. It does not claim a measured performance improvement, a closed research gap, or a validated deployment.

Section 2.3 identifies two papers, [18,19], whose methods overlap substantially with the protocol proposed here; benchmarking against both is the step that would establish relative merit, and is treated as required future work rather than as already accomplished. Section 2.4 situates CERT against two more recent alternatives, [22, 23]; retaining CERT here is a comparability decision, not a claim that CERT is the most realistic dataset available. A targeted search did not surface peer-reviewed work on SHAP/LIME instability specifically within tree-based anomaly detection (Section 2); this absence of identified evidence

is the motivation for the Explainability Inversion hypothesis, which extends a classifier-context result [12] into a setting that, on the basis of this search, has not previously been tested, rather than a confirmed gap in the field.

8 Limitations

Scope limitations. This paper specifies an evaluation protocol; it does not report an empirical run of it. Every quantitative figure that would ordinarily appear in a results section (ECR, APFS, r_{EI}, EOR) is therefore absent by design, and execution of the protocol in Section 5 is the principal piece of future work.

Dataset limitations. CERT is a synthetic benchmark with known limitations in realism, imbalance, and currency (Section 2.4); it is not claimed to represent CNI telemetry directly. The ground-truth indicator mapping in Table 1 is derived from published scenario descriptions and dataset documentation [2,17] rather than the dataset's raw answer files, and should be re-verified against those files before the protocol is executed.

Empirical validation limitations. The Explainability Inversion hypothesis rests on extending a classifier-context finding [12] to an anomaly-detection context where a targeted search did not identify direct prior evidence; it is plausible, not established. The Lucid design pattern has not been implemented or evaluated; no claim about its effect on analyst behaviour is made.

Baseline benchmarking limitations. Benchmarking against Grzeczkowicz et al. [18] and T. S.P. [19] has not been performed and is required before any claim of contribution or improvement over the closest existing work could be made.

CNI generalisability limitations. Indicator Visibility Level 3 (HR or psychometric signal) is excluded from the protocol on fairness and privacy grounds [11]. The CAF 4.0 citation [21] supports the claim that AI-risk coverage was expanded in the August 2025 update; it does not support a claim that CAF 4.0 mandates explainability for automated alert-ranking, and no such claim is made, nor does the NIST AI RMF citation [16] support such a claim. Results obtained on CERT would still require validation against operational CNI telemetry before any claim of CNI applicability could be made.

9 Conclusion

This paper specifies an evaluation protocol for testing whether SHAP-based feature attribution can support alert prioritisation in insider threat detection for CNI environments. It defines two purpose-built metrics against an explicit, sourced ground-truth indicator mapping; states a falsifiable hypothesis grounded in documented but, at this specific intersection, untested SHAP behaviour; names two existing studies against which results must be benchmarked; situates the dataset choice against the current CERT-alternative landscape and the current UK cyber-assurance context (CAF 4.0); and illustrates, through the Lucid design pattern, how the metrics could inform an analyst-facing workflow. The

principal future work is to execute the protocol in Section 5, report whichever result the Explainability Inversion test produces, and benchmark the outcome against the named prior work.

References

1. A. Sanzgiri and D. Dasgupta, "Classification of Insider Threat Detection Techniques," in *Proc. ACM CISRC*, 2016.
2. J. Glasser and B. Lindauer, "Bridging the Gap: A Pragmatic Approach to Generating Insider Threat Data," in *Proc. IEEE Security and Privacy Workshops*, pp. 98–104, 2013.
3. F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," in *Proc. IEEE ICDM*, pp. 413–422, 2008.
4. L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
5. S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Proc. NeurIPS*, pp. 4765–4774, 2017.
6. D. C. Le and N. Zincir-Heywood, "Evaluating Insider Threat Detection Workflow Using Supervised and Unsupervised Learning," in *Proc. IEEE Security and Privacy Workshops*, 2018.
7. G. Vilone and L. Longo, "Explainable Artificial Intelligence: A Systematic Review," *arXiv:2006.00093*, 2020.
8. P. Singh et al., "A Deep Learning Approach for Insider Threat Detection," in *Proc. IEEE ICACCS*, 2020.
9. National Cyber Security Centre, "Cyber Assessment Framework (CAF) Guidance," *NCSC*, 2024. [Online]. Available: `https://www.ncsc.gov.uk/collection/cyber-assessment-framework`
10. MITRE, "ATT&CK: Adversarial Tactics, Techniques and Common Knowledge," 2024. [Online]. Available: `https://attack.mitre.org`
11. Information Commissioner's Office, "Guidance on AI and Data Protection," *ICO*, 2023.
12. D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation Methods," in *Proc. AAAI/ACM AIES*, pp. 180–186, 2020.
13. M. Bishop and C. Gates, "Defining the Insider Threat," in *Proc. ACM CSIIRW*, 2008.
14. CISA, "Insider Threat Mitigation Guide," *CISA*, 2020.
15. C. Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions," *Nature Machine Intelligence*, vol. 1, pp. 206–215, 2019.
16. NIST, "AI Risk Management Framework (AI RMF 1.0)," *NIST*, 2023.
17. J. Glasser and B. Lindauer, "CERT Insider Threat Dataset Documentation," *Carnegie Mellon University Software Engineering Institute*, 2016.
18. R. Grzeczkowicz, C. Neal, N. Baghalizadeh-Moghadam, N. Boulahia Cuppens, and F. Cuppens, "Classifying Insider Threat Scenarios Through Explainable Artificial Intelligence," in *Risks and Security of Internet and Systems (CRiSIS 2024)*, LNCS vol. 15456, pp. 157–172, Springer, Cham, 2025.
19. T. S.P., "Explainable AI (XAI) for Insider Threat Detection: Balancing Security and Transparency in Cloud Computing," *Concurrency and Computation: Practice and Experience*, vol. 37, no. 25-26, e70390, 2025.

20. S. Tariq, M. Baruwal Chhetri, S. Nepal, and C. Paris, "Alert Fatigue in Security Operations Centres: Research Challenges and Opportunities," *ACM Computing Surveys*, vol. 57, no. 9, Article 224, 2025.
21. National Cyber Security Centre, "Cyber Assessment Framework v4.0 Released in Response to Growing Threat," *NCSC Blog*, Aug. 2025. [Online]. Available: `https://www.ncsc.gov.uk/blog-post/caf-v4-0-released-in-response-to-growing-threat`; full text: `https://www.ncsc.gov.uk/files/NCSC-Cyber-Assessment-Framework-4.0.pdf`
22. D. Álvarez Muñiz, L. Pérez Miguel, A. Mateo Muñoz, X. Larriva-Novo, M. Álvarez-Campana, and D. Rivera, "Design and Generation of a Dataset for Training Insider Threat Prevention and Detection Models: The SPEDIA Dataset," *Computers & Security*, 2025, Art. 104743. [Online]. Available: `https://doi.org/10.1016/j.cose.2025.104743`
23. J. Flynt, "OrgForge-IT: A Verifiable Synthetic Benchmark for LLM-Based Insider Threat Detection," *arXiv:2603.22499*, 2026.
24. "EXACT: An Explainable Anomaly Classification Tool," Master's thesis, 2025. [Online]. Available via institutional repository (DiVA portal). Grey literature; not peer-reviewed; used only as a weak supporting signal in Section 3.
25. F. Jalalvand, M. Baruwal Chhetri, S. Nepal, and C. Paris, "Alert Prioritisation in Security Operations Centres: A Systematic Survey on Criteria and Methods," *ACM Computing Surveys*, vol. 57, no. 2, Article 42, 2025.