Detecting and Mitigating Insider Threats with Artificial Intelligence

Paras Shigvan[1], Vishal Auti[2], Gaurav Kumar[3], Proff. Sheetal Shevkar[4]
[1,2,3]*Student, Maeers MIT Arts Commerce and Science College Alandi*
[4]*Assistant Professor, Maeers MIT Arts Commerce and Science College Alandi*

***Abstract*— Insider threat-threatening or irresponsible conduct of authorized users is an old and continuously developing danger to organizations. Traditional perimeter-based security tools have limited ability to identify this type of attack because insiders have valid access and understanding of the system. The latest achievements in artificial intelligence (AI) and machine learning (ML), behavioral analytics, anomaly detection, natural language processing (NLP), graph-based models, and federated learning provide the new opportunities to detect, prioritize, and reduce insider risks. The present paper outlines AI- based insider threat detection techniques, synopticizes the key issues (data scarcity, privacy, false positives, concept drift, and interpretability), integrates an insider threat detection pipeline, and finally, evaluates the performance criteria and future opportunities. Suggestions are also given on the need to combine technical controls with the organizational policies and human supervision.**

***Index Terms*— Insider Threat Detection, Artificial Intelligence, Machine Learning, Behavioral Analytics, Anomaly Detection, Natural Language Processing, Graph Neural Networks, Federated Learning, Privacy-Preserving AI, Explainable AI.**

I. INTRODUCTION

Insider threats are those attacks by people who have been given the right to access a system and use that right to damage the organization either with the intention or going against it. They can rob information, destroy property, defraud, or accidentally leave confidential information. Since such users are not only have valid credentials, but also know the architecture of systems, traditional security systems based on external signature or external perimeter protection are inadequate to identify such threats. As a result, companies will need to implement high-tech tools that can examine behavioral patterns, situational indicators and trends. According to numerous government researches and work of laboratories, one strategy that can be best used is a combination of technical surveillance equipment, organizational policies and a culture of awareness and responsibility.

II. THREAT TAXONOMY AND PROBLEM SPACE.

Insider attacks may be defined by intent (malicious and negligent), capability (privileged and regular user), and attack type (data access, system manipulation, or physical intrusion). Research has shown that users with privileges and system administrators are more dangerous because they have privileged statuses. Indicators can be given by the behavior and contextual signals of their displeasure, e.g. a breach of policies that have been previously implemented or signs of discontent. Thus, the overall detection plans should address technical elements (e.g., system logs, access history) as well as non-technical ones (e.g., HR incidents, behavioral patterns, and workplace environment).

III. AI APPROACHES FOR DETECTION

3.1 Behavioral analytics and anomaly detection.

We develop a model that gets to know what normal user behavior should look like through such information as file access, network activity, logins, commands and device activity. Machine-learning methods that are unsupervised or semi-supervised (such as clustering, autoencoders, one-class SVMs, isolation forests or deep-learning reconstruction models) are used to detect abnormal patterns that can be used to signal insider abuse. The models work well in cases where the instances of bad insiders are few but

they should be able to keep up with the changing roles, processes and tools.

3.2 Classification that is monitored and hybrid models.
When there is an input of labeled data (e.g. a history of past events) then we can apply supervised classifiers, including random forests, gradient-boosted trees, or neural networks, to distinguish malicious and normal behaviour. The system is divided into hybrid systems that combine the scores of anomalies with supervised risk scores to minimize false positives along with the rules, which are presented by the business domain. Relational models and graph neural networks are useful since they show connections between users, computers and files.

3.3 Sentiment analysis, natural language processing (NLP).
The intent of people within a system including requests to break a rule, passwords or swear words, are evidenced in the code comments as well as the ticket messages, chat logs and emails. Simple key words spotting systems, transformer systems and so on, can show the dangerous content, topic or mood change. Scanning communications must have privacy saving techniques and tough measures in order to make sure that we do not overstep the legal and ethical limit to be within the legal limits.

3.4 Relational and graph-based approaches.
The idea of insider activities may be viewed as weird relationships, including a weird chain of user-file-computer or a weird burst of lateral mobility. Graph techniques are directly used to map these relationships. They are based on graph neural networks, centrality measures and sub-graph anomaly detection which indicates suspicious patterns not detected by plain-feature methods. Research completed recently suggests that graphs can be used in complicated instances of insiders.

3.5 Privacy preserving federated learning and ML.
The sensitivity of the data- logs and communications of the behavior is confidential and legal thus one of the highest barriers. Federated learning enables businesses to develop a shared model without transmitting unprocessed data as well. Training the local model is done in every single organisation after which the model is simply shared with other organisations that then combine it to form global model. Early studies show that federated learning can be used to maximize the precision of insider threat detection and maintain the privacy of data. Other privacy assurances are given in terms of differential privacy, secure aggregation and homomorphic encryption.

IV. CHALLENGES AND LIMITATIONS

4.1 Data scarcity and labeling
The insider attacks are not a common occurrence as opposed to normal activity, hence an unbalanced data. It is expensive in terms of time and inconsistent when labeling this data. Thus, we apply unsuperved techniques and generate artificial data, but we should ensure that the counterfeit data is natural so that we do not deceive the models.

4.2 Privacy, legal and ethical limitations.
Monitoring users and their actions and documents they work with presents privacy and legal issues, such as labor laws and privacy regulations. We should have policies which are protective to the employees but at the same time maintain a safe workplace. Anonymization of data, maintaining records as long as necessary, ensuring that only particular roles are notified of the data, and good governance (transparency, oversight, legal basis) are useful. Governance, according to agencies like CISA and NIST is key in solving these problems.

4.3 False positive and alert fatigue rates are high.
Not every unusual action of the user is bad, such as a change of job or a legitimate task done rarely. False alarms are too much to damage trust and saturate security personnel. Combining noise with additional information, such as job title, planned work schedule, and ticketing, reduces the score of anomalies. Allowing humans to go through alerts also is useful. Explainability of the model allows the analysts to focus on alerts first.

4.4 Concept drift and adversarial behavior.
Individuals evolve their behaviors over time due to the new roles or software updates or change of processes. Bad insiders may attempt to avoid detection. To ensure that the model continues to work, we should continue updating it, employing online learning, and adding resiliency to attacks.

4.5 Interpretability and trust

Decisions on security should have explanations. A black box model may be able to arrive at the correct solution but when it comes to problem correction or legal action the model cannot be explained. Most explicable AI approaches such as SHAP, LIME, or rule extraction, as well as displaying features easily understandable by humans (e.g. 3 odd file transfers to an external host in 2 hours), make analysts feel more assured.

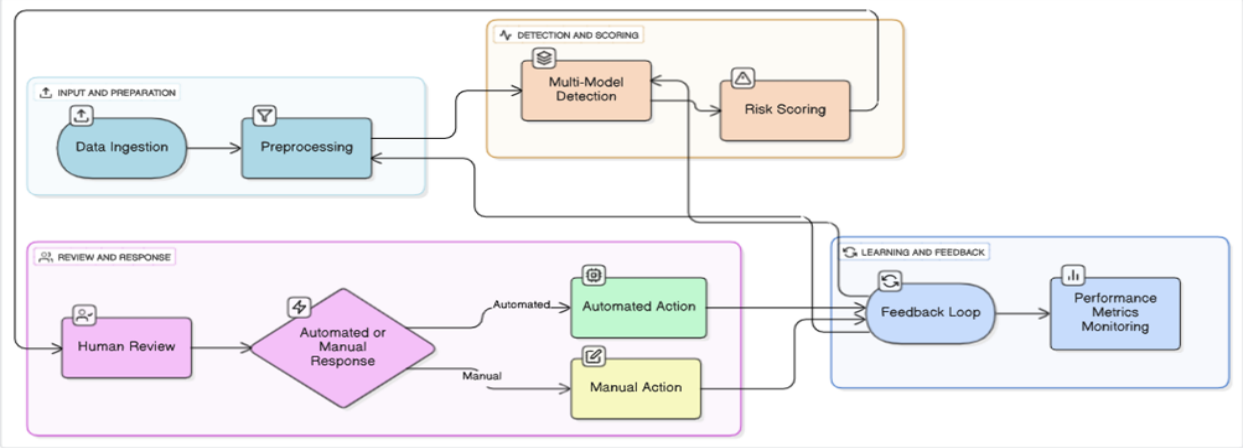
V. PROPOSED INTEGRATIVE DETECTION-AND-MITIGATION PIPELINE

A good insider threat management system integrates data ingestion, preprocessing, multi- model detection, risk fusion, explainability, and human oversight into a unified pipeline.

The process is initiated by a thorough data gathering on end points, authentication records, HR systems and network resources. This information is standardized into a single schema. Privacy preserving preprocessing guarantees pseudonymization, minimization of data and controlled access. Various detection models are run concurrently anomaly detectors detect baseline violations, supervised models measure known patterns of threats, NLP models can detect textual communications, and graph-based detectors can detect relational abnormalities.

The ensemble techniques and business logic are used to fuse outputs to produce a prioritized risk score. Interpretability explanations and confidence scores are provided to the analysts along with contextualized alerts. Each high-confidence alert results in automated actions (e.g. prevent data exfiltration), whereas lower-confidence instances are submitted to human consideration. Nevertheless, continuous improvement is achieved with the help of a feedback mechanism, where analyst labels and retraining of models are conducted.

Figure 1: Conceptual Figure 1: Conceptual Diagram- Integrative AI-Based Insider Threat Detection Pipeline.



VI. MEASURES OF EVALUATION AND BENCHMARKING.

Performance measures are concerned with recall (true positive rate), precision (false positive control), AUC-ROC and AUC-PR (when working with imbalanced data sets), Mean Time to Detect (MTTD), and Mean Time to respond (MTTR). Also, such variables as the workload of the analyst, the average lifetime of the alert, the ratio of false alerts allow to assess the utility of operations. A combination of synthetic datasets and anonymized real world events should be used in benchmarking and federated cross-organizational testing should be used to test scalability.

Metric	Description	Importance
Recall	Correctly identified threat	High
Precision	Correctly classified alerts	High
MTTD	Time to identify a threat	Medium
MTTR	Time to mitigate or respond	Medium
False Positive Rate	Frequency of wrong alerts	High

VII. EXAMPLES OF CASES AND PRACTICAL RESULTS

The regulatory and enforcement companies and the actors in the industry are initiating to pilot AI on insider threat and insider-like threat. An example of this is the application of AI by the market supervisors to detect insider trading trends faster and proving to save big in terms of executing early screening. Real studies and recommended controls that organizations may adapt are also reported in CERT and CISA guidance documents. These pilots in the real world demonstrate that AI will reduce the delay in detection and uncover patterns that are difficult to observe by humans alone, but again highlight the importance of close governance and analyst participation.

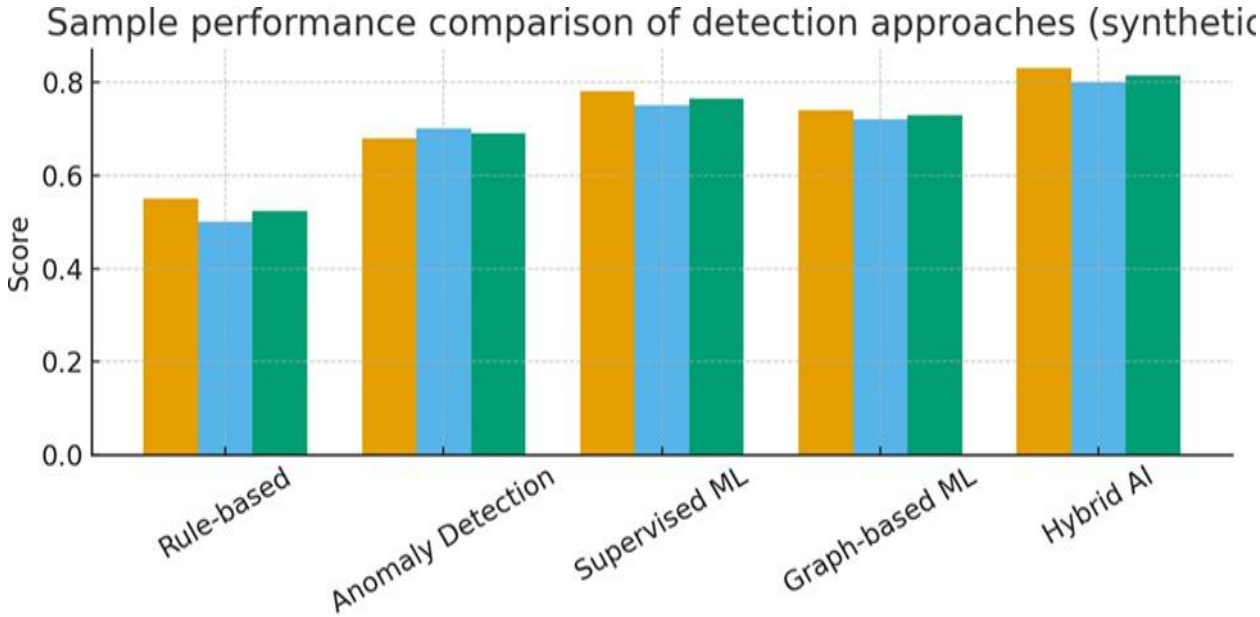


Figure 2: Sample performance comparison of detection approaches

VIII. BEST PRACTICES AND RECOMMENDATIONS.

Instead of using a single security solution, organizations are advised to use a layered security model, which combines the use of access controls and data loss prevention systems, as well as AI-based behavioral monitoring. All deployment steps (such as pseudonymization and controlled access to sensitive monitoring outputs) should be based on the principle of privacy-by-design. Human analysts should not be marginal to review and interpret AI alerts to make sure that the context is valid and fair. Performance should be monitored continuously to overcome concept drift and stay on track. Lastly, the overall insider threat program should be steered by cross-functional governance between HR, legal, IT and senior management to bring a balance between the security of the organization and the rights of employees.

IX. FUTURE DIRECTIONS

Several of these avenues of research are promising: Safe shared learning in which companies can pool insights without releasing logs; improved simulations and synthetic data used to learn on rare events; models which cannot be fooled by insiders; and more company context to detection models (such as project assignments and business events). It will be necessary to test such in real working environments in order to convert research wins into actual risk cuts.

X. CONCLUSION

AI will enlarge the tools to detect and prevent insider threats by analyzing behavior, relationships, and text in bulk. However, tech alone is not sufficient: privacy, legislation, its functionality in the day-to-day processes, the fact that we can describe the models, governance of it all. The combination of privacy-

saving AI, transparent guidelines, human auditors, and consistent verification can provide the most significant opportunity to reduce insider risks without losing trust or breaking the law.

REFERENCES

[1] Bin Sarhan, B., Insider Threat Detection Using Machine Learning Approach, Applied Sciences (MDPI), 2022

[2] SEI CERT Insider Threat Center, Common Sense Guide & Reports, Carnegie Mellon University.

[3] NIST Computer Security Resource Center — Insider Threat Definitions and Controls, csrc.nist.gov.

[4] CISA, Insider Threat Mitigation Guide, 2022.

[5] Gong, Y., Graph-Based Insider Threat Detection: A Survey, Elsevier, 2024.

[6] Ye, X., Research on Insider Threat Detection Based on Federated Learning, Nature Scientific Reports, 2025.

[7] Yilmaz, E., Harnessing Artificial Intelligence for Insider Threat Detection, ETASR, 2024.

[8] Reuters, Italy's Consob Tests AI for Market Supervision and Insider Trading Detection, 2024.

[9] Sharma, A. & Banerjee, P., Adversarial Robustness in Insider Threat Detection Systems, IEEE Access, 2024.

[10] Li, Z., Federated Graph Learning for Privacy-Preserving Insider Detection, ACM Transactions on Privacy and Security, 2025.

[11] Chen, L., Explainable Deep Learning for Insider Threat Analytics, Springer AI Journal, 2023.

[12] Ortiz, M., Synthetic Data Generation for Rare Cybersecurity Events, Journal of Cyber Intelligence, 2024.

[13] Wang, J., Human-in-the-Loop Models for Insider Threat Governance, Elsevier Computers & Security, 2025.