



doi:10.5281/zenodo.18907306

Explainability Gaps in Machine Learning-Based Insider Threat Detection: A Critical Perspective

Kitoye Ebire Okonny¹, Eseosa Omorogiuwa², Matthew Ehikhamenle²

¹ICT Centre, Ignatius Ajuru University of Education, Port Harcourt, Nigeria

²Department of Electrical and Electronics Engineering,
University of Port Harcourt, Port Harcourt, Nigeria

ABSTRACT

Machine learning has become integral to insider threat detection due to its capacity to model complex user behaviours and identify anomalous activities within organisational environments. Despite their predictive power, many machine learning systems operate as “black boxes,” generating risk scores that are difficult for security analysts to interpret, justify, or trust. Although explainable artificial intelligence (XAI) techniques have been introduced to improve transparency, explainability in insider threat detection remains fragmented and insufficiently aligned with operational needs. This paper presents a critical perspective on explainability gaps in machine learning-based insider threat detection. Rather than concentrating solely on detection accuracy, it examines limitations across three interconnected dimensions: technical adequacy, human trust, and practical usability. The review analyses how behavioural data is modelled, how black-box systems are deployed, and how existing explanation techniques often rely on shallow feature attribution while neglecting temporal dynamics and organisational context. It argues that many current approaches prioritise algorithmic transparency over actionable insight, resulting in explanations that are difficult to operationalise. The paper concludes by advocating for human-centred, context-aware, and analyst-in-the-loop frameworks to support more trustworthy and effective insider threat detection systems.

Keywords: Insider Threat Detection; Explainable Artificial Intelligence; Machine Learning Interpretability; Behavioural Analytics, Trustworthy AI

1. INTRODUCTION

Insider threats remain one of the most multifaceted and destructive security challenges facing modern organisations. Unlike external attackers, insiders operate within established trust boundaries and possess legitimate access to sensitive systems, data, and operational resources (Zeng et al., 2023; Inayat et al., 2024). These insiders may be malicious actors intentionally exploiting their privileges, negligent employees inadvertently exposing vulnerabilities, or compromised users whose credentials have been hijacked by external adversaries (Saxena et al., 2020). Because insider activities often resemble normal workplace behaviour, detecting malicious intent becomes significantly more difficult than identifying external intrusions. As organisations increasingly rely on digital infrastructures, cloud platforms, and distributed work environments, the scale and subtlety of insider risk continue to grow (Pitkar, 2025; Prasad et al., 2025).

Machine learning (ML) has emerged as a powerful tool for addressing insider threat detection challenges. ML-based systems analyse large volumes of user activity logs, access records, and behavioural patterns to identify anomalies or suspicious deviations from established norms (Asmar & Tuqan, 2024; Qawasmeh et al., 2025). Supervised, unsupervised, and semi-supervised models are commonly used to classify insider behaviour, detect abnormal access patterns, and flag potential misuse (Kamatchi & Uma, 2025). Compared to traditional rule-based or signature-based detection systems, machine learning approaches offer improved adaptability, scalability, and the ability to capture complex behavioural relationships. As a result, they have become central to modern insider threat detection research and practice. However, despite their predictive capabilities, many machine learning models function as “black boxes,” producing outputs without clear explanations of how decisions were reached (Hassija et al., 2024; Liu et al., 2025). In high-stakes security environments, this opacity presents a significant problem. Security analysts must evaluate alerts, justify decisions, and take actions that may affect employees’ careers or organisational operations. When a detection system flags a user as suspicious without providing understandable reasoning, analysts may struggle to assess the credibility of the alert. This lack of transparency can reduce trust in automated systems, increase reliance on manual investigation, and ultimately limit operational adoption (Mitchell, 2025; Ahangar et al., 2025).

In response to concerns about model opacity, the field of explainable artificial intelligence (XAI) has gained attention as a means of improving transparency and interpretability in machine learning systems. Explainability techniques aim to clarify which features influenced a prediction, how a model arrived at a decision, or why a particular behaviour was flagged as anomalous (Shifa et al., 2025; Tzionis et al., 2025; Attai et al., 2025a; Amannah et al., 2025; Attai et al., 2025b). In the context of insider threat detection, explainability is particularly important because alerts often involve distinct behavioural patterns rather than clear-cut rule violations. Current research often emphasises improving model accuracy while treating explainability as an auxiliary feature rather than a central design requirement. Many explanation methods focus narrowly on feature importance scores or post-hoc model interpretations, which may not provide meaningful insight into behavioural intent, contextual relevance, or risk severity (Kabir et al., 2025; Nastoska et al., 2025). Furthermore, explanations are frequently designed from a technical perspective, without sufficient consideration of how security analysts interpret, trust, and operationalise them. This disconnect creates what can be described as an explainability gap, a mismatch between what machine learning models can technically explain and what human decision-makers need to trust and act upon those explanations.

This paper adopts a critical perspective on explainability in machine learning-based insider threat detection. Rather than surveying algorithms exhaustively, it examines the broader conceptual and practical limitations that hinder explainable systems from achieving genuine trust and usability. Specifically, the paper explores how explainability gaps emerge across three interconnected dimensions: technical limitations in behaviour modelling and explanation methods, human trust challenges in analyst interpretation, and operational barriers to real-world deployment. By shifting attention from model-centric transparency toward trust-aware and context-sensitive design, this review highlights the need for more human-centred approaches to explainable insider threat detection. Through this analysis, the paper contributes a focused and critical understanding of why explainability in insider threat detection remains an unresolved challenge and identifies directions for developing systems that are not only accurate but also interpretable, trustworthy, and practically deployable.

2. Behaviour Modelling in Insider Threat Detection

At the core of machine learning-based insider threat detection lies behaviour modelling. Unlike traditional cybersecurity threats that often involve identifiable external attack signatures, insider threats manifest through patterns of legitimate system use (Inayat et al., 2024). As a result, detection mechanisms must focus not merely on identifying rule violations, but on analysing how users interact with organisational systems over time. Behaviour modelling, therefore, aims to construct representations of normal and

abnormal user activities, enabling the identification of subtle deviations that may indicate malicious intent, negligence, or credential compromise (Qawasmeh et al., 2025).

2.1 Nature of Behavioural Data

Insider threat detection systems are fundamentally data-driven, relying on continuous streams of behavioural data generated through employees' routine interactions with organisational systems. Unlike external cyberattacks that may involve identifiable intrusion signatures, insider threats often manifest as subtle deviations within legitimate access boundaries (Zeng et al., 2023). Consequently, detection mechanisms depend on detailed logging of user activities to construct behavioural baselines and identify anomalous patterns over time (Zhao et al., 2025). This data is typically high-volume, high-velocity, and heterogeneous, requiring sophisticated preprocessing and modelling techniques.

Authentication logs represent one of the primary data sources in insider threat detection. These logs record login frequency, timestamps, geographic or network locations, failed login attempts, and session durations (Al-Mhiqani et al., 2020). Deviations such as unusual login hours, repeated failed attempts, or access from atypical locations may signal compromised credentials or malicious intent. However, such anomalies must be interpreted carefully, as legitimate role changes or remote work arrangements can produce similar patterns.

File access records provide insight into how users interact with organisational data assets. These logs track downloads, uploads, modifications, deletions, and transfers of files, particularly those classified as sensitive or confidential. Sudden spikes in bulk downloads, repeated access to restricted directories, or unusual copying behaviour may indicate potential data exfiltration. Because insiders often operate within authorised access rights, analysing file interaction patterns relative to historical behaviour is critical. Email and communication metadata, such as sender-recipient relationships, frequency of communication, attachment transfers, and external domain interactions, offer additional behavioural indicators (Sasi et al., 2024). While content inspection may raise privacy concerns, metadata analysis can reveal shifts in communication networks, unusual external contacts, or abnormal attachment activity. Such patterns may precede intellectual property theft or information leakage.

Network activity logs further contribute to behavioural profiling by capturing data flows, accessed servers, browsing activity, and unusual outbound traffic volumes. Abnormal bandwidth usage, connections to unrecognised domains, or encrypted outbound transfers may raise suspicion, particularly when correlated with file access activity (Okolie et al., 2025). Network-level analysis is especially valuable for detecting covert data exfiltration attempts. Device usage patterns provide information about how and when endpoints are utilised. This includes USB insertions, peripheral device usage, remote desktop sessions, and workstation activity durations. Unauthorised use of removable storage devices or extended after-hours workstation sessions may signal attempts to extract sensitive information outside standard oversight (Wasserman & Wasserman, 2022). Application usage statistics help identify deviations in software interaction behaviour, such as accessing tools unrelated to an employee's job role, repeated database queries outside routine tasks, or sudden use of administrative utilities, which may indicate suspicious activity. When combined with role-based expectations, application usage patterns can help contextualise anomalies (Mohamed, 2025). Privilege escalation events are particularly significant in insider threat detection. Attempts to gain elevated access rights, modify permissions, or bypass security controls may indicate malicious preparation stages (Inayat et al., 2024). Monitoring such events is critical because privilege abuse often precedes significant security breaches.

Collectively, these diverse data sources enable the construction of comprehensive behavioural profiles. However, their volume and complexity also introduce challenges related to data integration, noise reduction, contextual interpretation, and privacy protection. Effective insider threat detection, therefore, depends not only on collecting behavioural data but also on modelling it in ways that capture temporal patterns, contextual relevance, and subtle deviations from normative behaviour. This data is often high-dimensional, heterogeneous, and temporally structured. Unlike static classification tasks, insider detection must account for sequential and evolving behaviour, as malicious intent may develop gradually or manifest in small increments over time. Consequently, behaviour modelling is not simply about

identifying single anomalous events, but about detecting meaningful deviations across behavioural trajectories.

2.2 Modelling Normal Behaviour

A fundamental assumption in many insider detection systems is that malicious behaviour represents a deviation from established norms. Therefore, behaviour modelling often begins by constructing baseline profiles of “normal” user activity. Behavioural profiles in insider threat detection can be constructed at multiple levels of granularity, depending on the modelling objective and the organisational structure. Each level offers distinct advantages and trade-offs in terms of sensitivity, scalability, and contextual relevance (Feng et al., 2025). Individual-level modelling focuses on building user-specific baselines derived from a person’s historical activity patterns. The system learns what is “normal” for a particular employee, such as typical login times, file access frequency, communication patterns, and application usage, and flags deviations from that personal norm (Sasi et al., 2024). This approach is highly sensitive to subtle behavioural changes and is particularly effective for detecting gradual insider activity.

However, it requires sufficient historical data for each user and may be computationally intensive in large organisations. Additionally, legitimate changes in job responsibilities or work schedules can temporarily distort personal baselines, potentially increasing false positives. Role-based modelling groups users according to job function, responsibilities, or access privileges. Instead of comparing a user only to their own past behaviour, the system evaluates whether their activity aligns with others performing similar roles (Nyame & Qin, 2020). By establishing normative patterns at the role level, detection systems can identify users whose activities deviate significantly from peer behaviour. This approach improves scalability and provides contextual grounding, but it may overlook gradual misuse if an individual’s behaviour shifts slowly while remaining broadly within role-based expectations. Departmental or organisational-level modelling aggregates behaviour across larger units, such as departments or the entire enterprise. This broader perspective is useful for detecting systemic anomalies, coordinated insider activities, or organisation-wide shifts in behaviour (Pourhabibi et al., 2020). It supports high-level risk monitoring and strategic oversight. However, because it operates at a coarse granularity, it is less effective at identifying subtle, user-specific deviations. Such models are often best used in combination with more granular approaches to provide layered detection coverage. In practice, effective insider threat detection systems often combine these modelling levels. Individual-level analysis captures personal behavioural drift, role-based modelling adds contextual grounding, and organisational-level profiling supports broader anomaly detection (Wadinger & Kvasnica, 2024). Together, they create a multi-layered behavioural monitoring framework capable of balancing precision, scalability, and contextual awareness.

Individual-level modelling captures personalised work patterns, such as preferred login times or frequently accessed files. Role-based modelling, on the other hand, identifies expected behaviour patterns for employees performing similar functions, which can help detect cross-domain access anomalies (Li et al., 2022; Odufisan et al., 2025). Both approaches attempt to contextualise behaviour rather than treat all users uniformly. However, defining “normal” behaviour in organisational environments is inherently complex. Work patterns may shift due to new projects, promotions, remote work arrangements, or organisational restructuring. What appears anomalous statistically may, in fact, reflect legitimate operational change. This dynamic nature of behaviour introduces ambiguity into modelling processes and complicates interpretation.

2.3 Supervised vs. Unsupervised Behaviour Modelling

Machine learning approaches to behaviour modelling typically fall into supervised, unsupervised, or semi-supervised paradigms. In supervised behaviour modelling, labelled datasets containing confirmed insider incidents are used to train classification models. The system learns to differentiate between benign and malicious behavioural patterns based on historical examples (Al-Mhiqani et al., 2020). While this approach can achieve strong predictive performance under controlled conditions, its effectiveness depends heavily on the availability and quality of labelled insider incidents, which are often scarce and

imbalanced. In contrast, unsupervised behaviour modelling focuses on anomaly detection. Rather than learning explicit malicious patterns, these models learn normal behavioural distributions and flag deviations (Pinto et al., 2024). Clustering, isolation-based methods, and reconstruction-based models are frequently used to identify outliers in user activity. Although this approach addresses the scarcity of labelled data, it often produces a high number of benign anomalies, which can overwhelm analysts. Semi-supervised models attempt to balance these approaches by leveraging limited labelled data to refine anomaly detection boundaries (Sun et al., 2025). This allows models to incorporate domain knowledge while still capturing evolving behavioural patterns. Despite differences in methodology, all paradigms share a central objective, which is to translate complex behavioural activity into measurable features that can be analysed algorithmically.

2.4 Feature Engineering and Representation Challenges

The success of ML-based insider threat detection depends not only on the choice of algorithms but also on how behavioural data is transformed into meaningful features. Raw system logs are typically noisy, unstructured, and high-dimensional. Feature engineering, therefore, plays a critical role in translating low-level activity records into structured representations that models can learn from (Prakash & Rella, 2019; Atiea et al., 2025). Poorly designed features may obscure important behavioural signals, whereas well-crafted features can significantly enhance detection accuracy and interpretability.

Frequency-based metrics are among the most commonly used features. These include counts such as the number of file downloads per day, login attempts per session, or emails sent within a given time window (Qian & Cong, 2024; Amamra et al., 2025). Frequency features are intuitive and easy to compute, making them attractive for baseline modelling. However, frequency alone may not capture the context or intent behind actions. A high number of downloads, for instance, may be normal during project deadlines but suspicious in other contexts. Thus, frequency metrics are most effective when combined with temporal and contextual information.

Temporal features capture when activities occur, such as logins outside normal working hours, weekend system access, or unusually long session durations. Since insider threats often involve behavioural shifts over time, temporal characteristics are critical for detecting deviations from established work patterns (Mohamed, 2025). However, defining “normal working hours” can be complex in modern organisations with remote work, flexible schedules, and global teams. Temporal modelling must therefore account for dynamic and role-specific time patterns to avoid false alarms.

Statistical deviation features measure how far current behaviour diverges from historical baselines. These may include z-scores, variance measures, or anomaly scores derived from previous activity distributions (Ekle & Eberle, 2024). Such features are particularly useful in individual-level modelling, where deviations from personal norms can indicate potential risk. Nevertheless, statistical deviations depend heavily on the quality and stability of baseline data. Sudden legitimate changes, such as promotions or departmental transfers, can produce misleading deviation signals if models are not updated adaptively.

Sequence-based patterns focus on the order and progression of actions rather than isolated events, such as the sequence of system commands executed, the progression from privilege escalation to bulk file access, or the combination of login and file transfer activities can reveal more complex behavioural signatures (Maghanaki et al., 2025). Sequence modelling captures procedural intent and is valuable for identifying multi-step insider attack patterns. However, representing sequential data requires more advanced modelling techniques and increases computational complexity.

Contextual indicators enrich features with semantic meaning. These may include the sensitivity classification of accessed files, the criticality of target systems, or the user’s organisational role (Le et al., 2025). Contextual features help distinguish between routine and high-risk behaviours by embedding organisational knowledge into the modelling process. For example, accessing a highly confidential database carries different risk implications depending on the user’s role and authorisation level. While contextualization improves detection relevance, it also introduces challenges related to data integration, dynamic role changes, and privacy considerations.

3. Machine Learning and Explainability in Insider Threat Detection

The increasing reliance on machine learning for insider threat detection has significantly enhanced the ability of organisations to identify complex and subtle behavioural anomalies. However, as detection models grow more sophisticated, they also become more opaque. Many high-performing algorithms, particularly ensemble methods and deep neural networks, operate as computational “black boxes,” generating predictions without transparent reasoning processes (Hassija et al., 2024). In high-stakes security environments where decisions can impact employee reputations, legal outcomes, and organisational integrity, such opacity poses substantial challenges. Explainability in machine learning refers to the extent to which the internal mechanics of a model or the rationale behind its predictions can be understood by human users (Bobes-Bascarán et al., 2026). In insider threat detection, explainability serves not only a technical function but also an operational and ethical one. Security analysts must evaluate alerts, justify investigations, and determine appropriate response actions. Therefore, understanding why a system has flagged a particular user is as important as whether the prediction is accurate.

3.1 The Black-Box Nature of Machine Learning Models

In insider threat detection, ML models are often selected for their ability to achieve high predictive accuracy in complex and high-dimensional environments. ML algorithms can capture intricate relationships between user actions, temporal patterns, and contextual features. However, this predictive strength frequently comes at the cost of interpretability, giving rise to what is commonly described as the “black-box” problem (Hassija et al., 2024). Ensemble models, such as random forests and gradient boosting machines, operate by aggregating predictions from multiple decision trees. While a single decision tree can be relatively interpretable, an ensemble composed of dozens of trees becomes significantly more difficult to analyse (Yang & Wang, 2025; Azad et al., 2025). Each tree may contribute differently to the final prediction, and tracing how specific behavioural features influenced a particular alert can be complex. As a result, security analysts may receive a risk score without a clear, intuitive explanation of the reasoning behind it. Neural networks, particularly deep learning architectures, further intensify this opacity. These models transform input features through multiple hidden layers, learning high-dimensional representations of behavioural data (Mienye & Swart, 2024). Although this enables the detection of subtle and non-linear patterns, the learned internal structures often do not correspond to easily understandable behavioural concepts. The relationship between raw user activity and the final prediction becomes embedded within layers of mathematical transformations, making it difficult to articulate why a specific user was flagged as suspicious.

3.2 Forms of Explainability in Insider Threat Detection

3.2.1 Model-Level Explainability

Model-level, or global, explainability focuses on providing a high-level understanding of how a machine learning model operates across the entire dataset, rather than explaining individual predictions. The goal is to identify which features or inputs generally drive the model’s decisions and to gain insight into the overall decision-making logic (Linardatos et al., 2021). Common techniques include feature importance rankings, which quantify the relative influence of each input variable; partial dependence plots, which illustrate how changes in a specific feature affect predicted risk; and decision rules extracted from trained models, which summarise the patterns learned by the model in an interpretable form (Mathotaarachchi et al., 2024). In the context of insider threat detection, global explanations help security teams understand which behavioural indicators, such as frequent file downloads, unusual login times, or access to sensitive data, are most strongly associated with flagged alerts. By highlighting these patterns, analysts gain a broad understanding of the model’s priorities and the features it considers most relevant for risk assessment. This knowledge can be useful for validating models, designing monitoring policies, and guiding resource allocation for investigations. However, global explanations have limitations. They provide an aggregate view of model behaviour and may overlook individual nuances. Insider threat

detection often involves context-specific scenarios, where subtle behavioural deviations matter more than general trends.

3.2.2 Instance-Level Explainability

Instance-level explanations focus on individual predictions. Post-hoc explanation methods can identify which features contributed most significantly to a specific alert (Kabir et al., 2025). For instance, a model may indicate that a user was flagged due to increased after-hours activity combined with access to sensitive files. Local explanations are particularly valuable in insider detection because analysts must assess specific incidents rather than general model trends. However, these explanations often quantify feature contributions without explaining the broader behavioural narrative. Knowing that “feature X contributed 35% to the risk score” may not sufficiently clarify whether the behaviour reflects malicious intent or legitimate work activity.

3.2.3 Post-Hoc vs. Intrinsic Interpretability

Explainability approaches in machine learning can be broadly categorised into intrinsic interpretability and post-hoc explainability, each with distinct advantages and limitations. Intrinsic interpretability refers to inherently transparent models, meaning their internal decision-making logic is easily understandable without additional explanation tools (Kabir et al., 2025). Models such as decision trees, logistic regression, or linear models have intrinsic interpretability because analysts can trace how input features contribute to outputs directly, allowing straightforward reasoning about predictions. This transparency supports trust, accountability, and operational adoption, particularly in high-stakes environments like insider threat detection. However, intrinsic models may struggle with complex, high-dimensional behavioural data. Their simplicity can limit predictive performance when dealing with subtle patterns, non-linear relationships, or sequences of user activity that characterise insider threats (Bin-Sarhan & Altwaijry, 2023). Post-hoc explainability, on the other hand, applies to more complex “black-box” models such as ensemble methods or deep neural networks. Here, explanations are generated after model training using techniques like feature attribution, surrogate models, or local explanation methods. Post-hoc approaches allow organisations to retain the high predictive accuracy of complex models while providing some interpretability (Alvanpour et al., 2025). However, because these explanations approximate the model’s behaviour rather than revealing its true internal reasoning, they may sometimes be misleading or incomplete. Analysts might receive plausible-sounding rationales that do not fully reflect how the model arrived at a particular alert.

3.3 Explainability Objectives in Insider Threat Contexts

Explainability in insider threat detection serves multiple critical objectives that go beyond simply making machine learning models transparent. Transparency is essential because analysts need to understand how predictions are generated, which features or patterns contributed to a risk assessment, and why a particular user or activity has been flagged. Transparent models help bridge the gap between automated detection and human interpretation, allowing analysts to trust the system’s outputs (Nastoska et al., 2025). Justifiability is a key objective, since explanations must support analyst decisions during investigations by providing actionable reasoning. When a model flags potentially malicious behaviour, analysts rely on explanations to determine whether further investigation is warranted, to identify the nature of the threat, and to make informed recommendations for mitigation. Without justifiable explanations, analysts may disregard alerts or make errors in judgment (Hermosilla et al., 2025). Trust-building represents a third objective because security analysts often operate under high workload and time pressure, evaluating numerous alerts daily. Explanations that are clear, consistent, and contextually meaningful increase confidence in automated predictions, enabling analysts to rely on the system without constant manual verification. Trust is therefore not merely about understanding model mechanics but about integrating predictive outputs into analyst workflows effectively (Czekster et al., 2025). Accountability is critical because insider threat detection directly impacts individuals within the organisation. Alerts may lead to interviews, disciplinary action, or legal consequences. Explainability ensures that decisions can be audited, defended, and aligned with regulatory requirements. Models must provide explanations that are

clear, defensible, and ethically sound, supporting responsible decision-making while protecting both the organisation and its employees (Saxena et al., 2020; Inayat et al., 2024). Unlike domains such as image recognition or recommendation systems, where errors primarily affect system performance or user experience, insider threat detection carries ethical, legal, and operational stakes. Therefore, explainability must simultaneously achieve transparency, justifiability, trust, and accountability to ensure that security interventions are both effective and responsible.

3.4 Limitations of Current Explainability Practices

Although explainable AI has gained attention in insider threat detection, existing approaches exhibit significant limitations that reduce their practical utility. One major limitation is that explanations are often restricted to feature attribution, highlighting which inputs influenced a model's prediction without providing meaningful behavioural reasoning (Aysel et al., 2025). Analysts may know that "late-night logins" or "large file downloads" contributed to a risk score, but these explanations rarely convey the context, intent, or sequence of actions that would clarify whether behaviour is genuinely suspicious. Without behavioural narratives, alerts remain abstract and harder to interpret. Another key limitation is the inadequate representation of temporal and contextual factors. Insider threats frequently unfold over time, involving gradual deviations in patterns or contextual anomalies tied to role changes, project deadlines, or organisational events (Prasad et al., 2025). Current explanation methods tend to focus on single-instance predictions, ignoring the temporal evolution of behaviour and the situational factors that help analysts judge risk. This gap limits the relevance and actionability of explanations in real-world investigations. Additionally, explanation outputs can be too technical for non-expert analysts. Metrics such as feature weights, latent embeddings, or principal component contributions may be understandable to ML researchers but are often opaque to security practitioners. When explanations are not aligned with human cognitive models or operational workflows, analysts may struggle to use them effectively, undermining trust in the system. There is a limited empirical evaluation of explainability effectiveness. Most studies in insider threat detection focus on improving detection accuracy or reducing false positives, paying little attention to whether explanations actually enhance analyst performance, decision confidence, or situational awareness. Consequently, explainability is frequently treated as an optional visualisation layer rather than a core requirement of system design. This reinforces the perception that explanations are supplementary, rather than integral, and highlights the need for research that evaluates both technical and human-centred dimensions of explainable insider threat detection.

3.5 Transition Toward a Critical Perspective

The integration of machine learning and explainability in insider threat detection raises a fundamental question: Does providing feature-level transparency truly resolve the trust and usability challenges faced by security analysts? While explainable AI offers promising tools, there remains a noticeable gap between algorithmic explanation mechanisms and the practical needs of operational environments. This gap forms the central focus of the next section. The following critical analysis examines how explainability in machine learning-based insider threat detection falls short across technical, human, and operational dimensions, and why bridging these gaps is essential for developing trustworthy and deployable systems.

4. Explainability Gaps in Machine Learning-Based Insider Threat Detection: A Critical Perspective

Although machine learning and explainable AI techniques have advanced insider threat detection capabilities, significant gaps remain between algorithmic transparency and practical trust. Current explainability approaches often address technical opacity without fully resolving deeper issues related to behavioural reasoning, human interpretation, and operational deployment. These shortcomings can be understood across three interconnected dimensions: technical gaps, human trust gaps, and operational gaps.

4.1 Technical Gaps

Most explainability mechanisms in insider threat detection focus on feature attribution, identifying which variables contributed to a prediction. While feature importance scores or local explanation outputs provide insight into model behaviour, they often fail to capture meaningful behavioural narratives. For

instance, a system may indicate that a user was flagged due to increased file downloads, late-night logins, and access to sensitive directories. However, such explanations remain largely descriptive rather than interpretive. They quantify correlations but do not explain intent, contextual relevance, or behavioural progression. Insider threats frequently emerge gradually, involving subtle behavioural drift rather than abrupt anomalies. Current explanation methods rarely account for temporal evolution or cumulative risk patterns, leaving analysts with fragmented insights rather than coherent behavioural stories. Another technical limitation arises from the use of high-dimensional features. Many models rely on latent embeddings, statistical deviations, or composite anomaly scores that are difficult to interpret intuitively. Even when explanations highlight influential features, those features may lack semantic clarity. For instance, referencing “anomaly score deviation from principal component” provides little actionable understanding. This divide between mathematical representation and human meaning constrains the usefulness of explainability outputs. Furthermore, explainability techniques are often applied post hoc, meaning they approximate the reasoning of complex models rather than reflect their true internal logic. This can introduce inconsistencies between the explanation and the underlying model decision process. In high-stakes environments such as insider investigations, approximate explanations may undermine confidence rather than enhance it.

4.2 Human Trust Gaps

Even when technical explanations are provided, they do not automatically translate into trust. Insider threat detection systems operate within Security Operations Centres (SOCs) or governance teams, where analysts must evaluate alerts quickly and under pressure. Explanations that are overly technical, fragmented, or context-free can increase cognitive burden rather than reduce it. A key issue is the mismatch between algorithmic reasoning and human reasoning. Machine learning models detect statistical irregularities, whereas analysts interpret behaviours in terms of organisational norms, role expectations, and intent (Ali et al., 2021). An anomaly may appear statistically significant but operationally benign. Without contextual enrichment, such as information about ongoing projects, job role transitions, or authorised access changes, analysts may struggle to determine whether flagged behaviour warrants investigation. Moreover, explanations often focus narrowly on numerical contribution values, assuming that transparency equates to trust (Schilke & Reimann, 2025). However, trust is shaped by factors such as consistency, clarity, contextual relevance, and perceived fairness. If analysts repeatedly encounter false positives or ambiguous explanations, confidence in the system may erode, regardless of reported accuracy metrics. There is also limited empirical research examining how analysts interpret and utilise explanations in practice. Many studies evaluate model performance using metrics such as accuracy, precision, and recall, but few assess whether explanations improve analyst decision-making, reduce investigation time, or enhance situational awareness. As a result, explainability is frequently validated from a technical standpoint rather than a human-centred one. This reveals a critical gap: current explainable models are often designed for interpretability in theory, but not necessarily for usability in operational security environments.

4.3 Operational Gaps

Beyond technical and human considerations, explainability gaps extend into operational deployment. Insider threat detection systems must integrate with existing security infrastructures, governance frameworks, and compliance requirements. However, many explainable AI solutions remain at the research prototype stage and are not optimised for real-world integration. One operational challenge involves scalability. Organisational environments generate vast quantities of behavioural data daily. Producing detailed, instance-level explanations for every flagged event may be computationally intensive and difficult to manage at scale. Additionally, presenting extensive explanation data to analysts may overwhelm dashboards rather than streamline workflows. Another issue concerns accountability and documentation. Insider investigations may involve legal scrutiny or human resource actions. Explanations must therefore be defensible, consistent, and auditable. If explanation mechanisms rely on probabilistic approximations or unstable post-hoc methods, maintaining consistent justification across cases becomes problematic (Ortigossa et al., 2025). There is also a tension between privacy and explainability. Detailed

behavioural explanations may expose sensitive employee activity information. Balancing transparency with privacy protection adds further complexity to system design. Explainability is often implemented as an add-on component rather than as an integrated design principle. When explanation mechanisms are layered onto models after development, they may not align with operational objectives or risk management strategies. This reinforces the perception that explainability is supplementary rather than foundational.

4.4 The Core Disconnect

Across these dimensions, a central theme emerges, which is that transparency does not automatically produce trust. While current approaches attempt to make models more interpretable, they often neglect the broader socio-technical ecosystem in which insider threat detection operates (Sunkara, 2025). Trust depends not only on understanding feature contributions but also on contextual alignment, reliability, consistency, and operational usefulness (Aquilino et al., 2025). In many cases, explainability efforts focus on making models understandable to developers rather than to practitioners. This creates a persistent gap between algorithmic explanation and actionable intelligence. Bridging this gap requires shifting from model-centric transparency toward human-centred and deployment-oriented design strategies.

5. Future Directions and Research Opportunities

Addressing explainability gaps in machine learning-based insider threat detection requires more than incremental improvements in model transparency. As discussed in the previous section, the limitations are not purely technical; they involve human interpretation, organisational context, and operational integration. Future research must therefore adopt a more holistic and human-centred perspective if explainability is to meaningfully enhance trust and usability.

5.1 Toward Human-Centred Explainability

A major future direction for insider threat detection is the transition from model-centric to human-centred explainability. Many current systems emphasise mathematical transparency, such as feature importance scores or contribution weights, without fully considering how security analysts actually interpret and operationalise these outputs. While such metrics may clarify internal model behaviour, they do not necessarily align with the cognitive processes, time constraints, and investigative workflows of practitioners. As a result, explanations may be technically accurate but practically underutilised. Human-centred explainability requires designing systems around the needs of analysts rather than the structure of algorithms. This means understanding how analysts evaluate alerts, what contextual information they require, and how explanations influence their decisions. Future research should therefore move beyond measuring predictive performance alone and instead assess practical outcomes such as reductions in investigation time, improvements in decision confidence, decreases in false positive fatigue, and enhancements in situational awareness. These human-centred metrics provide a more realistic assessment of whether explainability adds operational value. Additionally, explanation interfaces should move beyond isolated feature contributions and instead provide coherent behavioural narratives. Rather than stating that “after-hours access contributed 30% to risk,” systems could present structured summaries that compare current activity to historical baselines, highlight deviations relative to role expectations, and situate behaviour within a broader organisational context. Narrative-driven explanations are more consistent with how humans reason about intent and risk, and they may therefore foster stronger trust and more effective decision-making. Ultimately, human-centred explainability acknowledges that insider threat detection is not solely a computational task but a socio-technical process. Explanations must support real analysts working in complex environments, ensuring that transparency translates into actionable insight rather than additional cognitive burden.

5.2 Context-Aware and Temporal Explanations

Insider threats rarely emerge as isolated events; instead, they often develop gradually through subtle behavioural drift over time. Employees may slowly increase data access, shift communication patterns, or escalate privileges in ways that appear benign when viewed individually but become suspicious when examined collectively. However, many current explainability techniques focus on single-instance

predictions, explaining why a particular action was flagged without accounting for its temporal progression. This snapshot-based perspective limits the ability of analysts to understand how risk accumulates over time. Future research should therefore prioritise temporal and context-aware explanations that reflect behavioural evolution. Sequential modelling approaches can help highlight how actions unfold in stages, revealing patterns such as privilege escalation followed by unusual file access. Instead of treating each alert independently, explanations could describe the progression of activity and its deviation from established baselines. This shift from static anomaly detection to dynamic behavioural storytelling would better mirror the real nature of insider threats. In addition, explanations should incorporate contextual enrichment, such as recent role changes, new project assignments, or authorised access modifications. For example, increased database access may be legitimate following a promotion but suspicious in the absence of such contextual factors. Embedding organisational context within explanations enables analysts to differentiate between operational shifts and potential malicious intent more effectively. Visual tools, such as risk trajectory representations, could further support understanding by illustrating cumulative deviations over time rather than isolated risk scores. By showing how risk evolves, these visualisations help analysts see patterns and assess whether behaviour reflects temporary irregularities or sustained escalation.

5.3 Hybrid and Analyst-in-the-Loop Frameworks

A promising future direction in insider threat detection is the development of hybrid, analyst-in-the-loop frameworks that integrate machine learning with human expertise. Rather than positioning ML systems as fully autonomous decision-makers, these frameworks emphasise collaboration between automated detection models and security analysts. This approach recognises that while algorithms excel at identifying statistical irregularities across large datasets, human analysts contribute contextual judgment, organisational awareness, and investigative reasoning that models alone cannot replicate. In such systems, analysts could provide structured feedback on false positives or confirmed cases, allowing models to refine behavioural baselines and reduce recurring errors. For instance, if certain after-hours activities are repeatedly deemed legitimate due to specific project requirements, the system could adjust its thresholds accordingly. This iterative feedback loop enhances adaptability and ensures that detection mechanisms remain aligned with evolving organisational practices. Additionally, explanations could become interactive rather than static. Analysts might query the system for clarification, request comparisons with historical behaviour, or explore alternative contributing factors. Adaptive explanation interfaces could respond dynamically, tailoring information based on analyst needs and expertise levels. This interactivity transforms explanations from passive outputs into investigative tools. Systems could also learn from investigation outcomes, incorporating case resolutions to improve future explanation relevance. By analysing which explanatory elements were most useful in past decisions, models can prioritise similar contextual insights in subsequent alerts.

5.4 Standardised Evaluation of Explainability

A significant research gap in insider threat detection lies in the lack of standardised methods for evaluating explainability. While most studies rigorously assess predictive performance using metrics such as accuracy, precision, recall, and F1-score, far less attention is given to how well explanations function in practice. As a result, explainability is often assumed to be beneficial without being systematically measured. This imbalance creates uncertainty about whether explanatory components genuinely improve analyst understanding or merely provide superficial transparency. Future research should develop structured evaluation frameworks that assess explanation quality across multiple dimensions. One important criterion is clarity and coherence; explanations should be understandable, logically structured, and free from unnecessary technical jargon. Another critical factor is stability and consistency, ensuring that similar cases produce similar explanations. Inconsistent outputs may undermine trust, particularly in sensitive investigations where fairness and defensibility are essential. Explanations should also demonstrate alignment with domain knowledge. If model explanations contradict established organisational policies or expert understanding of user behaviour, analysts may question their validity. Ensuring that explanations reflect realistic behavioural reasoning strengthens both credibility and

practical relevance. Evaluation must consider the impact on human decision-making. This includes measuring whether explanations improve analyst accuracy, reduce investigation time, increase confidence, or lower cognitive workload. Without empirical evidence of such benefits, explainability remains largely theoretical. In the absence of robust and standardised evaluation methodologies, explainability risks being treated as a conceptual enhancement rather than a demonstrable improvement. Establishing measurable benchmarks is therefore essential to ensuring that explainable insider threat detection systems deliver tangible operational value.

5.5 Balancing Transparency, Privacy, and Ethics

Finally, future work must address the ethical and privacy implications of explainable insider threat detection systems. Detailed behavioural explanations may inadvertently expose sensitive employee information. Transparent systems must therefore balance accountability with data minimisation and fairness. Research should explore privacy-preserving explanation mechanisms that maintain interpretability while protecting personal data. Additionally, ethical guidelines should be incorporated into system design to ensure that explanations do not lead to unjust profiling or biased decision-making.

5.6 Moving Beyond Transparency Toward Trustworthy Systems

Improving explainability in insider threat detection requires moving beyond the narrow goal of transparency toward the broader objective of trustworthy system design. Trust emerges when systems are accurate, understandable, consistent, context-aware, and aligned with organisational practices. It is shaped not only by algorithmic clarity but also by reliability, usability, and fairness. Future research must therefore integrate technical innovation with human factors engineering, workflow design, and governance considerations. Only by addressing these interdependent dimensions can explainable machine learning systems transition from theoretical prototypes to trusted components of operational insider threat detection frameworks.

CONCLUSION

Machine learning has significantly advanced insider threat detection by enabling organisations to analyse vast volumes of behavioural data and identify subtle deviations that may signal malicious activity. However, as detection systems have become more sophisticated, they have also grown increasingly opaque. This paper critically examined the resulting explainability gaps in machine learning-based insider threat detection and argued that current approaches remain insufficient for establishing genuine human trust and operational reliability. While explainable AI techniques have improved transparency through feature attribution methods, local explanation tools, and interpretable model designs, these solutions often address only surface-level opacity. They frequently clarify which variables influenced a prediction but fail to provide meaningful behavioural narratives or contextual reasoning. As a result, there remains a disconnect between algorithmic outputs and the interpretive needs of security analysts. Explanations that are mathematically precise but cognitively misaligned do not necessarily enhance trust or decision quality. The analysis highlighted three major gaps: technical limitations in translating statistical anomalies into coherent behavioural reasoning, human trust gaps stemming from cognitive burden and contextual ambiguity, and operational challenges related to scalability, accountability, and integration within organisational governance frameworks. Together, these gaps reveal that explainability in insider threat detection is not merely a technical problem but a socio-technical one. A central argument of this paper is that transparency alone does not guarantee trust. Trust emerges from systems that are understandable, context-aware, consistent, and aligned with human workflows. Bridging the explainability gap, therefore, requires a shift from model-centric transparency toward human-centred and deployment-oriented design strategies. This includes integrating contextual enrichment, developing temporal explanations, establishing standardised evaluation metrics for explainability, and incorporating analyst-in-the-loop frameworks. Ultimately, insider threat detection operates at the intersection of technology, human behaviour, and organisational risk management. For explainable machine learning to fulfil its promise in this domain, future research must move beyond isolated interpretability techniques

and toward holistic, trustworthy system design. Only then can explainability transition from a theoretical enhancement to a practical foundation for reliable and accountable insider threat detection systems.

REFERENCES

- Ahangar, M. N., Farhat, Z. A., & Sivanathan, A. (2025). AI Trustworthiness in Manufacturing: Challenges, Toolkits, and the Path to Industry 5.0. *Sensors*, 25(14), 4357. <https://doi.org/10.3390/s25144357>
- Ali, M., Khattak, A. M., Ali, Z., Hayat, B., Idrees, M., Pervez, Z., Rizwan, K., Sung, T.-E., & Kim, K.-I. (2021). Estimation and Interpretation of Machine Learning Models with Customized Surrogate Model. *Electronics*, 10(23), 3045. <https://doi.org/10.3390/electronics10233045>
- Al-Mhiqani, M. N., Ahmad, R., Zainal Abidin, Z., Yassin, W., Hassan, A., Abdulkareem, K. H., Ali, N. S., & Yunus, Z. (2020). A Review of Insider Threat Detection: Classification, Machine Learning Techniques, Datasets, Open Challenges, and Recommendations. *Applied Sciences*, 10(15), 5208. <https://doi.org/10.3390/app10155208>
- Alvanpour, A., Acun, C., Spurlock, K., Robinson, C. K., Das, S. K., Popa, D. O., & Nasraoui, O. (2025). Comparative Analysis of Post Hoc Explainable Methods for Robotic Grasp Failure Prediction. *Electronics*, 14(9), 1868. <https://doi.org/10.3390/electronics14091868>
- Amamra, A., Anunwah, J. C., & Louafi, H. (2025). IoT Device Fingerprinting via Frequency Domain Analysis. *Electronics*, 14(16), 3248. <https://doi.org/10.3390/electronics14163248>
- Amannah, C., Attai, K. F., & Uzoka, F.-M. (2025). A Data-Driven Intelligent Methodology for Developing Explainable Diagnostic Model for Febrile Diseases. *Algorithms*, 18(4), 190. <https://doi.org/10.3390/a18040190>
- Aquilino, L., Di Dio, C., Manzi, F., Massaro, D., Bisconti, P., & Marchetti, A. (2025). Decoding Trust in Artificial Intelligence: A Systematic Review of Quantitative Measures and Related Variables. *Informatics*, 12(3), 70. <https://doi.org/10.3390/informatics12030070>
- Asmar, M., & Tuqan, A. (2024). Integrating machine learning for sustaining cybersecurity in digital banks. *Heliyon*, 10(17). <https://doi.org/10.1016/j.heliyon.2024.e37571>
- Atiea, M. A., Reda, R., Ataya, S., & Ibrahim, M. (2025). Explainable AI and Feature Engineering for Machine-Learning-Driven Predictions of the Properties of Cu-Cr-Zr Alloys: A Hyperparameter Tuning and Model Stacking Approach. *Processes*, 13(5), 1451. <https://doi.org/10.3390/pr13051451>
- Attai, K. F., Amannah, C., Ekpenyong, M., Asuquo, D. E., Akputu, O. K., Obot, O. U., ... & Uzoka, F. M. (2025a). Developing an explainable artificial intelligence system for the mobile-based diagnosis of febrile diseases using random forest, LIME, and GPT. *Healthcare Informatics Research*, 31(2), 125-135. <https://doi.org/10.4258/hir.2025.31.2.125>
- Attai, K. F., Amannah, C., Ekpenyong, M., Baadel, S., Obot, O., Asuquo, D., Attai, E., Uzoka, F.-V., Dan, E., Akwaowo, C., & Uzoka, F.-M. (2025b). Predicting Predisposition to Tropical Diseases in Female Adults Using Risk Factors: An Explainable-Machine Learning Approach. *Information*, 16(7), 520. <https://doi.org/10.3390/info16070520>
- Aysel, H. I., Cai, X., & Prugel-Bennett, A. (2025). Explainable Artificial Intelligence: Advancements and Limitations. *Applied Sciences*, 15(13), 7261. <https://doi.org/10.3390/app15137261>
- Azad, M., Nehal, T. H., & Moshkov, M. (2025). A novel ensemble learning method using majority based voting of multiple selective decision trees. *Computing*, 107(1), 42.
- Bin-Sarhan, B., & Altwaijry, N. (2023). Insider Threat Detection Using Machine Learning Approach. *Applied Sciences*, 13(1), 259. <https://doi.org/10.3390/app13010259>
- Bobes-Bascarán, J., Mosqueira-Rey, E., Fernández-Leal, Á., Alonso-Ríos, D., Figueirido-Arnoso, I., & Vidal-Ínsua, Y. (2026). Evaluating Explanatory Capabilities of Machine Learning Models in Medical Diagnostics: A Human-in-the-Loop Approach. *Mathematics*, 14(3), 497. <https://doi.org/10.3390/math14030497>
- Czekster, R. M., Webber, T., Furstenu, L. B., & Marcon, C. (2025). Dynamic risk assessment approach for analysing cyber security events in medical IoT networks. *Internet of Things*, 29, 101437. <https://doi.org/10.1016/j.iot.2024.101437>
- Ekle, O. A., & Eberle, W. (2024). Anomaly detection in dynamic graphs: A comprehensive survey. *ACM Transactions on Knowledge Discovery from Data*, 18(8), 1-44.
- Feng, W., Cao, Y., Chen, Y., Wang, Y., Hu, N., Jia, Y., & Gu, Z. (2025). Multi-Granularity User Anomalous Behavior Detection. *Applied Sciences*, 15(1), 128. <https://doi.org/10.3390/app15010128>

- Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., ... & Hussain, A. (2024). Interpreting black-box models: a review on explainable artificial intelligence. *Cognitive Computation*, 16(1), 45-74. <https://doi.org/10.1007/s12559-023-10179-8>
- Hermosilla, P., Berríos, S., & Allende-Cid, H. (2025). Explainable AI for Forensic Analysis: A Comparative Study of SHAP and LIME in Intrusion Detection Models. *Applied Sciences*, 15(13), 7329. <https://doi.org/10.3390/app15137329>
- Inayat, U., Farzan, M., Mahmood, S., Zia, M. F., Hussain, S., & Pallonetto, F. (2024). Insider threat mitigation: Systematic literature review. *Ain Shams Engineering Journal*, 15(12). <https://doi.org/10.1016/j.asej.2024.103068>
- Kamatchi, K., & Uma, E. (2025). Insights into user behavioral-based insider threat detection: systematic review. *International Journal of Information Security*, 24(2), 88. <https://doi.org/10.1007/s10207-025-01002-6>
- Kabir, S., Hossain, M. S., & Andersson, K. (2025). A Review of Explainable Artificial Intelligence from the Perspectives of Challenges and Opportunities. *Algorithms*, 18(9), 556. <https://doi.org/10.3390/a18090556>
- Le, T. D., Le-Dinh, T., & Uwizeyemungu, S. (2025). Cybersecurity Analytics for the Enterprise Environment: A Systematic Literature Review. *Electronics*, 14(11), 2252. <https://doi.org/10.3390/electronics14112252>
- Li, Y., Du, Z., Fu, Y., & Liu, L. (2022). Role-Based Access Control Model for Inter-System Cross-Domain in Multi-Domain Environment. *Applied Sciences*, 12(24), 13036. <https://doi.org/10.3390/app122413036>
- Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy*, 23(1), 18. <https://doi.org/10.3390/e23010018>
- Liu, X., Huang, D., Yao, J., Dong, J., Song, L., Wang, H., Yao, C., & Chu, W. (2025). From Black Box to Glass Box: A Practical Review of Explainable Artificial Intelligence (XAI). *AI*, 6(11), 285. <https://doi.org/10.3390/ai6110285>
- Maghanaki, M., Keramati, S., Chen, F. F., & Shahin, M. (2025). Generation of a Multi-Class IoT Malware Dataset for Cybersecurity. *Electronics*, 14(21), 4196. <https://doi.org/10.3390/electronics14214196>
- Mathotaarachchi, K. V., Hasan, R., & Mahmood, S. (2024). Advanced Machine Learning Techniques for Predictive Modeling of Property Prices. *Information*, 15(6), 295. <https://doi.org/10.3390/info15060295>
- Mienye, I. D., & Swart, T. G. (2024). A Comprehensive Review of Deep Learning: Architectures, Recent Advances, and Applications. *Information*, 15(12), 755. <https://doi.org/10.3390/info15120755>
- Mitchell, T. (2025). Trust and Transparency in Artificial Intelligence: T. Mitchell. *Philosophy & Technology*, 38(3), 87. <https://doi.org/10.1007/s13347-025-00916-2>
- Mohamed, N. (2025). Artificial intelligence and machine learning in cybersecurity: a deep dive into state-of-the-art techniques and future paradigms. *Knowledge and Information Systems*, 67(8), 6969-7055. <https://doi.org/10.1007/s10115-025-02429-y>
- Nastoska, A., Jancheska, B., Rizinski, M., & Trajanov, D. (2025). Evaluating Trustworthiness in AI: Risks, Metrics, and Applications Across Industries. *Electronics*, 14(13), 2717. <https://doi.org/10.3390/electronics14132717>
- Nyame, G., & Qin, Z. (2020). Precursors of Role-Based Access Control Design in KMS: A Conceptual Framework. *Information*, 11(6), 334. <https://doi.org/10.3390/info11060334>
- Odufisan, O. I., Abhulimen, O. V., & Ogunti, E. O. (2025). Harnessing artificial intelligence and machine learning for fraud detection and prevention in Nigeria. *Journal of Economic Criminology*, 7, 100127. <https://doi.org/10.1016/j.jeconc.2025.100127>
- Okolie, S. A., Amadi, C. A., Odii, J. N., Nwokorie, E. C., & Onyemauche, U. C. (2025). Anomaly Detection in Heterogeneous Cybersecurity Data. *Franklin Open*. <https://doi.org/10.1016/j.fraope.2025.100426>
- Ortigossa, E. S., Dias, F. F., Barr, B., Silva, C. T., & Nonato, L. G. (2025). T-explainer: A model-agnostic explainability framework based on gradients. *IEEE Intelligent Systems*.
- Pinto, A., Herrera, L. C., Donoso, Y., & Gutierrez, J. A. (2024). Enhancing critical infrastructure security: Unsupervised learning approaches for anomaly detection. *International Journal of Computational Intelligence Systems*, 17(1), 236. <https://doi.org/10.1007/s44196-024-00644-z>
- Pitkar, H. (2025). Cloud Security Automation Through Symmetry: Threat Detection and Response. *Symmetry*, 17(6), 859. <https://doi.org/10.3390/sym17060859>
- Pourhabibi, T., Ong, K. L., Kam, B. H., & Boo, Y. L. (2020). Fraud detection: A systematic literature review of graph-based anomaly detection approaches. *Decision Support Systems*, 133. <https://doi.org/10.1016/j.dss.2020.113303>
- Prakash, B., & Rella, R. (2019). The role of feature engineering in machine learning: Techniques challenges and automation with data engineering. *International Journal of Data Engineering*, 2(9), 224-242.

- Prasad, N., Diro, A., Warren, M., & Fernando, M. (2025). A survey of cyber threat attribution: Challenges, techniques, and future directions. *Computers & Security*, 104606.
- Qawasmeh, S. A.-D., & AlQahtani, A. A. S. (2025). Beyond Firewall: Leveraging Machine Learning for Real-Time Insider Threats Identification and User Profiling. *Future Internet*, 17(2), 93. <https://doi.org/10.3390/fi17020093>
- Qian, L., & Cong, L. (2024). Channel Features and API Frequency-Based Transformer Model for Malware Identification. *Sensors*, 24(2), 580. <https://doi.org/10.3390/s24020580>
- Sasi, T., Lashkari, A. H., Lu, R., Xiong, P., & Iqbal, S. (2024). A comprehensive survey on IoT attacks: Taxonomy, detection mechanisms and challenges. *Journal of Information and intelligence*, 2(6), 455-513. <https://doi.org/10.1016/j.jiixd.2023.12.001>
- Schilke, O., & Reimann, M. (2025). The transparency dilemma: How AI disclosure erodes trust. *Organizational Behavior and Human Decision Processes*, 188, 104405. <https://doi.org/10.1016/j.obhdp.2025.104405>
- Saxena, N., Hayes, E., Bertino, E., Ojo, P., Choo, K.-K. R., & Burnap, P. (2020). Impact and Key Challenges of Insider Threats on Organizations and Critical Businesses. *Electronics*, 9(9), 1460. <https://doi.org/10.3390/electronics9091460>
- Shifa, N., Saleh, M., Akbari, Y., & Al Maadeed, S. (2025). A review of explainable AI techniques and their evaluation in mammography for breast cancer screening. *Clinical Imaging*, 123. <https://doi.org/10.1016/j.clinimag.2025.110492>
- Sun, Y., Keung, J. W., Yang, Z., Liu, S., & Liao, Y. (2025). SemiSMAC: A semi-supervised framework for log anomaly detection with automated hyperparameter tuning. *Information and Software Technology*. <https://doi.org/10.1016/j.infsof.2025.107869>
- Sunkara, G. (2025). Explainable AI for cyber threat Intelligence: Enhancing analyst trust.. <https://doi.org/10.53022/oarjst.2025.14.2.0091>
- Tzionis, G., Mouratidis, P., Kougka, G., Gialampoukidis, I., Vrochidis, S., Kompatsiaris, I., & Vlachopoulou, M. (2025). A review of explainable AI methods and their application in manufacturing systems. *Discover Applied Sciences*. <https://doi.org/10.1007/s42452-025-07908-z>
- Wadinger, M., & Kvasnica, M. (2024). Adaptable and Interpretable Framework for Anomaly Detection in SCADA-based industrial systems. *Expert Systems with Applications*, 246. <https://doi.org/10.1016/j.eswa.2024.123200>
- Wasserman, L., & Wasserman, Y. (2022). Hospital cybersecurity risks and gaps: Review (for the non-cyber professional). *Frontiers in digital health*, 4, 862221. <https://doi.org/10.3389/fdgth.2022.862221>
- Yang, Y., & Wang, H. (2025). Random Forest-Based Machine Failure Prediction: A Performance Comparison. *Applied Sciences*, 15(16), 8841. <https://doi.org/10.3390/app15168841>,
- Zeng, M., Dian, C., & Wei, Y. (2023). Risk Assessment of Insider Threats Based on IHFACS-BN. *Sustainability*, 15(1), 491. <https://doi.org/10.3390/su15010491>
- Zhao, X., Guo, K., Huang, M., Qiu, S., & Lu, L. (2025). ELFA-Log: Cross-System Log Anomaly Detection via Enhanced Pseudo-Labeling and Feature Alignment. *Computers*, 14(7), 272. <https://doi.org/10.3390/computers14070272>