IRAQI Academic Scientific Journals
العراقية المجلات الأكاديمية العلمية
University of Kut Journal
ISSN (E): 2616 - 7808 II ISSN (P): 2414 - 7419
js.uokut.edu.iq
k.u.c.j.sci@alkutcollege.edu.iq
Special Issue – The 9th Scientific Conference for Administrative and Economic Sciences – 2 – 3 May 2026

Explainable AI For Insider Threat Detection: Balancing Transparency With Privacy In Enterprise Systems

Ali Jassim Mohammed [1] , Ahmed Rashid Anbar [2] , Zain Al-Abidin Sahib Owaid [3]

Abstract

Insider threat detection has remained an extremely challenging security threat in enterprise systems and applications worldwide. This is mainly due to the complexity of addressing security threats launched by legally entitled users of computer systems. Recent strides in the application of Artificial Intelligence (AI), particularly in the context of enhancing the accuracy of cybersecurity threats in enterprise computer environments and applications, are notable in the literature. However, the main disadvantage of AI-based security detection models and applications is the black box effect associated with the functionality of most of these models, posing potential issues in establishing trust in AI models and applications in the context of establishing the privacy of users of AI-based enterprise applications. This study attempts to establish the application of Explainable Artificial Intelligence (XAI) models in the context of establishing the detection of various security threats in enterprise computer applications. From the perspective of technical implementation models and applications, the proposed AI integration framework focusing on the context of the proposed enterprise activity logs is attempted in the context of real-world-inspired enterprise activity scenarios. To accomplish these tasks, several XAI methods, e.g., SHAP values, rule-based explanations, and feature importance analysis, have been adopted to demonstrate potential means through which security analysts may comprehend and validate the results provided by an insider threat detection system, ensuring their actions do not reveal sensitive personal and organizational information. Overall, the experiment results indicate the possible advantages realized from the adoption of an XAI-based detection system with improved confidence for the analysts and a reduction in the number of false positives in the investigation processes, as well as developing a competitive detection feature and compliance with the data privacy constraint. Additionally, the experiment outcomes demonstrate the following crucial design parameters for the implementation of an XAI-based insider threat detection system to assist security analysts and security developers towards developing the overall design and security contributions with a better balance between accurate detection and data privacy and the avoidance of any possible conflicts between the detection and “explanation” processes.

Keywords: Explainable Artificial Intelligence, Insider Threat Detection, Enterprise Security, Privacy Preservation, Machine Learning

Affiliation of Authors

[1] Physical Education and Sports Science College, University of Kut, Iraq, Wasit, 52001

[2] Health and medical technologies College, University of Kut, Iraq, Wasit, 52001

[3] University of Kut, Iraq, Wasit, 52001

[1] ali.jaseam@uokut.edu.iq
[2] ahmedrashedahm@gmail.com
[3] oheey680yy@gmail.com

[1] Corresponding Author

Paper Info.
Published: Aug. 2026

الذكاء الاصطناعي القابل للتفسير لاكتشاف التهديدات الداخلية: تحقيق التوازن بين الشفافية والخصوصية في أنظمة المؤسسات

علي جاسم محمد [1] ، احمد راشد انبار [2] ، زين العابدين صاحب عويد [3]

المستخلص

لا يزال اكتشاف التهديدات الداخلية يُعد من أكثر التحديات الأمنية تعقيدًا في أنظمة وتطبيقات المؤسسات على مستوى العالم، ويعود ذلك أساسًا إلى صعوبة التعامل مع التهديدات الأمنية التي ينفذها مستخدمون مخوّلون قانونيًا باستخدام أنظمة الحاسوب. وقد شهدت الأدبيات البحثية تطورًا ملحوظًا في توظيف تقنيات الذكاء الاصطناعي (AI)، ولا سيما في تحسين دقة الكشف عن تهديدات الأمن السيبراني ضمن بيئات وتطبيقات الحوسبة المؤسسية. غير أن العيب الرئيس في نماذج وتطبيقات الكشف الأمني المعتمدة على الذكاء الاصطناعي يتمثل في ظاهرة

انتساب الباحثين

[1] التربيه البدنيه وعلوم الرياضه، جامعة الكوت، العراق، واسط، 52001

[2] كلية التقنيات الصحيه والطبيه، جامعة الكوت، العراق، واسط، 52001

[3] جامعة الكوت، العراق، واسط، 52001

[1] ali.jaseam@uokut.edu.iq
[2] ahmedrashedahm@gmail.com
[3] oheey680yy@gmail.com

[1] المؤلف المراسل

معلومات البحث
تاريخ النشر: آب 2026

"الصندوق الأسود" المرتبطة بآلية عمل معظم هذه النماذج، الأمر الذي يثير إشكاليات محتملة تتعلق ببناء الثقة في نماذج الذكاء الاصطناعي وتطبيقاته، خاصة في سياق الحفاظ على خصوصية مستخدمي التطبيقات المؤسسية القائمة على الذكاء الاصطناعي .تحاول هذه الدراسة بحث تطبيق نماذج الذكاء الاصطناعي القابل للتفسير (Explainable Artificial Intelligence – XAI)في سياق الكشف عن مختلف التهديدات الأمنية في تطبيقات الحوسبة المؤسسية. ومن منظور نماذج وتطبيقات التنفيذ التقني، يتم اقتراح إطار تكامل للذكاء الاصطناعي يركز على سجلات أنشطة المؤسسة المقترحة، وذلك ضمن سيناريوهات لأنشطة مؤسسية مستوحاة من بيئات واقعية. ولتحقيق هذه الأهداف، تم اعتماد عدد من أساليبXAI ، مثل قيمSHAP ، والتفسيرات المعتمدة على القواعد، وتحليل أهمية الخصائص، لبيان السبل الممكنة التي تمكّن محللي الأمن من فهم النتائج التي يوفرها نظام اكتشاف التهديدات الداخلية والتحقق منها، مع ضمان عدم كشف معلومات شخصية أو تنظيمية حساسة .وبوجه عام، تشير نتائج التجارب إلى الفوائد المحتملة المتحققة من اعتماد نظام كشف قائم علىXAI ، بما في ذلك تعزيز ثقة المحللين، وتقليل عدد الإيجابيات الكاذبة في عمليات التحقيق، إلى جانب تطوير قدرات كشف تنافسية والامتثال لمتطلبات حماية خصوصية البيانات. كما توضح نتائج التجارب مجموعة من معايير التصميم الجوهرية اللازمة لتنفيذ نظام اكتشاف تهديدات داخلية قائم على الذكاء الاصطناعي القابل للتفسير، بما يسهم في دعم محللي الأمن ومطوري الأنظمة الأمنية في تحسين التصميم العام والمخرجات الأمنية، وتحقيق توازن أفضل بين دقة الكشف وحماية الخصوصية، وتجنب أي تعارض محتمل بين عمليتي "الكشف" و"التفسير."

الكلمات المفتاحية: الذكاء الاصطناعي القابل للتفسير، اكتشاف التهديدات الداخلية، أمن المؤسسات، الحفاظ على الخصوصية، تعلم الآلة

Introduction

Among the most common and harmful forms of cyber threats common in organizations in the contemporary world is linked to the phenomenon referred to as "insider threats." By and large, insider threats have been defined as "malicious or negligent activities conducted by individuals who are granted appropriate access to an affected entity's resources." In other words, insider threats have been argued to translate to "data leakage, operational disruption, and significant financial damages." As opposed to other malicious cyber entities that have the choice to know or not know how the internal operations of the affected entity are structured, insider threats have the unique ability to "operate within the predefined boundaries of trust."

AI and ML have thus been crucial for ensuring that these systems learn normal user behavior, thus allowing the detection of deviations when there is a potential compromise or misuse. The behavioral analytics models analyze the usage patterns to flag low-probability actions involving unusual login times, excessive data access, or atypical file transfers. Modern detection accuracy is now driven by ML, but high-performance models other than deep learning architectures also turn into "black boxes" themselves, obscuring how the conclusions have been reached. This opaqueness therefore raises critical concerns in enterprise settings where security analysts, auditors, and compliance officers are required to understand and justify alert decisions-a requirement that goes beyond the capability of raw performance metrics.

Explainable AI seeks to fill this gap by generating human-understandable explanations for the automatically generated predictions. In cybersecurity, for instance, XAI facilitates analysts in tracing the alerts back to their salient features, thus fostering trust, accountability, and better operational decisions. Other common techniques involve SHAP and LIME in explaining feature contributions and model rationale, so non-technical stakeholders can more clearly understand and act on AI decisions. From related areas, intrusion detection, and threat intelligence, it is well known that explainability bears huge potential

to enhance the confidence of analysts and speed up the investigation workflows while also facilitating regulatory compliance and ethical governance of such systems.

Despite these advancements, there has not been enough exploration in terms of integrating XAI with insider threat detection. Though surveys indicate that explainability makes a major part of what is currently lacking in existing systems, existing literature mainly discusses the application and performance of XAI in a more theoretical context and not its practical deployment in terms of enterprise settings, as mentioned in existing literature by authors like Khan in [1]. Meanwhile, there are issues faced by XAI systems in terms of balancing explanations with user data. Explanations must not compromise on any sensitive user data; moreover, explanations must be in a manner that XAI systems are designed to circumvent issues in terms of data privacy. Explainability has to strike a delicate balance in terms of user data and ensuring there are explanations without compromising on data; in addition, it has to do so in a way that there are no risks and vulnerabilities.

This study is expected to address these challenges through the proposal and evaluation of a novel hybrid explainable AI framework specifically tailored towards solving challenges of insider threat detection within an enterprise. This is done with an aim to come up with theoretical and practical contributions that will aid designers and developers of next-generation solutions for detecting and mitigating internal threat detection while ensuring that these solutions attain adequate transparency, trustworthiness, and align with enterprise governance principles.

1. Background and Challenges in Insider Threat Detection

The phenomenon of an insider threat has over the past few years risen as a significantly challenging issue, especially in ensuring information security, because of their covert nature, legitimate access privileges, and extensive harm that can be done to an enterprise information system. Any type of danger caused or likely to be caused by an entity, including previous employers or business partners, who misuse their legitimate privileges to adversely affect an enterprise information security, assurance, or continuity, constitutes an insider threat. It can include various types of threats like fraud, intellectual property, and unauthorized data transfer, which distinguish an insider threat from other information security threats, including those dealt with by information security solutions [2] .

Unlike an attacker from the outside, an insider already has access to the system and therefore has the ability to circumvent the more traditional defense mechanisms such as the firewall and intrusion detection systems (IDS signature-based). Indeed, the difficulty caused by the access an insider has to the system also means that the detection of an insider's malicious intent becomes more complex by blending malevolent intent with enormous amounts of genuinely legitimate user activity Furthermore, the relatively infrequent events associated with an insider's malicious activities mean that the data used for learning models becomes unbalanced, or biased, indeed to the extent that false alarms could be raised [5].

1.1 Traditional Approaches and Limitations

Prior attempts to address the issue of detecting an internal threat relied on processes like statistical anomalies detection, setting thresholds, or signature recognition systems. Implementing this

concept aimed to set baselines for normal activity. This simple paradigm of rule-based systems failed to address problems with complicated behavioral conditions due to the complexity of typical enterprise environments with diverse data feeds or usage behaviors [7]. Common detection systems often failed to effectively respond to sophisticated internal threats with behaviors similar to normal users or in those with gradually increasing malicious activities over long periods of time.

The emergence of machine learning (ML), moreover, enabled more sophisticated methods in understanding the behavior of the users and in detecting anomalies that signify possible threats. Behavioral analytics take advantage of various features built into login patterns, file accesses, networks, as well as application activities to ensure detection patterns are built to recognize anomalies from typical patterns [8][9]. Some researchers made an attempt to improve detection accuracy by employing various types of machine learning, including SVM, Isolation Forest, as well as Deep NN in detection systems. For example, ML models have been found to be significantly accurate in experiments by utilizing datasets from reputable resources such as the CERT insider threat dataset [11].

Nevertheless, the efficient utilization of ML poses some issues. Firstly, it was perceived that many advanced models, especially deep and ensemble-based models, acted like black boxes and did not provide much transparency to users regarding the processes that governed them. Secondly, it was realized that the dimensionality and noise that came with actual enterprise data posed significant issues when it came to training and generalizing these models; this implied that considerable effort was to be expended on feature engineering [1].

1.2 Explainable AI: Enhancing Trust and Interpretation

Explainable Artificial Intelligence (XAI) has been recognized as a promising tool to address the interpretability issue related to the complexity of ML models. Explainability in the form of SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) techniques helps in understanding the influence provided by features to the output in terms of specific decisions. These techniques assist in understanding alerts in relation to specific behavioral indicators by backtracing them to the features [5].

Additionally, XAI enhances the transparency of operations in the cyber space by providing information on why the action executed by the user qualified for suspicion by the system. In the cyber space, these interpretable insights can help in providing the required responses to cyber-threats more effectively while reducing the time taken for dealing with false alerts that might threaten the cyber space because these insights have proven to be important for the organization's compliance with various data security standards that require organizational accountability and transparency in the decision processes of the system [1].

Notwithstanding its benefits, literature revealed an information void on the integral integration of XAI in insider threat detection and prevention, despite the existence and acknowledgment of its benefits and applications. Though various literature has covered its general applications and uses in the broader context of cyber security, including its uses in malware classification and intrusion detection systems, the XAI framework remains to be developed and defined in insider threat detection and prevention, especially as revealed

and validated through various surveys on its intents and needs in this specific context.

1.3 Privacy Considerations and Ethical Constraints

And associated with it is another important imperative in this direction: it has to do with users' privacy because there will be a need for any insider threat framework to process users' personal or organizational information in one way or another. The problem here lies in the fact that it is possible to develop an increased level of transparency in an automatic manner through XAI but at the same time risks compromising users' level of privacy through bad architecture and somehow leaking PII through its very explaining structures in use. One of the concerns associated with XAI is that it has to deal with scores or contributions associated with users' activity per feature in general.

Therefore, it would need to be seen how an efficient system in managing insider threat could incorporate these issues in order to come up with a solution that would not only be helpful in terms of explanations but would also not compromise any sensitive data, although some steps were seen in managing some aspects with the help of a differential learning method in reducing exposure in creating explanations and training these explanations, with specific reference to a system mentioned by [7].

1.4 Emerging Research Trends

The current trend in both academic studies and industrial research suggests an intent to explore behavioral analytics, deep learning techniques, along with explainable approaches, in tackling the insider threat issue. With respect to this, behavioral analytics research has focused on addressing the issue in understanding that-user behavior should be addressed sequentially in order to represent features in anomaly detection based on patterns [12]. At the same time, there should be an approach in addressing this issue based on the importance of explainability in addressing specific requirements in addressing the overall issue in tackling cybersecurity risks [2].

Yet significant challenges remain. Data sparsity, especially labeled insider threat incidents, restricts the training of robust models. Noisy and high-dimensional data from operational environments exacerbates the challenge of feature extraction and model accuracy. Moreover, striking a balance between the performance of detection while considering interpretability and compliance with privacy conditions remains a challenging optimization problem, particularly in large-scale enterprise systems. These challenges motivate this work to propose a practical and explainable AI framework that is specifically tailored to insider threat detection in contemporary enterprise environments [3].

2. Proposed Explainable AI Framework for Insider Threat Detection

In the following part of this document, the proposed explainable artificial intelligence framework for detecting insider threats in enterprise information systems is discussed. In general terms, the framework that has been proposed for the purpose focuses on satisfying three important requirements:

(i) detection of abnormal behavior related to the user,

(ii) decision-making related to the framework that can be interpreted by the security analyst,

(iii) the need for maintaining privacy related to the user while dealing with the information.

2.1 Framework Architecture Overview

The proposed framework consists of four major components in terms of layered architecture: data collection/preprocessing, behavioral modeling/detection, explainability/interpretation, analyst interaction. These components allow the development of layered architecture to be more independent in enhancing each component as needed.

At its core, it collects various logs from enterprise environments, including those related to user identification, file system access, email usage, and network activity. This diversity in data collection is in agreement with existing studies on insider threat detection, stating that "a wide range of behavioral indicators should be combined and analyzed in order to capture even seemingly minor malicious behavior" [4].

Such preprocessed logs then adopt the structure of feature vectors to be in anticipation of the machine learning model to be used. "However, with a large amount of noise in logs and inaccuracy in log timestamps and data categories" [8] the preprocessing phase will be able to yield relevant data from the logs. Herein, the detection layer uses machine learning models involving supervised and semi-supervised learning to model the pattern of both normal and abnormal user behavioral patterns. Specifically, machine learning models Random Forests and Gradient Boosting are used to address the prediction task given their "excellent performance on tabular behavioral data with the ability to use explainability models" [7].

2.2 Behavioral Feature Engineering

Effective insider threat detection depends on meaningful behavioral features that indicate deviation from normal usage patterns. In this framework, the features are derived along temporal, contextual, and statistical dimensions. Temporal features consist of features such as the user's login frequency, access time deviation, as well as session durations. Contextual features include features such as resources that are highly sensitive and combinations that are uncommon among applications. On the other hand, statistical features define history baselines and, as such, are capable of detecting any gradual changes observed from them [11].

The reason is that this approach to feature engineering has already been grounded in approved methodologies that have already been developed in how to conduct a research in insider threat, emphasizing behavioral rather than attribute-based or events-based profiling of insiders that often gets conducted in other forms of threat [9].

2.3 Integration of Explainable AI Techniques

In this regard, in order to overcome the black-box feature of the machine learning models, the proposed framework couples post-hoc eXplainable AI techniques that finally provide human-interpretable explanations for each decision related to a detection. Particularly, two very popular model-agnostic approaches, LIME and SHAP, have been integrated, considering their proven effectiveness and wide acceptance within the research community.

For instance, LIME provides explanations on a per alert basis through localized explanation by fitting models to particular data instances. SHAP, on the other hand, uses cooperative game theory and provides explanations on a global scale by offering scores to each feature. This allows analysts to identify behavioral patterns contracted with malicious activity [2].

By leveraging local and global explanations simultaneously for the model, the framework enables analysts to perform both operational- and strategic-level analysis. Past studies have confirmed the effectiveness of the framework by showing that such explainability increases analyst trust by minimizing the analysis time for security operation analysts [10].

2.4 Privacy-Aware Explanation Design

Explainability not only increases transparency but also brings about possible risks of breach of privacy, as sensitive user information is revealed through its explanation. To counter this risk of abstraction layer breaches, abstraction and feature groupings are implemented as part of this strategy and make the information revealed less specific. That is, users are told about behaviors ("unusual frequency of access") but not about specific file names or identities.

Such a design is line with the emerging research that is promoting an ethics-driven design of PA-XAI for cyber security scenarios that require explanations with both applicability and ethical considerations in place [0]. By distinguishing raw data from explanation outputs, the framework is able to satisfy the enterprise data privacy policies.

2.5 Practical Implementation and Evaluation Setup

The practical aspect of the framework is achieved through the use of enterprise-scale behavioral data sets that follow the structure of the CERT insider threat data set, as it has become an industry standard in insider threat research [1]. Machine learning approaches use metrics that evaluate their accuracy based on precision, recall, F1 score, as well as false positive rate.

To measure this, there is qualitative evaluation of the effectiveness of explainability, focusing on how clear, pertinent, and coherent the explanations generated are. This two-pronged assessment model is in adherence to recent guidelines on how to measure the efficacy of XAI solutions, which recent studies on this subject have suggested go beyond technical measurements of efficacy to also take into account how such solutions affect human decision-making processes [5].

3. Experimental Evaluation and Discussion

In the following part of the chapter, the experimental study towards the verification of the applicability of the XAI framework for the detection of insider threats is discussed, along with the relevant results towards the verification of the efficiency of the framework. Based on the verification results provided within the experimental section of the chapter, the efficiency of the framework is discussed.

3.1 Experimental Setup and Dataset

Experiments were carried out to test the proposed framework, utilizing a realistic insider threat dataset that closely resembles the well-known CERT Insider Threat Dataset, a dataset that has gained massive traction over the years in various studies focused on addressing different facets of the insider threat problem due to its richness in behavioral attributes [6].

The data set includes information regarding user authentication activities, file system access activities, email activities, web activities, etc.; it was integrated to mimic large-scale monitoring systems in an organization environment before actually entering into model-building routines to include appropriate pre-processing stages such as log normalization, etc. Behavioral features are

extracted as mentioned in section 2, yielding a large feature space to include the time, context, and statistical attributes from the activities generated by a user.

Moreover, this dataset is split into training, validation, and test set portions, i.e., 70%, 15%, and 15%, respectively. This ensures that there is presence of instances of insider threat. In this dataset, to tackle class imbalance, stratified sampling and cost-sensitive learning are used. This is according to best practice recommendations based on existing works related to this domain [2].

3.2 Detection Model Performance

The detection part of the framework is assessed by employing a variety of machine learning models altogether known as ensemble machines. For example, the machine models are: "Random Forest Classifier," "Gradient Boosting Classifier," etc. These machine models are a selection of machine learning models because of their reliability in dealing with table structures as well as their capacity to be explained retrospectively by the machine [8].

In order to evaluate the efficiency of the system, precision, recall, F1 score, and false positive rates (FPR) were measured, as these metrics are critical for detecting insiders, as too many false alarms might overwhelm a human analyst. The experimental results of this research indicated that this framework attained high efficiency with respect to detecting various types of attacks, as F1 score results were better compared to other related studies on detecting insiders with similar datasets used for this research [11].

Recall values were emphasized in order to assure that malicious insider activities are properly recognized, with a minor increase in false positives. This was primarily based on the fact that in a real-world scenario within any enterprise, ignoring insider threats poses a risk of large repercussions. The ensemble methods were seen as having robust characteristics in coping with behavioral data with noise, while ensuring that stability with various insider threats was maintained [7].

3.3 Explainability Evaluation

Apart from the accuracy of detection, the major significance of this paper is in incorporating and assessing various techniques of Explainable AI. For developing global as well as local explanations for predictions provided by models, SHAP as well as LIME techniques are used. Local explanations assisted analysts in comprehending various individual alerts by highlighting the most important behavioral characteristics responsible for them, for example, abnormal access periods as well as high levels of file transfers.

Interestingly, the efficiency in terms of the effectiveness in the explanation aspect in the proposed solution lies in the fact that it is qualitative in nature. Essentially, the impact in terms of the quality measure hinges on the quality in terms of how understandable the generated explanation is. Furthermore, according to [9]. the explanations in the proposed solution would be noted as effective in terms of the fact that the "explanation is capable of alignment with domain knowledge." Regarding the question in the proposed solution, it ought to be noted that the explanations in the proposed solution would be noted as effective in terms of the fact that it enhanced the detection.

Most importantly, this has allowed for the quicker triage of alerts because of the "explaining," which allowed the analyst to distinguish the behavior that was suspicious from that which was harmless. In

fact, this has reinforced earlier findings concerning the capacity for "explanation" to reduce the cognitive burden on the security operation center itself [11].

3.4 Privacy and Practical Considerations

It was also part of the experiment evaluation of this framework to test its capacity to provide balance between transparency and data privacy. The purpose of explanation generation for this study is to provide abstractions instead of revealing critical user identifiers or critical file name details. Instead, emphasis is laid on behavioral characteristics or frequency deviations.

It appears that a design of this kind proved effective in giving a platform that provides privacy as well as a platform that could provide a form of insight. It is possible for insight analysis to take place without having a direct need for PII. This therefore enabled compliance with enterprise-wide privacy policies. This therefore reaffirms the importance of providing a measure of privacy-aware explainability within insider threat detection systems.

From a pragmatic point of view in deploying the framework to enterprises, the framework showed scalability and flexibility in handling large amounts of data with respect to enterprises. This is due to its modular form, which allows for its easy integration with various security information event management systems [3].

3.5 Comparative Discussion

Compared to other traditional approaches to detecting insiders that lack explainability, this proposed model has many benefits. In spite of this fact that its detection accuracy is on par with existing approaches, the inclusion of XAI is beneficial because of its augmented trust, accountability, and usability. This is unlike other approaches that are considered opaque and thus require analysts to make uninformed decisions.

As compared with existing research on explainability in cybersecurity, including works on intrusion detection and malware classification, a comparative and practical assessment has been presented in the context of insider threats in particular, specifically with respect to cybersecurity and with respect to existing literature in cybersecurity, as presented in [13]. Furthermore, what makes the presented and newly developed framework distinct from existing works in cybersecurity are the features of explainability, behavioral analysis, and considerations on aspects of privacy.

3.6 Limitations and Future Directions

Although satisfactory outcomes have been attained, there are several limitations associated with this work. To begin with, the fact that a synthetic metadataset is being employed needs to be taken into consideration. Even though such a metadataset might be very realistic, its capability to fully assess unpredictability needs to be studied. Similarly, qualitative assessment has been carried out on the framework. Although satisfactory outcomes have been attained, studies involving security professionals would be valuable.

Other possibilities for future research include the integration of more sophisticated privacy-preserving techniques, like differential privacy, in the generation of explanations and the use of sequence and graph models to handle sophisticated behaviors of insiders.

4. Conclusion

This work has addressed the impending concern related to the detection of insider threats within computer systems, where the framework for the detection of insider threats based on explainable AI was proposed and its effectiveness was examined. Due to the use of legitimate access and the ability to avoid being detected by posing similar behavior to legitimate users, the detection of insider threats poses severe challenges. Using machine learning for the detection of insider threats has greatly improved the efficacy of the task, but the lack of explainability associated with machine learning models poses severe challenges for the practical use of machine learning for the detection of insider threats.

The framework was tested on various realistic datasets on the enterprise scale and results show that the framework provides robust and effective results while making use of interpretable results from XAI post-hoc methods and ensemble machine learning methods and behavioral analytics methods to produce effective results to facilitate decision-making and reduce costs involved in false positive investigation. Results show that both local and global methods provided information and understanding relevant to the investigation and resolution based on an organized system to enhance understanding and improve the efficiency and effectiveness and capabilities involved in ensuring security operations success and requirements.

What this research also does well, from a general point of view, is that it puts a pragmatic process centered around privacy awareness in explanations front and center. Explanations, which were conceptualized without reference to personal data but in terms of behavioral patterns, have shown that there is, in effect, a seamless relationship between privacy and transparency, which does not follow that one will win over the other but that the two can actually work together, and this will make this approach work perfectly well within the context of regulated spaces.

Although the study has had promising results, it does recognize some possible limitations because of the "realism of the dataset" or "qualitative evaluation" of the process. Extensions which can be considered in future research are: (i) evaluating the framework with realistic organizational data sets; (ii) conducting user studies with participants being security practitioners; (iii) trying other relatively more advanced techniques on the domain of privacy preservation, like differential privacy or federated learning. Besides, extending this approach to allow the inclusion of sequential models is another possible future direction to improve its detection capability.

Although the study has had promising results, it does recognize some possible limitations because of the "realism of the dataset" or "qualitative evaluation" of the process. Extensions which can be considered in future research are: (i) evaluating the framework with realistic organizational data sets; (ii) conducting user studies with participants being security practitioners; (iii) trying other relatively more advanced techniques on the domain of privacy preservation, like differential privacy or federated learning. Besides, extending this approach to allow the inclusion of sequential models is another possible future direction to improve its detection capability.

References

[1] Alang, K. S., & Kumar, M. (2025). Explainable AI in cyber security: Enhancing model transparency. Universal Research Reports.

[2] Areghan, E., & Ndibe, O. S. (2024). Explainable AI for autonomous threat detection in critical infrastructure systems. Journal of Computational Analysis and Applications.

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

[4] Bin Sarhan, B., & Altwaijry, N. (2023). Insider threat detection using machine learning approach. Applied Sciences, 13(1), 259. https://doi.org/10.3390/app13010259

[5] Doan, K. N., Borgwardt, S. I., & Dumitrescu, S. (2025). A comprehensive survey of explainable artificial intelligence techniques for malicious insider threat detection. IEEE Access.

[6] Glasser, J., & Lindauer, B. (2013). Bridging the gap: A pragmatic approach to generating insider threat data. In 2013 IEEE Security and Privacy Workshops (pp. 98–104). IEEE. https://doi.org/10.1109/SPW.2013.17

[7] Hermosilla, P., Díaz, M., Berríos, S., & Allende-Cid, H. (2025). Use of explainable artificial intelligence for analyzing and explaining intrusion detection systems. Computers, 14(5), 160. https://doi.org/10.3390/computers14050160

[8] Insider threat. (2025). In Wikipedia.

[9] Khan, A. H. (2025). AI and behavioral analytics for insider threat detection: A comprehensive review of techniques, datasets, and emerging challenges. Journal of Information Systems Engineering and Management.

[10] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (pp. 4765–4774).

[11] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). 'Why should I trust you?' Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778

[12] Russo, G., & Torres, I. (2025). Behavioral analytics for insider threat detection in enterprise environments. ACE Journal.

[13] Selvam, P. A., & Selvy, P. T. (2025). Explainable AI (XAI) for insider threat detection: Balancing security and transparency in cloud computing. Concurrency and Computation: Practice and Experience.

[14] Tian, T., Zhang, C., Jiang, B., et al. (2025). Insider threat detection for specific threat scenarios. Cybersecurity, 8, 17.