Insider Threat Detection Using AI–Based Behaviour Analytics

Rushdha V[1], Shada Ali Kuzhikattil[2], Dr.Priya.P. Sajan[3]
[1,2]*Member, UG student Computer Science and Engineering (AI & Data Science), Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology Avadi, Chennai, Tamil Nadu, India*
[3]*Member, Senior Project Engineer, C-DAC Thiruvananthapuram*

***Abstract*—Insider threats pose a unique cyber security challenge, due to the authorized nature of access making malicious actions difficult to distinguish from normal behaviour. With the rising impact of insider incidents and the limitations of rule-based approaches, there is a critical need for adaptive and intelligent detection solutions. This research proposes an AI-driven behaviour analytics framework that analyses patterns in user activity to detect anomalies suggestive of insider threats. Using the CERT Insider Threat Dataset and a Random Forest classifier, the proposed system achieves accurate, interpretable, and computationally efficient threat detection. This Adaptive AI-driven framework emphasizes early threat identification, minimizing false positives, and ensures feasibility for real-world deployment in resource constrained environments.**

***Index Terms*—Insider Threat Detection, Behaviour Analytics, Machine Learning, Random Forest, Anomaly Detection.**

I. INTRODUCTION

As digital transformation advances across industries, insider threats have become one of the most persistent and damaging cyber security challenges. Unlike external attacks, insider threats originate from individuals within an organization such as employees, contractors, or third-party vendors who may misuse their authorized access intentionally or unintentionally.

According to the 2024 Verizon Data Breach Investigations Report (DBIR), insider actions were responsible for approximately 18% of all reported cyber security breaches, with “Privilege Misuse” and “Error” identified as the most common contributing factors [2]. In a related finding, the Ponemon Institute estimated the average cost of an insider-related security incident to be around USD 15.4 million [1]. These incidents affect critical sectors such as finance, healthcare, government, and national infrastructure, where even a minor breach can lead to significant financial, operational, or strategic harm.

The rise of remote and hybrid work models, along with increasing dependence on cloud platforms and digital collaboration tools, has expanded the landscape in which insider threats can occur. These shifts make such threats more pressing and complex.

Traditional security mechanisms, such as rule-based access control, Security Information and Event Management (SIEM) systems, and manual audits, often fall short in identifying the subtle and dynamic behavioural patterns of insider activity. These approaches typically rely on static thresholds, lack contextual awareness, and often generate a high volume of false positives, leading to alert fatigue among security analysts. Furthermore, insider actions are frequently concealed within routine activities, making it difficult to detect using traditional methods.

In response to these limitations, behaviour-based detection techniques powered by Artificial Intelligence (AI) and Machine Learning (ML) have gained attention. These methods offer adaptive and scalable solutions that can learn from historical user activity and flag anomalous behaviour in real time. This research explores such an AI-based approach that utilizes the CERT Insider Threat Dataset and a Random Forest classifier to improve detection accuracy, support interpretability, and ensure feasibility for deployment in practical, resource-constrained environments.

II. LITERATURE SURVEY

Insider threats have become a growing concern in

cyber security research, due to their stealthy nature and high financial impact. As organizations increasingly adopt digital-first operations and remote work environments, malicious insiders with authorized access have exploited their privileges to exfiltrate data or sabotage systems. According to the 2024 Ponemon Institute report, insider incidents cost organizations an average of USD 15.4 million annually, marking a 40% increase over the past five years [1]. The 2024 Verizon Data Breach Investigations Report (DBIR) reported that approximately 18% of all reported security breaches to insider actions, with ”Privilege Misuse” and ”Error” identified as the leading action varieties in such incidents [2, p. 18].

Sector-specific studies highlight this threat across industries such as healthcare, finance, and government, where sensitive data exposure can result in severe legal and operational disruptions [3], [4]. Conventional defense mechanisms, including Security Information and Event Management (SIEM) systems, static access controls, and manual audits, are often reactive and inadequate for identifying subtle changes in user behaviour that signal insider activity [5]. These methods suffer from static thresholds, limited contextual awareness, and high false positive rates, leading to operational inefficiencies and delayed incident response.

To address these challenges, recent studies have explored the use of Artificial Intelligence (AI) and Machine Learning (ML) for insider threat detection. Pal et al. [6] introduced a temporal feature aggregation model with attention mechanisms using the CERT v6.2 dataset and demonstrated that ensemble models such as Random Forests are effective in capturing behavioural patterns. Rashid et al. [7] demonstrated use of Hidden Markov Models to track sequential access patterns that may indicate malicious intent. Salem et al. [8], in a broader survey of insider threat detection strategies, highlighted deep learning-based anomaly detection techniques using neural networks.

However, many of these AI-driven methods lack transparency and high computational resources, making them impractical for smaller organizations. Moreover, several models fail to incorporate contextual behavioural factors such as abnormal access times or sudden spikes in USB or email activity that could enhance detection accuracy. Our research addresses these gaps by proposing a behavioural analytics framework based on a Random Forest classifier trained on the CERT Insider Threat Dataset. This approach aims to provide a practical balance between detection performance, explainability, and computational efficiency, supporting its deployment in environments with constrained security infrastructure.

III. EXISTING SYSTEM

Organizations have traditionally depended on standard securit measures such as rule-based mechanisms, manual audits, and Security Information and Event Management (SIEM) platforms to detect and manage insider threats. While these systems serve as the foundational layer of enterprise security, they often struggle to detect insider threats, mainly because malicious actions can closely resemble routine activity carried out by users with legitimate access.

Rule-based systems generally operate on predefined conditions and thresholds. While they are effective in flagging known attack patterns, they often fail to recognize evolving and context-specific behaviours often exhibited by insiders. Similarly, SIEM platforms and manual audit processes are primarily designed to collect and review security events after an incident has occurred, offering limited support for real-timedetection. These approaches often generate excessive false positives, contributing to alert fatigue among analysts and reducing overall response effectiveness.

To address these limitations, recent academic efforts have introduced Artificial Intelligence (AI) and Machine Learning (ML) techniques into insider threat detection. For instance, Pal et al. introduced a temporal feature aggregation model with attention mechanisms, which showed improved detection accuracy using the CERT v6.2 dataset [6]. Rashid et al. utilized Hidden Markov Models to analyse user activity sequences and uncover behavioural anomalies suggestive of insider intent [7]. Similarly, Salem et al. explored deep learning-based anomaly detection using neural networks to identify deviations from normal user behaviour [8].

Despite these advancements, many existing models still face challenges when deployed in practical settings. Many AI-driven models lack transparency in

their decision-making processes, which complicates the validation of alerts and reduces trust in operational settings. In addition, some approaches require high computational overhead, making them less suitable for deployment in smaller or resource-constrained organizations. Moreover, a number of these systems fail to account for contextual behavioural factors such as irregular access times or unusual spikes in USB and email activity which, if integrated, could enhance the accuracy and relevance of threat detection.

These limitations highlight the need for an efficient, explainable, and behaviour-aware solution. To address this, the present research proposes a behavioural analytics model leveraging a Random Forest classifier trained on the CERT Insider Threat Dataset. The proposed framework seeks to improve detection performance, reduce false positives, and offer an interpretable and resource-conscious solution suitable for diverse organizational contexts.

IV. PROPOSED SYSTEM

A. Overview

This research presents a behaviour-driven framework for insider threat detection, utilizing machine learning to identify deviations from established user activity patterns. Unlike conventional rule-based or signature-based systems, the proposed approach employs a Random Forest classifier to learn typical user behaviours from historical log data and flag anomalies indicative of potential threats. This method aims to reduce false positives and improve interpretability, thereby offering a practical solution for organizations with limited cybersecurity infrastructure.

To evaluate the effectiveness of the framework, the CERT Insider Threat Test Dataset v6.2 developed by the CERT Division at Carnegie Mellon University's Software Engineering Institute was used. This dataset is widely regarded as a benchmark in insider threat research due to its detailed simulation of real-world organizational environments [9].

B. System Architecture and Workflow

The system architecture is structured into four primary stages: data pre-processing, feature engineering, threat detection, and risk scoring & interpretation.

1) Data Pre-processing: Raw activity logs from the CERT v6.2 dataset are first cleaned and organized to extract relevant events, including user logins, file access operations, email interactions, and USB device usage. This stage ensures that unstructured log data is converted into a format suitable for behavioural analysis.

2) Feature Engineering: From the pre-processed logs, a set of behavioural indicators is derived. These include anomalies such as irregular login times, unexpected USB insertions, sudden increases in file access activity, and elevated outbound email usage. These features are aggregated on a per-user, per-day basis to form the input vectors for the classification model.

3) Threat Detection: A Random Forest classifier is trained to distinguish between normal and suspicious user behaviour. The algorithm is selected for its high classification accuracy, resilience to noise, and balance between performance and interpretability—attributes that make it suitable for deployment in enterprise environments. To improve model transparency, SHAP (SHapley Additive exPlanations) values are employed to explain the influence of each feature on the model's predictions. This allows analysts to understand why a particular user was flagged, supporting traceability and trust in the system's outputs [6].

4) Risk Scoring and Interpretation: Based on the model's output, each user is assigned a dynamic risk score that reflects the severity and frequency of behavioural anomalies. These scores aid in prioritizing alerts for further investigation and can be adapted to suit different operational risk thresholds. To ensure resource efficiency, model parameters such as tree depth and the number of estimators are optimized to support deployment in low-computation environments, making the framework accessible to organizations with limited IT resources.

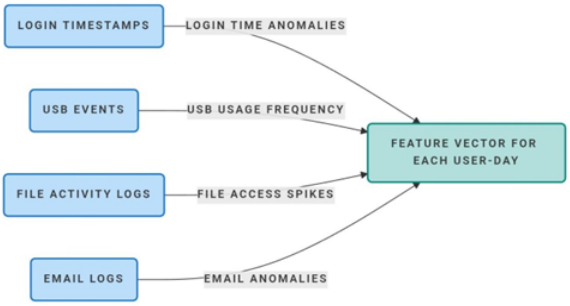

Fig. 1. Process of feature extraction from user activity log.

V. CONCLUSION

Insider threats continue to be one of the most difficult challenges in cyber security, not because of technical complexity alone, but because they involve trusted individuals operating within legitimate boundaries. Traditional systems, though widely used, often fail to catch subtle behavioural changes that signal potential insider misuse. They rely on predefined rules and thresholds that do not adapt well to real-world dynamics.

In response to this gap, our research introduced a behaviour-based detection framework powered by a Random Forest classifier trained on the CERT Insider Threat Dataset. By analysing real indicators such as unusual login times, sudden email surges, or unexpected file access patterns, the system brings a deeper understanding of user behaviour and enables more context-aware threat detection.

Our results show promising improvements in reducing false positives and maintaining high detection accuracy, even in environments where computing resources are limited. What makes this approach particularly relevant for today's organizations is its balance between effectiveness and transparency. With the integration of SHAP-based explainability and a dynamic risk scoring system, the model supports security teams not just in detecting threats, but in understanding them.

This work serves as a practical step toward building insider threat solutions that are not only intelligent but also interpretable and deployable in real-world scenarios. Looking ahead, future enhancements could involve real-time detection with streaming data, use of Natural Language Processing to understand behavioural intent from communications and the incorporation of multi-source data to handle the growing complexity of hybrid and remote work settings.

VI. FUTURE SCOPE

Insider threats continue to be one of the most difficult challenges in cybersecurity, not because of technical complexity alone, but because they involve trusted individuals operating within legitimate boundaries. Traditional systems, though widely used, often fail to catch subtle behavioural changes that signal potential insider misuse. They rely on predefined rules and thresholds that do not adapt well to real-world dynamics.

In response to this gap, our research introduced a behaviour-based detection framework powered by a Random Forest classifier trained on the CERT Insider Threat Dataset. By analyzing real indicators such as unusual login times, sudden email surges, or unexpected file access patterns, the system brings a deeper understanding of user behaviour and enables more context-aware threat detection.

Our results show promising improvements in reducing false positives and maintaining high detection accuracy, even in environments where computing resources are limited. What makes this approach particularly relevant for today's organizations is its balance between effectiveness and transparency. With the integration of SHAP-based explainability and a dynamic risk scoring system, the model supports security teams not just in detecting threats, but in understanding them.

This work serves as a practical step toward building insider threat solutions that are not only intelligent but also interpretable and deployable in real-world scenarios. Looking ahead, future enhancements could involve real-time detection with streaming data, use of Natural Language Processing to understand behavioural intent from communications, and the incorporation of multi-source data to handle the growing complexity of hybrid and remote work settings

REFERENCE

[1] Ponemon Institute, "2024 Cost of Insider Threats Global Report," Sponsored by Proofpoint, Mar. 2024. [Online]. Available: https://www.proofpoint.com/sites/default/ files/ponemon-2024-cost-insider-threats.pdf

[2] Verizon, "2023 Data Breach Investigations Report," Verizon, 2023. [Online]. Available: https://www.verizon.com/business/resources/reports/dbir/

[3] IBM Security," Cost of a Data Breach Report 2023," IBM, 2023. [Online]. Available: https://www.ibm.com/reports/ data-breach

[4] McKinsey & Company," Cybersecurity Trends: Looking Over the Horizon," McKinsey & Company, Oct. 2023. [Online]. Available:

https://www.mckinsey. com/business-functions/risk-and-resilience/our-insights/ cybersecurity-trends-looking-over-the-horizon

[5] M. Bishop," Position: Insider is relative," in Proc. New Security Paradigms Workshop, Lake Arrowhead, CA, USA, 2005, pp. 77–78. P. Pal, P. Chattopadhyay, and M. Swarnkar, "Temporal feature aggregation with attention for insider threat detection from activity logs," Expert Systems with Applications, vol. 224, Art. no. 119925, Mar. 2023, doi: 10.1016/j.eswa.2023.11992

[6] T.Rashid, I. Agrafiotis, J. R. Nurse, MIST 16 (ACM Workshop)," Insider threat detection using behavioral modeling and hidden Markov models," *Computers & Security*, vol. 102, pp. 1–14, 2020.

[7] M. Salem, S. Hershkop, and S. J. Stolfo," A survey of insider attack detection research," in *Insider Attack and Cyber Security: Beyond the Hacker*, S. J. Stolfo, S. M. Bellovin, and A. D. Keromytis, Eds., Boston, MA, USA: Springer, 2010, pp. 69–90.

[8] CERT Division, "Insider Threat Test Dataset v6.2," Software Engineering Institute, Carnegie Mellon University. [Online]. Available: https://resources.sei.cmu.edu/library/ asset-view.cfm?assetid=508099

[9] CrowdStrike, 2024 Global Threat Report: Enter the Era of Data-Driven Security, CrowdStrike Holdings, Inc., Feb. 2024. [Online]. Available: https://www.crowdstrike.com/ global-threat-report/