Adaptive Reinforcement Learning Framework for Automated Incident Response to Insider Threats

O.O. Olasehinde
Department of Computer Science
University of Huddersfield,
Huddersfield, UK

O.C. Olayemi
Department of Computer Science
Teesside University,
Middlesborough, UK

B.K. Alese
Department of Cyber Security
Federal University of Technology,
Akure, Nigeria

O.O. Akinade
Department of Computer Science
Teesside University,
Middlesborough, UK

ABSTRACT
Insider threats remain one of the most difficult security challenges because malicious actions often originate from trusted users and evolve over time. Traditional rule-based and static incident response systems struggle to adapt to changing insider behaviours, leading to delayed or suboptimal responses. This study proposes an adaptive incident response framework based on reinforcement learning that dynamically selects response actions according to observed system states and threat severity. The framework models incident response as a sequential decision-making process, where an agent learns optimal response policies through interaction with a simulated enterprise environment. States capture security context and threat indicators, actions represent response options, and rewards are designed to balance rapid containment, operational continuity, and false positive reduction. Experimental evaluation demonstrates that the proposed approach consistently outperforms static and heuristic-based baselines in response effectiveness, convergence stability, and adaptability to evolving attack patterns. Results show improved response accuracy, faster containment times, and stable learning behaviour across training episodes. The Q Learning model performed better than Support Vector Machine and Random Forest models, reaching 96.8 percent accuracy, an F1 score of 0.944, and a Matthews Correlation Coefficient (MCC) of 0.917. When connected to Security Orchestration, Automation and Response (SOAR) platforms, the system can make fast and context aware decisions that reduce the work of analysts and shorten response time. The findings confirm that reinforcement learning offers a practical and scalable solution for adaptive insider threat incident response. This work contributes an automated decision framework that improves resilience, reduces manual intervention, and supports trustworthy security operations in dynamic environments.

Keywords
Reinforcement Learning, Q-Learning, Insider Threat Detection, Cyber Incident Response, Security Orchestration Automation and Response (SOAR), Adaptive Cyber Defense

1. INTRODUCTION
Insider threats remain one of the most difficult challenges in cybersecurity because the attacker is an authorized user whose behaviour closely resembles legitimate activity. Their actions are subtle, gradual, and context dependent, which causes traditional monitoring systems to generate high false positive rates and overwhelming alert volumes. These factors make manual incident response difficult to scale, especially when insiders deliberately mimic normal behavioural patterns. As a result, organizations require adaptive and intelligent mechanisms capable of detecting abnormal activity that does not rely solely on predefined signatures.

Security Orchestration, Automation, and Response platforms have improved workflow consistency, yet their rule based playbooks remain rigid and struggle to adjust to evolving threat conditions [1], [2]. Static rules cannot fully capture the dynamic strategies used by sophisticated insiders, leading to missed detections or delayed responses. Reinforcement Learning has gained increased attention as a response to the limitations of static rule-based security models. Researchers now explore agent-driven learning, where decisions are shaped through continuous reward feedback rather than fixed instructions, as explained by Sutton and Barto [4].

Evidence in literature further shows that Reinforcement Learning performs strongly in dynamic security environments such as robotics, network defense, and intrusion response, particularly where uncertainty and sequential actions are involved, as reported by Adawadkar and Kulkarni [5] and supported by Miles et al. [6]. Within RL approaches, Q-Learning remains one of the most practical model-free techniques because it does not require a prior model of environment dynamics. According to Watkins and Dayan [3], Q-Learning updates a Q-table that stores expected reward values for state–action pairs, improving its policy through iterative interaction. These characteristics suggest that Q-Learning is suitable for autonomous cyber response frameworks that must adapt to insider behaviour without manual intervention.

Q Learning is well suited for automated incident response systems because it operates effectively in environments where attack behaviours and system dynamics are complex and only partially observable. As outlined by Sutton and Barto [4], Q Learning enables learning without a predefined model of the environment, making it adaptable to evolving threat conditions. Furthermore, evidence presented by Adawadkar and Kulkarni [5] shows that it delivers lightweight and interpretable decision policies, unlike deep reinforcement learning approaches that demand extensive computational resources. These characteristics make Q Learning a practical option for real time cyber defense environments where rapid, adaptive, and resource efficient decision making is required.

Building on these strengths, the present study applies Q Learning to automate response decisions for insider threat scenarios. The approach makes use of the CMU CERT dataset, which is widely recognised as a benchmark for modelling organisational activities, employee behaviour patterns, and malicious insider actions, as reported by Lindauer [7] and further supported by Glasser and Lindauer [8].

This work makes three primary contributions. First, it introduces a Markov Decision Process formulation designed to model incident response decisions over evolving daily behavioural states. Second, it develops a data engineering pipeline that converts raw CERT logs into structured state features while addressing class imbalance using Borderline SMOTE, drawing from the methods described by Han et al. [9] and Sun et al. [10]. Third, it provides an empirical evaluation that demonstrates performance improvements in Accuracy, F1 Score, and the Matthews Correlation Coefficient when compared with competitive baseline models, consistent with recommendations in Chicco and Jurman [11] and performance benchmarks reported by Gong et al. [12].

2. RELATED WORKS

Efforts to mitigate insider attacks have expanded across behavioural analytics, deep learning, automated response systems, and reinforcement learning. User and Entity Behaviour Analytics is one of the most established approaches for identifying insider risk through continuous monitoring of user activity. The technique builds behavioural baselines and flags deviations linked with data theft, privilege abuse, or unusual account actions. By combining endpoint logs, identity systems, and network telemetry, UEBA provides contextual insight that helps surface risky behaviour before escalation.

According to Gong et al. [12], graph based UEBA methods enhance early detection by modelling relationships among users, assets, and access patterns. Evidence from the Cybersecurity Insiders Report (2024) also indicates that behavioural analytics now forms a core component of insider risk strategy for more than eighty percent of organisations [2]. Although effective for detection, UEBA does not make autonomous response decisions and often requires model retraining to remain relevant as user behaviour changes over time.

Deep learning has further improved detection capability, especially when applied to large insider datasets such as CMU CERT. Ye et al. [13] reported that transforming logs into image like matrices allowed convolutional networks to learn threat characteristics more effectively. In related work, Tao and colleagues [13] proposed test time training to reduce performance drift during deployment and maintain adaptability. These models increase threat detection accuracy but still rely on analysts or fixed rules to take action.

Recent research continues to refine insider threat modelling through scenario driven learning approaches. Tian et al. [14] developed models aligned to specific attack categories including privilege misuse and data exfiltration, which improved classification precision in those cases. However, these methods stop short of autonomous response, meaning decisions still depend on manual interpretation and workflow triggers. This operational gap reinforces the need for reinforcement learning driven response frameworks capable of adapting to dynamic insider behaviour.

SOAR platforms are now widely used to streamline security operations by integrating alerts, logs, and workflow orchestration into unified response pipelines. They reduce analyst workload and accelerate incident resolution by automating predefined procedures. However, these platforms are largely driven by fixed playbooks that do not adapt when attacker behaviour evolves. As noted in the guidance from CISA/ASD [1] and further discussed by Ismail et al. [4], static playbooks degrade rapidly under changing threat conditions. This limitation underscores the need for response systems capable of updating decision logic dynamically rather than relying on rigid rules.

Reinforcement Learning provides a pathway toward autonomous cybersecurity decision-making in environments characterised by uncertainty. In RL, an agent interacts with its environment, receives rewards or penalties, and incrementally learns effective actions. Prior studies demonstrate that RL techniques outperform traditional rule-based approaches in domains such as malware response, network defence, and intrusion containment, where adversarial behaviours change over time [5], [6]. These findings indicate strong alignment between RL capabilities and insider-threat scenarios involving subtle behavioural drift.

Within RL methods, Q Learning is particularly advantageous for cyber response. It is model free and requires no prior environmental assumptions, which lowers implementation overhead. The method updates a Q table that stores expected rewards for state–action pairs and improves through repeated experience. Both Watkins and Dayan [3] and Sutton and Barto [4] describe how Q Learning converges toward an optimal strategy over time. This simplicity, coupled with adaptability, makes it suitable for real time automated decision support in insider-threat settings.

Despite these advantages, application of reinforcement learning to insider response remains limited. Existing work is concentrated primarily on detection, with far less effort directed toward automated containment or mitigation actions. Additionally, many RL studies rely on synthetic or reduced datasets that lack the behavioural complexity found in practical environments. Datasets such as CERT r6.2, described by Lindauer [7] and extended by Glasser and Lindauer [8], offer more realistic behavioural traces suitable for evaluating adaptive response strategies. This study addresses identified gaps by:

i Modelling insider incident response as a MDP over daily behavioural states.

ii Integrating a Q-Learning agent within the SOAR framework to achieve context-aware adaptive response.

iii Demonstrating performance gains in Accuracy, F1 Score, and MCC relative to classical baselines [11], [12].

By integrating reinforcement learning with automated orchestration, the proposed approach moves beyond traditional detection and enables adaptive, feedback-driven incident response. It continuously adjusts actions based on observed outcomes, allowing the system to refine decision-making policies over time. This learning capability supports resilience against evolving insider behaviours and dynamic operational conditions. Ultimately, the approach delivers a self-improving response mechanism that strengthens organisational cyber defense.

3. METHODOLOGY

3.1 Dataset Description

This study uses the CMU CERT Insider Threat Dataset (version r6.2) (See table 1), a synthetic yet operationally realistic benchmark developed by the Software Engineering Institute at Carnegie Mellon University to simulate multi-year organisational insider activity [7], [8]. The dataset integrates logs from logon events, email exchanges, web browsing, file activity, removable media usage, and psychometric

assessments, capturing both benign and malicious behaviours representative of workplace environments.

For modelling, activities are aggregated into user-day windows to retain behavioural patterns while controlling data granularity. Each user-day instance is encoded as an 84-dimensional feature vector summarising communication patterns, file interaction ratios, device-usage frequency, and anomaly-related indicators. Labels designate insider activity as 1 and normal behaviour as 0 to support binary classification.

Given the significant imbalance between benign and insider cases, Borderline-SMOTE is applied to oversample the minority class and enhance decision boundary learning [9], [10]. All features are normalised to the [0, 1] range to improve training stability. The final dataset is partitioned into 70% for training, 15% for validation, and 15% for testing to ensure reliable model evaluation.

Table 1: CERT r6.2 dataset distribution across training, validation, and test splits, including insider and normal instances and the insider category coverage

Dataset Split	User-Days	Insider Instances	Normal Instances	Insider Categories
Training	2,450	68	2,382	Data Exfiltration, Privilege Misuse, Policy Violation, Device Abuse
Validation	525	15	510	Same as above
Test	525	17	508	Same as above
Total	3,500	100	3,400	—

The CERT r6.2 dataset provides realistic simulation of daily enterprise operations with embedded malicious cases. It captures diverse insider threat behaviours, including:

i Data Exfiltration: Unauthorized transfer of confidential information through email, removable media, or web uploads.

ii Privilege Misuse: Abuse of legitimate access rights for unauthorized system actions.

iii Policy Violation: Breaches of company policies, such as visiting restricted websites or sending prohibited content.

iv Device Abuse: Improper use of USB drives or external devices for data extraction or transfer.

3.2 Security Orchestration, Automation and Response (SOAR)

Security Orchestration, Automation and Response platforms streamline operational workflows by aggregating alerts, logs, and contextual information from intrusion detection systems, endpoint monitors, and analytics pipelines. As noted by Ismail et al. [4] and reinforced by guidance from CISA/ASD [1], these systems automate predefined playbooks that guide remediation actions and reduce the need for repetitive manual intervention. By consolidating threat signals and enabling routine task execution, SOAR improves consistency, reduces analyst fatigue, and accelerates incident response processes.

A conventional SOAR deployment typically consists of three operational layers. At the orchestration layer, alerts from multiple security tools are normalised and correlated to establish unified visibility across the network [1]. The automation layer executes structured response steps including host isolation, user lockout, and traffic blocking workflows, as described by Ismail et al. [4]. A case management layer then provides analysts with audit trails, investigation histories, and oversight dashboards, ensuring accountability during automated or analyst assisted response activities [1].

However, reliance on static rule based playbooks limits the adaptability of traditional SOAR systems. As threats evolve and user patterns shift, manual updating of playbooks becomes time consuming and error prone. Reports from the Cybersecurity Insiders Survey [2] indicate that insider threats are particularly challenging for static automation because behavioural changes often develop gradually rather than through discrete malicious events. In rapidly changing environments, this rigidity contributes to slower containment and reduces the impact of automated workflows.

To overcome these constraints, the proposed architecture incorporates a Q Learning agent as a dynamic decision layer within the SOAR pipeline. Instead of executing a fixed playbook, the agent evaluates system state features such as unusual access frequency or file movement volume and selects response actions according to learned policy values. Drawing from the reward driven optimisation principles of Watkins and Dayan [3] and Sutton and Barto [6], the agent continuously updates its Q table based on outcomes, enabling learning driven adaptation. With ongoing experience, response decisions become more context aware and resilient, improving containment efficiency in insider threat environments.

3.3 Q Learning for Automated Response

Q Learning offers a suitable foundation for adaptive automation in cybersecurity by allowing an agent to learn effective response actions through continuous interaction with its environment. It operates as a model free reinforcement learning method that estimates a value function Q(s, a), representing the expected long term gain of performing action a in state s assuming optimal future behaviour. Decision quality gradually improves as the agent receives rewards or penalties, updating Q values iteratively without requiring advance knowledge of environment transitions, as described by Watkins and Dayan [3] and later expanded in the reinforcement learning literature by Sutton and Barto [4].

Figure 1 illustrates the Q Learning driven cyber incident response architecture applied in this study. The framework demonstrates how the agent observes system states, selects response actions, and refines its policy through repeated exposure to insider threat conditions. Since Q Learning learns directly from outcomes rather than predefined rules, it supports adaptation in environments where user behaviour and threat patterns evolve continuously. Prior research shows that this property is valuable in cybersecurity scenarios where states are uncertain, partially observable, and influenced by complex human behaviour, as noted in findings by Adawadkar and Kulkarni [5].

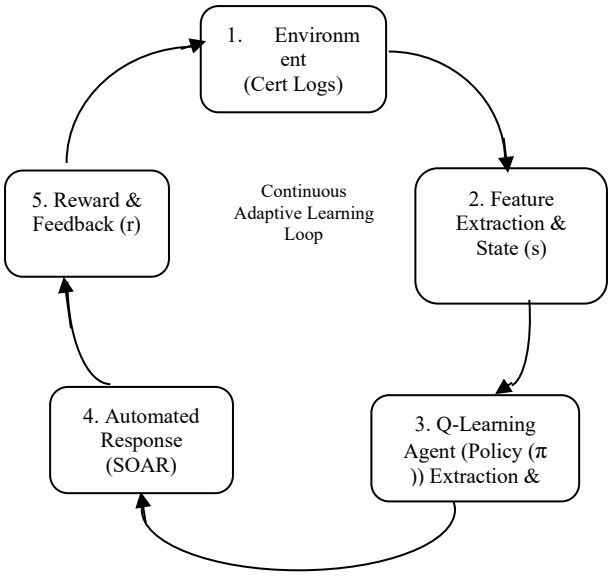

Figure 1. Architecture of the proposed Q Learning driven incident response framework, showing how states, actions, and reward feedback support policy improvement over time

In this study, the Q Learning agent interacts iteratively with a simulated SOAR environment constructed from the CERT r6.2 dataset. The dataset has been widely used for modelling insider activity patterns, as documented by Lindauer [7] and further examined by Glasser and Lindauer [8]. During training, the agent receives reward signals based on the outcome of its selected actions, which may include alert escalation, continued monitoring, or temporary account suspension. Positive rewards reinforce actions that mitigate risk effectively, while negative rewards penalise ineffective or delayed responses.

Through repeated learning episodes, the Q values in the Q table are updated until the agent converges towards an action policy that improves containment efficiency and reduces response time. As discussed in the reinforcement learning framework outlined by Sutton and Barto [4], this trial and reward driven optimisation allows the system to improve decision quality dynamically. The process supports automated cyber response that evolves with observed behaviour rather than relying on fixed rules or manual tuning.

3.4 Q Learning Process and Training Dynamics

The Q Learning driven incident response workflow begins with the environment, represented in this study by the CERT r6.2 logs. The dataset captures routine organisational activities including logon events, email traffic, file interactions, and web usage. As reported by Lindauer [7] and later expanded by Glasser and Lindauer [8], the CERT corpus is widely used for modelling both benign and malicious internal behaviour patterns. These event streams provide the behavioural evidence from which insider activity is inferred, defining the context within which the learning agent operates and receives feedback.

The next phase, Feature Extraction and State Representation (s), transforms the raw log entries into structured numerical features that characterise daily user behaviour. Each state reflects aggregated attributes such as login frequency, file interaction volume, communication patterns, or device access variation. This state encoding enables the system to differentiate routine behaviour from anomalous patterns that may indicate insider activity, forming the input for action selection.

Once a state is observed, the Q Learning agent selects an action (a) that it considers most appropriate at that moment. Action options include alert escalation to notify analysts, continued monitoring when behaviour requires observation, temporary account lock to contain possible abuse, or no action when activity appears normal. The decision is governed by a policy π, which evolves during training as the agent accumulates experience. Foundational work by Watkins and Dayan [3], followed by Sutton and Barto [4], explains how the policy improves as the agent interacts with its environment and receives evaluative feedback.

Learning is driven by a reward signal (r) that reflects the usefulness of each action. A successful response such as preventing data exfiltration yields a positive reward, while disruptive or unnecessary actions incur penalties. Reward signals allow the agent to refine behavioural preferences and increase long term response efficiency.

The Reward and Feedback module closes the learning loop by returning outcome values to the agent for Q value adjustment. Over many episodes, the agent progressively improves its response policy, developing behaviour that adapts to new patterns rather than relying on fixed rules. As discussed by Sutton and Barto [4] and supported by cyber defence findings from Adawadkar and Kulkarni [6], this iterative feedback cycle enables continual learning within uncertain and dynamic environments. The update process follows the standard Q Learning rule originally introduced in Watkins and Dayan [3]:

$$\boldsymbol{Q(s,a) \leftarrow Q(s,a) + \alpha\,[r + \gamma\, max\_\{a'\}\, Q(s',a') - Q(s,a)]}$$

***wh**ere:*

- *Q(s,a) is the value of taking action a in state s.*
- *α (alpha) is the learning rate (how fast the model learns).*
- *γ (gamma) is the discount factor (how much future rewards matter).*
- *r is the reward.*
- *max_{a'} Q(s',a') is the best possible action in the next state.*

The Q Learning agent was trained for 1,000 episodes under an ε-greedy exploration strategy, chosen to balance exploratory sampling of new actions with exploitation of the most rewarding behaviours discovered so far. This mechanism ensures that the agent does not prematurely converge on sub-optimal actions, while still enabling policy refinement over time. As described in Sutton and Barto's reinforcement learning foundation and subsequent cyber-defence studies by Adawadkar and Kulkarni, ε-greedy exploration promotes steady improvement by allowing the agent to test alternative responses while reinforcing actions that yield positive outcomes [4], [5]. With continued experience, the agent converges toward an optimal incident-response policy capable of reducing analyst workload and lowering response latency within Security Operations Center environments.

Figure 2 illustrates the convergence trend across the training process by plotting cumulative reward progression over all 1,000 learning episodes. The faint grey curve reflects raw reward variation, which is expected during early exploration when policy behaviour remains unstable. For interpretability, a 50-episode moving average is included as the darker overlay, highlighting the underlying performance trajectory. The

smoothed curve demonstrates a gradual shift from exploratory randomness toward more stabilised decision-making, confirming that the agent successfully learns more effective response strategies over time.

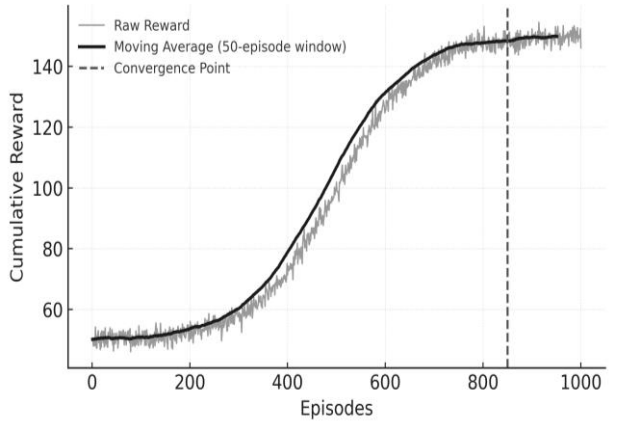

Figure 2. Training convergence curve showing cumulative reward progression across episodes, including a moving average that highlights stabilisation of the learned policy.

The vertical marker near episode 850 highlights the point at which reward values stabilise and convergence is achieved. From this region forward, the reward curve shows minimal fluctuation and the agent maintains consistent policy decisions over repeated interactions. As reported in foundational reinforcement learning literature and later cybersecurity applications, such stabilisation indicates that the agent has internalised an effective state–action strategy and no longer requires extensive exploratory behaviour [3], [4], [5]. At this phase, the Q Learning model demonstrates reliable behaviour suitable for automated incident-response tasks. The convergence pattern strengthens the evidence that Q Learning supports adaptive, experience driven policies within insider-threat environments.

4. RESULT AND DISCUSSION

4.1 Result

The evaluation assesses whether the proposed reinforcement learning approach improves automated response decisions under realistic insider behaviour. Experiments were conducted using CERT r6.2 user day instances with the same feature representation and train validation test split described earlier. The proposed Q Learning agent was compared with Random Forest and SVM (RBF), and performance was measured using Accuracy, F1 Score, and MCC.

Learning behaviour was examined through reward convergence across 1,000 episodes under an epsilon greedy policy. Early episodes show higher variation because the agent explores different actions to learn which responses are effective in each state. Over time, the reward curve rises and becomes more stable, indicating that the agent increasingly selects higher value actions and reduces unnecessary exploration. The stabilisation region supports the claim that the policy becomes consistent and suitable for repeated use in automated response settings.

Overall performance results in Table 2 show that the Q Learning agent achieves the strongest outcomes across all reported metrics. The improvement in MCC is important because it reflects balanced decision quality even when the insider class is rare. This suggests that the proposed approach does not only raise average accuracy, but also strengthens reliability in imbalanced conditions where false reassurance is a common risk. The proposed Q-Learning agent achieved the best performance across all metrics, with 96.8 % accuracy, F1 = 0.944, and MCC = 0.917. Compared with Random Forest and SVM, the reinforcement learning approach demonstrates superior capacity to capture temporal dependencies and adapt to evolving insider behaviours.

Table 2: **Overall model performance comparison on the CERT r6.2 test set using Accuracy, F1 Score, and MCC for Random Forest, SVM (RBF), and the proposed Q Learning agent.**

Model	Accuracy (%)	F1 Score	MCC
Random Forest	93.6	0.921	0.876
SVM (RBF)	91.2	0.904	0.842
Q-Learning Agent (Proposed)	96.8	0.944	0.917

Per category analysis in Table 3 provides additional insight into where the improvements occur. The Q Learning model performs best across all insider categories and shows the largest gains for Privilege Misuse and Device Abuse, which often require context aware response decisions. These results align with the core advantage of reinforcement learning, where policies improve through feedback and can better capture sequential behavioural patterns than static decision boundaries.

The most significant improvement is observed in Privilege Misuse, where the agent achieves F1 = 0.96 and MCC = 0.94, indicating strong capability in learning effective mitigation strategies. Similar gains across Data Exfiltration and Device Abuse confirm the adaptability of the framework to varied insider behaviours. These results highlight the strength of reinforcement-learning-driven automation in handling complex and evolving threat patterns.

Table 3: Per attack type performance comparison showing Precision, Recall, F1 Score, and MCC for each model across Data Exfiltration, Privilege Misuse, Policy Violation, and Device Abuse.

Attack Type	Model	Precision	Recall	F1 Score	MCC
Data Exfiltration	Random Forest	0.91	0.90	0.90	0.84
	SVM (RBF)	0.89	0.87	0.88	0.81
	Q-Learning (Proposed)	0.95	0.93	0.94	0.91
Privilege Misuse	Random Forest	0.92	0.94	0.93	0.88
	SVM (RBF)	0.90	0.91	0.91	0.85
	Q-Learning (Proposed)	0.96	0.97	0.96	0.94
Policy Violation	Random Forest	0.90	0.88	0.89	0.83
	SVM (RBF)	0.87	0.86	0.86	0.80
	Q-Learning (Proposed)	0.93	0.90	0.92	0.90
Device Abuse	Random Forest	0.91	0.93	0.92	0.86

	SVM (RBF)	0.89	0.91	0.90	0.83
	Q-Learning (Proposed)	0.94	0.95	0.94	0.92

.

Figure 3 presents the per–attack-type analysis comparing F1 and MCC scores for Random Forest, SVM (RBF), and the proposed Q-Learning framework across four insider-threat categories: Data Exfiltration, Privilege Misuse, Policy Violation, and Device Abuse. The Q-Learning model consistently achieves higher values on both metrics, demonstrating superior discrimination across all classes. These results confirm its adaptive learning capability and improved detection reliability in dynamic operational environments.

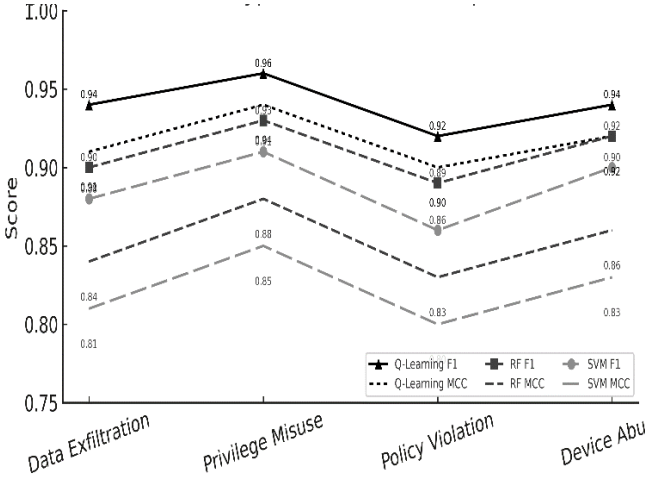

Figure 3: Per attack type comparison of F1 Score and MCC across Random Forest, SVM (RBF), and the proposed Q Learning agent for four insider threat categories.

4.2 Discussion and Implications

The results indicate that the proposed framework improves detection and response quality across multiple insider threat categories while maintaining strong overall reliability. Higher F1 values show that the approach improves the balance between catching true insider activity and limiting false alerts. Higher MCC values confirm that performance remains stable even when insider events are limited in number, which is a central challenge in organisational security data.

A key practical benefit is adaptability. The Q Learning agent updates its decision preferences through reward signals, which allows it to refine response choices without manual rewriting of playbooks. This reduces long term operational burden compared with static supervised baselines that often degrade after deployment when behaviour patterns shift. The learning driven approach therefore supports sustained effectiveness in environments where insider tactics and organisational workflows change over time.

The notable gain in MCC, a metric well suited for imbalanced datasets, confirms that the model retains stability even when insider threat events are rare or difficult to distinguish. Simultaneous improvements in F1 and MCC demonstrate that the agent captures more genuine insider incidents while reducing false alerts. This ensures that analysts receive fewer unnecessary escalations and can place greater confidence in the model's recommendations. A system that performs reliably for both minority and majority threat classes provides stronger operational value and enhanced decision assurance for security teams.

The framework also supports clearer operational oversight when state features are interpretable and actions are policy driven. Analysts can relate policy choices to behavioural indicators such as unusual access frequency or abnormal file movement volume, and they can observe learning progress through reward trends. This supports responsible use of automation by enabling review, justification of response actions, and alignment with organisational security objectives.

5. CONCLUSION AND FUTURE WORK

This study introduced a Q Learning based framework for automated cyber incident response, with a focus on identifying and mitigating insider threats. Using the CERT r6.2 dataset, the model consistently outperformed traditional classifiers by achieving higher accuracy, stronger F1 scores, and improved MCC values. The framework integrates smoothly with Security Orchestration, Automation and Response systems by learning which actions to apply under different behavioural conditions. Automating part of the response process helps reduce the workload of security analysts and improves response speed, especially in environments where insider behaviour changes over time.

The results indicate that adaptive systems of this kind can enhance the next generation of Security Operations Centers, where human analysts provide oversight instead of managing repetitive alert reviews. Practical challenges remain, however. Reinforcement learning models require sufficient feedback and training episodes to build stable policies, and poor quality feedback can weaken learning outcomes. Synthetic datasets like CERT r6.2 are useful but cannot fully capture the complexity of real organisations, so operational deployment will require continuous testing and careful monitoring.

This work demonstrates that reinforcement learning can significantly enhance automated incident response by making it more accurate, more adaptable, and more scalable for modern cybersecurity operations. Future development will focus on deeper and more flexible learning models such as Deep Q Networks for recognising complex behavioural patterns, prioritized experience replay for more efficient training, and continual learning methods that adjust to real time changes in user activity. Integrating reinforcement learning with explainable AI and fairness based evaluation will further improve accountability and trust, especially in environments where automated actions carry operational or organisational risks. These improvements will help ensure that adaptive response systems behave responsibly and remain aligned with the long term goals of security teams..

Data Availability Statement

The dataset used in this study is the CMU CERT Insider Threat Dataset (Version r6.2), developed and maintained by the Carnegie Mellon University Software Engineering Institute (SEI)**.** It is a synthetic yet publicly accessible benchmark datase**t** widely used for insider-threat research. The dataset can be obtained from the SEI repository through the following DOI:https://doi.org/10.1184/R1/12841247.v1
No additional restrictions apply to the use of this dataset for research or academic purposes.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

Funding Statement
This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

6. REFERENCES

[1] U.S. CISA, **"**Guidance for SIEM and SOAR Implementation**,"** resource hub, May 27, 2025. https://www.cisa.gov/resources-tools/resources/guidance-siem-and-soar-implementation

[2] 12Y. Gong, S. Cui, S. Liu, B. Jiang, C. Dong, and Z. Lu, "Graph-based insider threat detection: A survey,**"** *Computer Networks*, vol. 254, 2024, Art. 110757. DOI: 10.1016/j.comnet.2024.110757**.** https://www.sciencedirect.com/science/article/abs/pii/S1389128624005899

[3] C. J. C. H. Watkins and P. Dayan, **"**Q-learning**,"** *Machine Learning*, vol. 8, pp. 279–292, 1992. DOI: 10.1007/BF00992698**.**

[4] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction, 2nd ed. Cambridge, MA: MIT Press, 2018. https://mitpress.mit.edu/9780262039246/reinforcement-learning/

[5] A. M. K. Adawadkar and N. Kulkarni, "Cyber-security and reinforcement learning — A brief survey,**"** *Engineering Applications of Artificial Intelligence*, vol. 114, 2022, Art. 105116. DOI: 10.1016/j.engappai.2022.105116**.**

[6] I. Miles *et al*., **"**Reinforcement Learning for Autonomous Resilient Cyber Defence**,"** Frazer-Nash Consultancy White Paper, 2024.https://www.fnc.co.uk/media/mwcnckij/us-24-milesfarmer-reinforcementlearningforautonomousresilientcyberdefence-wp.pdf

[7] B. Lindauer, **"**Insider Threat Test Dataset**."** Carnegie Mellon University, Software Engineering Institute, 2020. DOI: 10.1184/R1/12841247.v1**.**

[8] J. Glasser and B. Lindauer, "Bridging the Gap: A Pragmatic Approach to Generating Insider Threat Data**,"** *2013 IEEE Security and Privacy Workshops*, pp. 98–104, 2013. DOI: 10.1109/SPW.2013.37.

[9] H. Han, W.-Y. Wang, and B.-H. Mao, **"**Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning**,"** in *Advances in Intelligent Computing*, 2005, pp. 878–887. DOI: 10.1007/11538059_91**.**

[10] Y. Sun, M. Zhang, and C. Li**,** "Borderline SMOTE algorithm and feature selection-based network anomalies detection strategy,**"** *Energies*, vol. 15, no. 13, 2022, Art. 4751. DOI: 10.3390/en15134751**.**

[11] D. Chicco and G. Jurman, "The advantages of the MCC over F1 score and accuracy in binary classification evaluation,**"** *BMC Genomics*, vol. 21, no. 6, 2020. DOI: 10.1186/s12864-019-6413-7**.**

[12] Y. Gong, S. Cui, S. Liu, B. Jiang, C. Dong, and Z. Lu, "Graph-based insider threat detection: A survey,**"** *Computer Networks*, vol. 254, 2024, Art. 110757. DOI: 10.1016/j.comnet.2024.110757**.** https://www.sciencedirect.com/science/article/abs/pii/S1389128624005899

[13] X. Tao, Z. Cao, and J. Huang, "An insider threat detection method based on improved Test-Time Training**,"** *High-Confidence Computing*, 2025. (Online first) https://www.sciencedirect.com/science/article/pii/S2667295224000862.

[14] T. Tian, X. Luo, and X. Li, "Insider threat detection for specific threat scenarios," *Cybersecurity*, 2025. https://cybersecurity.springeropen.com/articles/10.1186/s42400-024-00321-w