

# Explainable AI for Malicious Insider Threat Detection

Komal Tiwari<sup>1</sup> Shikha Gupta<sup>2</sup> Gauri Ansurkar<sup>3</sup>

<sup>1,2</sup>Student <sup>3</sup>Assistant Professor

<sup>1,2,3</sup>Department of Information Technology

<sup>1,2,3</sup>Keraleeya Samajam's Model College, Thakur, Dombivli, Maharashtra, India

**Abstract** — Malicious insider threats are a significant cybersecurity challenge because authorized users already have legitimate access to organizational systems and resources. As a result, malicious insider activities can appear similar to normal user behaviour, making them difficult to distinguish using conventional security mechanisms. Artificial Intelligence (AI) and Machine Learning (ML) can support the analysis of user behavioural activities and identify patterns associated with potentially malicious behaviour. However, many machine learning models operate as black boxes, making it difficult for security analysts to understand why a particular activity has been classified as suspicious. This lack of interpretability can reduce trust in automated detection systems and affect security decision-making. This paper proposes a conceptual Explainable Artificial Intelligence (XAI)-based framework for malicious insider threat detection. The proposed framework considers multiple sources of user activity data, including logon, file, web, and email activities. After data preprocessing and behavioural feature extraction, XGBoost is proposed for supervised threat classification, while SHAP is used to explain the contribution of behavioural features to individual predictions. A Deep Autoencoder is proposed as a separate anomaly detection component operating on behavioural features to identify unusual or previously unseen activities that may not be adequately captured by supervised classification. LIME may additionally be used as a local explanation technique. The present study follows a conceptual framework design supported by an exploratory and descriptive survey-based research approach. A structured online questionnaire is used to examine respondents' perceptions regarding XAI, explainability, insider threat detection, behavioural features, trust, interpretability, and the perceived usefulness and potential effectiveness of the proposed framework. The study does not claim actual implementation or experimental validation of the proposed machine learning components. Technical implementation and dataset-based performance evaluation are considered future work. The proposed framework aims to provide a more transparent and understandable approach to malicious insider threat detection and support security analysts in making informed decisions.

**Keywords:** Malicious Insider Threat, Explainable Artificial Intelligence, XAI, Machine Learning, XGBoost, SHAP, LIME, Anomaly Detection, Behavioural Analysis, Cybersecurity

## I. INTRODUCTION

Organizations increasingly depend on digital systems, interconnected platforms, cloud services, and information resources. This dependence creates security challenges not only from external attackers but also from individuals who already possess legitimate access to organizational systems. A malicious insider may misuse authorized privileges to access, modify, disclose, or otherwise misuse organizational

resources. The detection of such behaviour is difficult because legitimate and malicious activities may occur within the same authorized environment.

Malicious insider threat detection therefore requires analysis of user behaviour rather than relying only on predefined signatures or conventional security rules. Previous research has identified online activities and behavioural information as important sources for insider threat detection. At the same time, insider threat detection faces challenges such as changing behavioural patterns, false alarms, class imbalance, uncertainty, and the difficulty of identifying previously unseen activities (Gheyas & Abdallah, 2016).

Artificial Intelligence (AI) and Machine Learning (ML) provide opportunities to analyze large volumes of user activity data and identify behavioural patterns associated with potentially malicious activities. In the proposed research, relevant user activities include logon activity, file activity, web activity, and email activity. These activities can provide different perspectives on how an authorized user interacts with organizational resources.

Although ML can support automated threat classification, complex models can be difficult for security analysts to interpret. A system may indicate that a particular activity is suspicious without clearly explaining which behavioural characteristics contributed to that decision. This black-box problem can make it difficult for analysts to assess whether a prediction is reasonable and can reduce confidence in automated security systems.

Explainable Artificial Intelligence (XAI) addresses this limitation by providing information about the reasoning behind model predictions. SHAP, for example, assigns feature-importance values to individual predictions, allowing the contribution of different features to be examined (Lundberg & Lee, 2017). LIME provides another approach for producing local explanations of individual predictions (Ribeiro et al., 2016).

Based on these considerations, this study proposes a conceptual XAI-based framework for malicious insider threat detection. The framework combines behavioural activity analysis, XGBoost-based supervised classification, SHAP-based explanations, and a separate Deep Autoencoder-based anomaly detection component. XGBoost is proposed as the primary classification model, while the Deep Autoencoder is proposed to examine behavioural patterns for possible anomalies.

The present study is conceptual and survey-based. It does not claim that the proposed models have already been trained, implemented, tested, or experimentally validated. Instead, the study proposes the framework and uses a structured survey to examine perceptions regarding explainability, trust, interpretability, usefulness, and potential effectiveness.

## II. RESEARCH QUESTIONS

- RQ1: How can Explainable Artificial Intelligence improve the interpretability of malicious insider threat detection?
- RQ2: How can behavioural data from logon, file, web, and email activities support malicious insider threat detection?
- RQ3: How can XGBoost and SHAP be incorporated into an explainable malicious insider threat detection framework?
- RQ4: How can explainable predictions improve security analysts' trust and decision-making in insider threat detection?
- RQ5: What is the perceived usefulness and potential effectiveness of the proposed XAI-based malicious insider threat detection framework?

## III. LITERATURE REVIEW / RELATED WORK

### A. Malicious Insider Threat Detection

Gheyas and Abdallah (2016) systematically reviewed insider threat detection and found that insiders' online activities are the most widely used features, while key challenges include class imbalance, high false alarm rates, changing behavioural patterns, and undetected attacks. Homoliak et al. (2019) provided a structural taxonomy of insider threat incidents and organized the field into incidents and datasets, attacker analysis, simulations, and defence solutions, together with an overview of publicly available datasets for evaluating detection solutions. These studies establish that behavioural analysis is central to insider threat detection, but that data scarcity and imbalance remain open problems.

### B. Machine Learning for Insider Threat Detection

Machine-learning-based detection can be supervised (classification of known patterns) or unsupervised (anomaly detection). Nasir et al. (2021) proposed a deep-learning approach using logon/logoff, user role, and functional unit features on the CERT r4.2 dataset, reporting 90.60% accuracy, 97% precision, and 94% F1-score, although the method still depends on labelled behavioural data. Chattopadhyay et al. (2018) performed time-series classification of user activities from logon, device, file, web, and e-mail logs using a deep autoencoder classifier and reported high precision, recall, and F-score on the CMU insider threat data, while noting that anomalous activity is not necessarily malicious. These studies show that combining multiple behavioural sources improves detection, but most methods rely on labelled data that is scarce for insider threats.

### C. Explainable AI in Cybersecurity

Zhang et al. (2022) surveyed XAI applications in cybersecurity and observed that most machine-learning and deep-learning models are deployed as black boxes, which reduces human users' confidence in AI-based defences. Explainability is therefore needed to help analysts validate predictions and build trust. However, the survey also notes that XAI methods are not yet widely integrated into operational security workflows.

### D. SHAP and LIME for Model Explainability

SHAP provides feature-attribution values indicating the contribution of each feature to an individual prediction (Lundberg & Lee, 2017), while LIME explains individual predictions by approximating a complex classifier locally (Ribeiro et al., 2016). Wang et al. (2020) combined local and global SHAP explanations for intrusion detection systems and showed that SHAP improves model transparency for security analysts, although their framework was evaluated on network intrusion data (NSL-KDD) rather than insider threat data.

### E. XGBoost for Threat Classification

XGBoost is an efficient and scalable gradient-boosted decision tree system (Chen & Guestrin, 2016). Jiang et al. (2018) applied XGBoost to multi-domain behavioural features extracted from audit logs and reported an F-measure up to 99.96% for insider threat identification, outperforming SVM and random forest. A limitation of such supervised approaches is that they require sufficient labelled malicious samples, which are rare in insider threat settings.

### F. Autoencoder-Based Anomaly Detection

Autoencoders learn compressed representations of input data by reconstructing it (Hinton & Salakhutdinov, 2006). Liu et al. (2018a) trained an ensemble of deep autoencoders on categories of audit data representing normal user behaviour and used reconstruction error to flag anomalous insider activity on a benchmark dataset. This unsupervised strategy is well suited to detecting previously unseen behaviour, but it can produce false positives because anomalous activity is not necessarily malicious.

### G. Explainable AI for Insider Threat Detection

Alketbi and Mehmood (2025) surveyed XAI techniques for malicious insider threat detection and concluded that, although XAI improves interpretability and analyst trust, challenges persist, including class imbalance, heterogeneous data streams, and the absence of standardized metrics for evaluating explainability. Abbasi et al. (2025) combined a bidirectional LSTM autoencoder with SHAP-based explanations on CERT v5.2 data, demonstrating that explanations help analysts assess and dismiss false positives, while noting that deep models remain difficult to interpret directly.

### H. Existing Limitations and Research Gap

The reviewed literature shows that supervised classifiers such as XGBoost and unsupervised autoencoders have each been applied to insider threat detection, and that SHAP and LIME can make predictions more interpretable. However, few studies integrate behavioural analysis, supervised classification, anomaly detection, and explainability within a single framework, and standardized evaluation of explainability in insider threat detection is still lacking (Alketbi & Mehmood, 2025). The research gap addressed by this study is therefore the need for a conceptual framework that combines multiple behavioural sources (logon, file, web, and email activities), uses XGBoost for supervised classification, SHAP (with optional LIME) for explanation, and a Deep Autoencoder for anomaly detection, while

examining the role of explainability in analyst trust and decision-making.

#### IV. RESEARCH GAP

Existing literature demonstrates the importance of behavioural information and machine-learning-based methods for insider threat detection. It also identifies anomaly detection as a useful strategy for addressing previously unseen or difficult-to-label insider activities. However, there remains a practical need for approaches that connect behavioural analysis, supervised classification, anomaly detection, and explainability within a single conceptual framework.

- The research gap addressed by this study is therefore the need for a conceptual framework that:
- combines multiple user behavioural sources such as logon, file, web, and email activities;
- uses supervised classification for known behavioural patterns;
- incorporates XAI to make classification decisions understandable;
- separately considers anomaly detection for unusual or previously unseen behaviour; and
- considers the role of explainability in analyst trust, interpretability, and security decision-making.

The proposed framework addresses this gap by combining XGBoost, SHAP, and Deep Autoencoder-based anomaly detection while using a survey-based approach to examine perceptions of its usefulness and potential effectiveness.

#### V. CONCEPT OF XAI-BASED MALICIOUS INSIDER THREAT DETECTION

Explainable Artificial Intelligence refers to approaches that make AI and ML predictions more understandable to human users. In cybersecurity, explainability can help analysts understand the behavioural factors associated with a suspicious prediction.

This is particularly important for malicious insider threat detection because insiders already possess legitimate access to organizational systems. Their malicious activities may therefore appear similar to normal activities. The proposed approach uses multiple user activity sources to identify behavioural patterns.

- The proposed conceptual approach follows these principles:
- Analysis of multiple user activity sources.
- Behavioural feature extraction.
- Supervised threat classification using XGBoost.
- Explanation of XGBoost predictions using SHAP.
- Separate anomaly detection using a Deep Autoencoder.
- Optional local explanations using LIME.
- Generation of an explainable insider threat detection output.

#### VI. PROPOSED CONCEPTUAL FRAMEWORK

The proposed framework is a conceptual architecture rather than an experimentally implemented system.

##### A. User Activity Data

The framework considers four primary sources of user activity: logon activity, file activity, web activity, and email activity. These sources provide different types of information about user behaviour within an organizational environment.

##### B. Data Preprocessing

The collected activity data would first undergo preprocessing. This stage may include handling inconsistent information, preparing records for analysis, and organizing activity information into a suitable form for behavioural feature extraction.

##### C. Behavioural Feature Extraction

Relevant behavioural features would be derived from the activity logs. These features would represent patterns of user interaction with organizational systems and resources. The exact feature set would need to be determined during future technical implementation.

##### D. XGBoost Threat Classification

XGBoost is proposed as the supervised classification component of the framework. Its conceptual role is to analyze behavioural features and classify activities according to learned patterns of potentially malicious or normal behaviour. No claim is made that an XGBoost model has been trained or tested in the present study.

##### E. SHAP Explanation Layer

SHAP is proposed as the primary explanation technique for XGBoost predictions. For an individual prediction, SHAP can provide information about the contribution of different behavioural features to the model's output. This can help a security analyst understand why an activity was classified as potentially malicious.

##### F. Deep Autoencoder for Anomaly Detection

The Deep Autoencoder is proposed as a separate anomaly detection component operating on behavioural features. It is not proposed to operate directly on SHAP values. Instead, the autoencoder would examine behavioural representations and identify activities that differ substantially from learned patterns. This component is intended to complement supervised classification. XGBoost can be used for known classification patterns, while anomaly detection may provide an additional mechanism for identifying unusual or previously unseen behaviour.

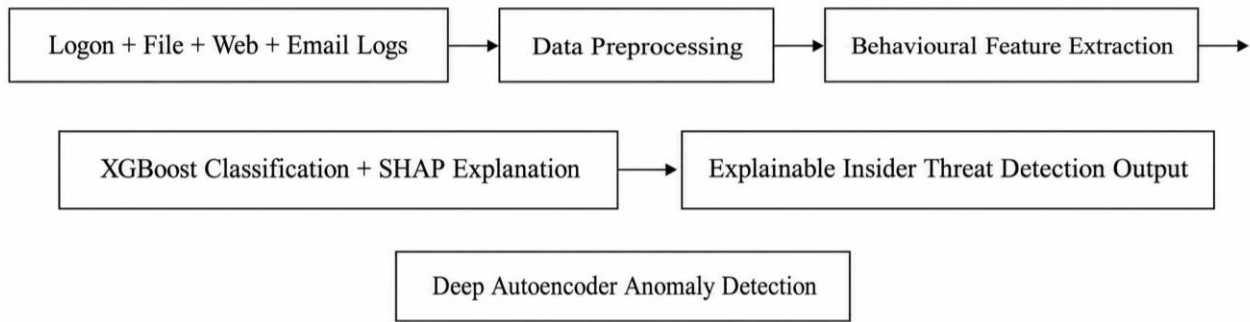
##### G. LIME as an Optional Explanation Technique

LIME is retained as an optional XAI technique. It may provide local explanations for individual predictions and can potentially be used in future work to compare explanation approaches with SHAP. SHAP remains the primary explanation technique in the proposed framework.

##### H. Explainable Insider Threat Detection Output

The final conceptual output should provide an indication of potentially malicious behaviour, indication of anomalous behaviour where applicable, important behavioural features, explanation of classification decisions, and information that can support analyst interpretation and decision-making.

## I. Proposed Framework Architecture



Logon + File + Web + Email Logs Data Preprocessing Behavioural Feature Extraction [XGBoost Classification + SHAP Explanation] & [Deep Autoencoder Anomaly Detection] Explainable Insider Threat Detection Output

## VII. OPERATIONAL SCENARIO

In a typical organizational environment, authorized users continuously generate activity records through system logons, file access, web activity, and email communication. The proposed framework would first organize and preprocess these activity records. Behavioural features would then be extracted from the available activity data.

The XGBoost component would conceptually classify behavioural patterns based on learned patterns. SHAP would then provide an explanation of the behavioural features contributing to an individual classification. At the same time, the Deep Autoencoder would examine behavioural features for unusual patterns. This separate anomaly detection path could potentially identify behaviour that differs from expected patterns and may not be adequately represented in supervised classification.

The final conceptual output would combine the classification and anomaly information and provide interpretable information for security analysts. LIME may additionally be considered when a local explanation of an individual prediction is required. This scenario represents the proposed operation of the framework, not an experimentally validated implementation.

## VIII. RESEARCH METHODOLOGY

### A. Research Design

The study follows a conceptual framework design supported by an exploratory and descriptive survey-based research approach. The research has two connected components: (1) a conceptual framework component in which an XAI-based framework is proposed for malicious insider threat detection, and (2) a survey component in which a structured online questionnaire is used to understand respondents' perceptions.

### B. Conceptual Framework Development

- User activity data collection.
- Data preprocessing.
- Behavioural feature extraction.
- Proposed XGBoost-based classification.
- SHAP-based explanation.
- Proposed Deep Autoencoder-based anomaly detection.
- Optional LIME-based local explanation.
- Explainable insider threat detection output.

### C. Survey-Based Study

The structured survey examines respondents' perceptions regarding awareness of AI and XAI, importance of explainability, malicious insider threat detection, behavioural activity features, usefulness of SHAP explanations, usefulness of XGBoost, anomaly detection, trust, interpretability, usefulness, and perceived potential effectiveness.

### D. Scope of the Present Study

The present study does not claim actual XGBoost training, actual SHAP implementation, classification results, actual Deep Autoencoder training, use of a specific technical dataset, achieved classification results, achieved anomaly detection results, or experimental validation of the framework. Actual technical implementation and performance evaluation are reserved for future work.

## IX. SURVEY QUESTIONNAIRE

The questionnaire is separate from the research questions. The following questions are intended for a structured online survey and can be adapted for Google Forms.

- Q1. What is your current level of knowledge or experience in Artificial Intelligence and cybersecurity?  
Options: Very Low; Low; Moderate; High; Very High
- Q2. How familiar are you with Explainable Artificial Intelligence (XAI)?  
Options: Not at all familiar; Slightly familiar; Moderately familiar; Very familiar; Extremely familiar
- Q3. How important is explainability when using AI-based systems for malicious insider threat detection?  
Options: Not Important; Slightly Important; Moderately Important; Very Important; Extremely Important
- Q4. How effective do you think traditional security methods are in detecting malicious insider activities?  
Options: Very Ineffective; Ineffective; Neutral; Effective; Very Effective
- Q5. To what extent do you agree that AI/ML can improve malicious insider threat detection?  
Options: Strongly Disagree; Disagree; Neutral; Agree; Strongly Agree
- Q6. How useful would explanations of AI predictions be for understanding why a user's activity is classified as potentially malicious?

Options: Not Useful; Slightly Useful; Moderately Useful; Very Useful; Extremely Useful

- Q7. How important is the analysis of logon, file, web, and email activities for identifying potentially malicious insider behaviour?

Options: Not Important; Slightly Important; Moderately Important; Very Important; Extremely Important

- Q8. To what extent do you agree that SHAP explanations can help identify behavioural features contributing to an insider threat prediction?

Options: Strongly Disagree; Disagree; Neutral; Agree; Strongly Agree

- Q9. To what extent do you agree that anomaly detection can help identify unusual or previously unseen insider activities?

Options: Strongly Disagree; Disagree; Neutral; Agree; Strongly Agree

- Q10. To what extent do you agree that XGBoost can be useful for classifying behavioural patterns associated with malicious insider threats?

Options: Strongly Disagree; Disagree; Neutral; Agree; Strongly Agree

- Q11. To what extent do you agree that an AI-based insider threat detection system would be more trustworthy if it provided clear explanations for its predictions?

Options: Strongly Disagree; Disagree; Neutral; Agree; Strongly Agree

- Q12. To what extent do you agree that XAI explanations can improve your understanding of why a particular user activity is classified as suspicious?

Options: Strongly Disagree; Disagree; Neutral; Agree; Strongly Agree

- Q13. To what extent do you agree that combining supervised classification, explainability, and anomaly detection could improve malicious insider threat detection?

Options: Strongly Disagree; Disagree; Neutral; Agree; Strongly Agree

- Q14. To what extent do you agree that the proposed XAI-based framework could support security analysts in making better-informed insider threat decisions?

Options: Strongly Disagree; Disagree; Neutral; Agree; Strongly Agree

- Q15. Overall, how effective do you perceive the proposed XAI-based malicious insider threat detection framework to be?

Options: Very Ineffective; Ineffective; Neutral; Effective; Very Effective

## X. RESEARCH OBJECTIVES

- To propose a conceptual Explainable Artificial Intelligence-based framework for malicious insider threat detection.
- To identify how behavioural data such as logon, file, web, and email activities can support malicious insider threat detection.
- To examine the role of XGBoost and SHAP in improving the explainability and interpretability of insider threat detection.

- To examine the perceived usefulness of Explainable Artificial Intelligence for improving trust and security decision-making.
- To assess respondents' perceptions regarding the potential effectiveness of the proposed XAI-based malicious insider threat detection framework.

## XI. VARIABLES OF THE STUDY

### A. Independent Variable

Use/Availability of Explainable Artificial Intelligence (XAI) Explanations. The independent variable represents the presence and usefulness of explanations provided by XAI techniques such as SHAP and, optionally, LIME.

### B. Dependent Variables

Perceived Trust: the extent to which respondents believe that an AI-based insider threat detection system can be trusted when it provides understandable explanations. Perceived Interpretability: the extent to which respondents believe that XAI explanations make the reasons behind a model's prediction understandable. Perceived Usefulness: the extent to which respondents believe that XAI explanations and behavioural information would be useful for insider threat detection and security decision-making. Perceived Effectiveness: respondents' perception of the potential ability of the proposed framework to improve malicious insider threat detection.

## XII. HYPOTHESIS

- Null Hypothesis (H0): XAI-based explanations do not significantly improve perceived trust and interpretability in AI-based malicious insider threat detection.
- Alternative Hypothesis (H1): XAI-based explanations significantly improve perceived trust and interpretability in AI-based malicious insider threat detection.
- Note: The proposed hypotheses (H0 and H1) are formulated based on the theoretical foundation of XAI and insider threat detection. These hypotheses have been empirically tested and statistically evaluated using the survey data collected for this study.

## XIII. RESULTS AND DATA ANALYSIS

### A. Survey Response Profile.

A total of 51 valid responses were obtained through the structured online survey. The responses were analysed using frequency and percentage distributions, descriptive statistics, and a one-sample t-test for the hypothesis-related perception measures. The analysis represents respondents' perceptions and does not represent measured machine-learning detection performance.

#### Respondent role distribution

| Role                            | Frequency | Percentage |
|---------------------------------|-----------|------------|
| Student                         | 21        | 41.2%      |
| IT Administrator                | 9         | 17.6%      |
| Security Analyst / Practitioner | 7         | 13.7%      |
| Academic / Researcher           | 6         | 11.8%      |
| Software Engineer               | 2         | 3.9%       |
| Software Developer              | 2         | 3.9%       |

|                             |   |      |
|-----------------------------|---|------|
| Auditor                     | 1 | 2.0% |
| Executive                   | 1 | 2.0% |
| Technical Support Executive | 1 | 2.0% |
| Software Tester             | 1 | 2.0% |

### B. Descriptive Analysis.

Mean scores were calculated on the 1–5 response scale. Higher scores indicate higher knowledge, familiarity, importance, usefulness, agreement, or perceived effectiveness according to the wording of each item. The highest mean was observed for the importance of explainability ( $M = 4.16$ ), followed by the importance of analysing logon, file, web and email activity ( $M = 4.06$ ). Overall perceived framework effectiveness had a mean of

| Survey measure                                       | Mean | SD   | Median |
|------------------------------------------------------|------|------|--------|
| AI knowledge/experience                              | 3.63 | 1.18 | 4      |
| Cybersecurity knowledge/experience                   | 3.35 | 1.07 | 3      |
| XAI familiarity                                      | 3.14 | 1.36 | 3      |
| Importance of explainability                         | 4.16 | 0.88 | 4      |
| Effectiveness of traditional security methods        | 3.33 | 1.26 | 3      |
| AI/ML can improve insider threat detection           | 3.75 | 0.96 | 4      |
| Usefulness of AI explanations                        | 3.75 | 1.04 | 4      |
| Importance of logon/file/web/email activity analysis | 4.06 | 0.90 | 4      |
| SHAP helps identify behavioural features             | 3.35 | 1.05 | 3      |
| Anomaly detection identifies unusual activities      | 3.45 | 1.06 | 3      |
| XGBoost classifies behavioural patterns              | 3.59 | 1.13 | 4      |
| Explanations make system more trustworthy            | 3.51 | 1.21 | 4      |
| XAI improves understanding of suspicious activity    | 3.47 | 1.24 | 4      |
| Combining classification, XAI and anomaly detection  | 3.92 | 0.93 | 4      |
| Framework supports analyst decisions                 | 3.61 | 0.92 | 4      |
| Overall perceived framework effectiveness            | 3.69 | 1.03 | 4      |

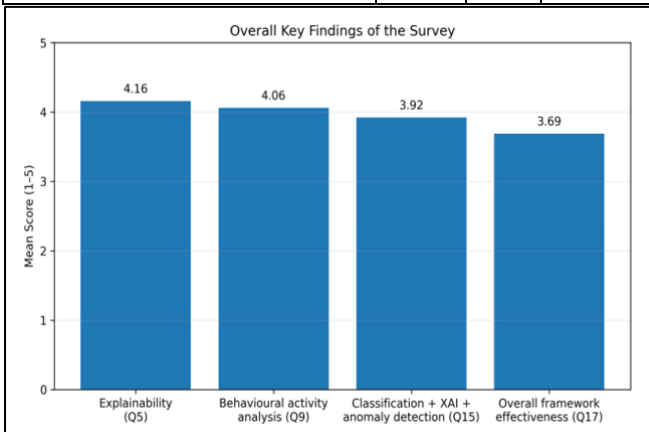


Fig. 1: Overall key findings of the survey showing mean scores for explainability, behavioural activity analysis,

combined classification + XAI + anomaly detection, and overall framework effectiveness.

### C. Key Response Patterns

- Importance of explainability: 72.5% selected the two highest response categories (4–5), with a mean of 4.16.
- Logon, file, web and email activity analysis: 74.5% selected categories 4–5, with a mean of 4.06.
- Usefulness of AI explanations: 64.7% selected categories 4–5, with a mean of 3.75.
- AI/ML improvement: 56.9% selected categories 4–5, with a mean of 3.75.
- Combined classification, XAI and anomaly detection: 64.7% selected categories 4–5, with a mean of 3.92.
- Overall framework effectiveness: 56.9% selected categories 4–5, with a mean of 3.69.

### D. Hypothesis Testing.

The hypothesis concerns perceived trust and interpretability associated with XAI-based explanations. The neutral midpoint (3) was used as the reference value for a one-sample t-test. Q13 measured whether explanations make the system more trustworthy, while Q14 measured whether XAI improves understanding of suspicious activity.

| Measure                        | Mean | t     | df | p-value |
|--------------------------------|------|-------|----|---------|
| Perceived trust (Q13)          | 3.51 | 3.018 | 50 | 0.0040  |
| Perceived understanding (Q14)  | 3.47 | 2.713 | 50 | 0.0091  |
| Combined trust + understanding | 3.49 | 3.019 | 50 | 0.0040  |

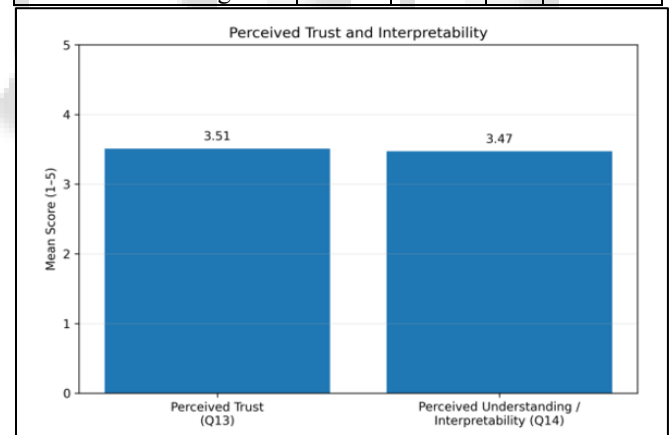


Fig. 2: Perceived trust and interpretability mean scores (Q13 and Q14). Both measures were significantly above the neutral midpoint of 3.

Interpretation. Perceived trust was above the neutral midpoint ( $M = 3.51$ ,  $t(50) = 3.018$ ,  $p = 0.0040$ ). Perceived understanding was also above the midpoint ( $M = 3.47$ ,  $t(50) = 2.713$ ,  $p = 0.0091$ ). The combined measure had a mean of 3.49, with a 95% CI of 3.16–3.82,  $t(50) = 3.019$ ,  $p = 0.0040$ . At the 5% significance level, the null hypothesis is rejected for the survey-based measures of perceived trust and interpretability.

### E. Discussion of Findings.

The responses indicate generally positive perceptions of explainability and the proposed framework. Respondents gave relatively high scores to the importance of explainability and behavioural activity analysis. The responses also show

moderate-to-positive perceptions of SHAP, anomaly detection, XGBoost, and the combined framework. The hypothesis test provides statistical evidence that the average perception of trust and understanding was above the neutral response level. These findings should be interpreted as survey-based perceptions rather than evidence that the proposed XGBoost, SHAP, or Deep Autoencoder components have achieved a particular detection accuracy, because the technical framework has not been experimentally implemented or validated in the present study.

#### F. Summary of Results.

Overall, the analysis of 51 responses supports the study's perception-oriented research objectives. Respondents generally viewed explainability, behavioural activity analysis, and the combination of classification, XAI and anomaly detection positively. The statistical test of the trust and understanding measures was significant at the 5% level. Therefore, the survey results provide empirical support for the proposed XAI-based framework at the level of respondent perception, while technical performance evaluation remains future work.

### XIV. CONCLUSION

Malicious insider threats present a difficult cybersecurity problem because authorized users can perform activities that appear similar to legitimate organizational behaviour. The analysis of behavioural information from login, file, web, and email activities can provide useful information for identifying potentially suspicious patterns. Machine Learning can support automated classification of behavioural patterns, but the black-box nature of many ML models can make their predictions difficult for security analysts to understand. Explainable Artificial Intelligence provides an opportunity to address this limitation by making model decisions more interpretable.

This paper proposes a conceptual framework that combines XGBoost for supervised threat classification, SHAP for explanation, and a Deep Autoencoder for separate behavioural anomaly detection. LIME is retained as an optional local explanation technique. The framework is designed to provide not only an indication of potentially malicious or anomalous behaviour but also understandable information about the behavioural factors associated with detection.

A structured survey of 51 respondents was conducted to examine perceptions regarding XAI, explainability, trust, interpretability, behavioural features, usefulness, and potential effectiveness. The results indicate generally positive perceptions, with the importance of explainability receiving the highest mean score ( $M = 4.16$ ). Hypothesis testing confirmed that perceived trust and understanding were significantly above the neutral midpoint. The primary contribution of this paper remains its conceptual architecture, supported by empirical perception-based evidence, while technical model training and dataset-based evaluation remain future work.

### XV. LIMITATIONS OF THE STUDY

- Conceptual nature of the framework: The proposed XGBoost, SHAP, and Deep Autoencoder components have not been technically implemented or experimentally validated in the present study.
- Survey-based perception: The study uses respondents' perceptions to examine usefulness, trust, interpretability, and perceived effectiveness rather than measuring actual detection performance.
- Variation in technical knowledge: Respondents may have different levels of knowledge about AI, cybersecurity, XAI, and machine learning, which may influence their responses.
- Self-reported responses: Survey findings depend on the accuracy and understanding of respondents' answers.
- Absence of organizational deployment: The proposed framework has not been tested in a real organizational environment.
- No current dataset-based evaluation: The study does not currently use an insider-threat dataset to train or evaluate the proposed models.

### XVI. FUTURE WORK

#### A. Framework Implementation

The complete architecture can be implemented using appropriate machine-learning and XAI tools.

#### B. Dataset-Based Evaluation

An appropriate insider-threat dataset, such as the CERT Insider Threat Dataset, may be considered for future technical evaluation, subject to dataset availability, suitability, and research requirements.

#### C. XGBoost Training and Testing

XGBoost can be trained using appropriate behavioural features and evaluated for malicious insider threat classification.

#### D. Deep Autoencoder Implementation

A Deep Autoencoder can be implemented separately for behavioural anomaly detection to examine its ability to identify unusual or previously unseen activities.

#### E. SHAP and LIME Comparison

Future work can compare SHAP and LIME in terms of the usefulness and interpretability of explanations generated for insider threat predictions.

#### F. Performance Evaluation

The implemented framework can be evaluated using suitable classification metrics such as Accuracy, Precision, Recall, F1-score, and ROC-AUC. These metrics would represent future experimental evaluation and are not achieved results of the present study.

#### G. Organizational Case Studies

Future research can investigate the framework in realistic organizational environments to determine whether the proposed combination of behavioural analysis, classification,

anomaly detection, and explainability can support practical security operations.

#### REFERENCES

- [1] Abbasi, A. M., Muqtadir, F., Basir, O., & Cao, S. (2025). Explainable bidirectional LSTM autoencoder for insider threat detection in various scenarios. 2025 IEEE International Conference on Cyber Security and Computer Science (ICCST). <https://doi.org/10.1109/icst63435.2025.11295479>
- [2] Alketbi, K., & Mehmood, A. (2025). A comprehensive survey of explainable artificial intelligence techniques for malicious insider threat detection. IEEE Access, 13. <https://doi.org/10.1109/access.2025.3587114>
- [3] Chattopadhyay, P., Wang, L., & Tan, Y.-P. (2018). Scenario-based insider threat detection from cyber activities. IEEE Transactions on Computational Social Systems, 5(3). <https://doi.org/10.1109/tcss.2018.2857473>
- [4] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794. <https://doi.org/10.1145/2939672.2939785>
- [5] Gheyas, I. A., & Abdallah, A. E. (2016). Detection and prediction of insider threats to cyber security: A systematic literature review and meta-analysis. Big Data Analytics, 1, 6. <https://doi.org/10.1186/s41044-016-0006-0>
- [6] Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. Science, 313(5786), 504-507. <https://doi.org/10.1126/science.1127647>
- [7] Homoliak, I., Toffalini, F., Guarnizo, J., Elovici, Y., & Ochoa, M. (2019). Insight into insiders and IT: A survey of insider threat taxonomies, analysis, modeling, and countermeasures. ACM Computing Surveys, 52(2). <https://doi.org/10.1145/3303771>
- [8] Jiang, W., Tian, Y., Liu, W., & Liu, W. (2018). An insider threat detection method based on user behavior analysis. IFIP International Conference on Intelligent Information Processing. [https://doi.org/10.1007/978-3-030-00828-4\\_43](https://doi.org/10.1007/978-3-030-00828-4_43)
- [9] Liu, L., De Vel, O., Chen, C., Zhang, J., & Xiang, Y. (2018a). Anomaly-based insider threat detection using deep autoencoders. 2018 IEEE International Conference on Data Mining Workshops (ICDMW). <https://doi.org/10.1109/icdmw.2018.00014>
- [10] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. arXiv:1705.07874. <https://doi.org/10.48550/arXiv.1705.07874>
- [11] Nasir, R., Afzal, M., Latif, R., & Iqbal, W. (2021). Behavioral based insider threat detection using deep learning. IEEE Access, 9. <https://doi.org/10.1109/access.2021.3118297>
- [12] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144. <https://doi.org/10.1145/2939672.2939778>
- [13] Wang, M., Zheng, K., Yang, Y., & Wang, X. (2020). An explainable machine learning framework for intrusion detection systems. IEEE Access, 8. <https://doi.org/10.1109/access.2020.2988359>
- [14] Zhang, Z., Al Hamadi, H., Damiani, E., Yeun, C. Y., & Taher, F. (2022). Explainable artificial intelligence applications in cyber security: State-of-the-art in research. IEEE Access, 10. <https://doi.org/10.1109/access.2022.3204051>