Machine Learning Based Control Framework for Mitigating Insider Threats in Proxy-Controlled Environments

1st Balasirisha Juluri
Scientist/Engineer-'SC'
National Remote Sensing Centre, ISRO
Hyderabad, India
balasirisha_j@nrsc.gov.in

2nd Mahim Saxena
Scientist/Engineer-'SC'
National Remote Sensing Centre, ISRO
Hyderabad, India
mahim_saxena@nrsc.gov.in

3rd Mounika K
Scientist/Engineer-'SC'
National Remote Sensing Centre, ISRO
Hyderabad, India
mounika_k@nrsc.gov.in

4th Kishore SVSRK
Scientist/Engineer-'SC'
National Remote Sensing Centre, ISRO
Hyderabad, India
kishore_svsr@nrsc.gov.in

doi: https://doi.org/10.37285/bsp.SACAD2025.51

***Abstract*— Insider threats are one of the most significant cyber security risks facing organisations today. Unlike external attacks, which exploit vulnerabilities in systems, these threats originate from authorised users. These users could be malicious employees, negligent staff or external attackers who have hijacked legitimate credentials. Insiders can misuse their access intentionally or unintentionally compromise the organisation's confidentiality, integrity, or availability of information and systems. Insider threats are difficult to identify with rule-based traditional security mechanisms. This paper presents a Machine Learning based control mechanism for the detection and mitigation of insider threats by identifying anomalous user activities from gateway proxy logs. These machine learning models establish behavioural baselines and detect deviations from legitimate usage. The proposed approach incorporates a control mechanism that not only generates alerts but also enforces real-time responses, such as restricting or disconnecting anomalous users from accessing the Internet. It offers an adaptive solution for securing enterprise systems through proactive anomaly detection and control-driven threat mitigation.**

Keywords*— *Cyber Insider Threats, Cyber Security, Machine Learning, behaviour Analysis, Anomaly detection

I. INTRODUCTION

When it comes to the cyber security chain, people who have legitimate access to organisation facilities, systems, and data are found to be the weakest link, whether intentionally or unintentionally. They are the "insiders". An insider is any person who has or had authorised access to or knowledge of an organisation's resources, including personnel, facilities, information, equipment, networks, and systems[1].

There is a general consensus among most experts that insider threats are prevalent and continue to grow by the day. It is found that two-thirds of all security events are caused by insiders. However, the vast majority of these are caused by unintentional insider threats. These threats are difficult to identify because traditional security devices and solutions are designed primarily to detect malicious activities. The disruptive impacts of insider-related incidents jeopardise safety, undermine value, and degrade operations[2]. Despite many threats, a large number of organisations do not have an insider threat mitigation program to manage the risk.

Most of the time, it is observed that procedures exist to monitor organisation networks for intrusions and the collection of various logs for network analysis[5], but they do not have an equivalent capability to proactively detect and respond to insider activities, especially malicious disruptions that are a frequent expression of insider threats. One of the main challenges is the lack of controls designed to identify and respond effectively to insider behaviour, especially unintentional threats.

Hence, it becomes quite essential for organisations to implement appropriate security measures and policies to mitigate the risk of insider threats, which need to be able to detect and identify improper or illegal actions, assess threats to determine levels of risk, and implement solutions to manage and mitigate the potential consequences of an insider incident[1]. This calls for the usage of machine learning or deep learning models that are capable of protecting organisations from these threats.

In our organizational environment, the sensitivity and confidentiality of data make the volume of data transfer a critical indicator of potential risk. Unusual spikes in upload or down load volumes, particularly during non-business hours or from atypical user accounts, may indicate suspicious or unauthorized activities. Therefore, focusing on data

movement volume is a contextually strong and relevant metric for detecting anomalous user behaviour in our environment.

Machine learning techniques helps in creating behavioural baselines for individual employees and detects deviations by using data-driven methodologies to identify malicious intent. They are capable of recognising any peculiar or suspicious behaviour or instances where there are irregularities from normal everyday patterns or usage[4] and sends out an alert if it sees abnormal user behaviour. For example, if a particular employee downloads files of 100 MB regularly during particular time of the day in the network every day but starts downloading 10 GB of files at some other time, then the designed system would consider this an abnormal behaviour and alerts an IT administrator, or if automations are in place, automatically disconnect that user from the network.

ML models exhibit adaptive learning capabilities, enabling them to evolve with changing access behaviours and emerging attack strategies. ML solutions are capable of detecting threats, those that might be undetectable day to day, but over time display a different pattern. It helps organisations assess risks and mitigate threats before bad actors can traverse networks and cause serious damage.

II. LITERATURE SURVEY

A. Insider Threats in Organisational Security

1) Definition and Types of Insider Threats: Insider threats are cybersecurity risks that originate from individuals within an organisation who have access to sensitive data, systems, or networks. Based on the intentions of the individuals, insider threats can be of 2 types: Intentional, where individuals deliberately harm the organisation, or Unintentional, which arises from careless or negligent behaviour on part of the individuals. The primary types of insider threats include ma licious insiders, ignorant insiders, and compromised insiders. Malicious insiders are employees who steal, leak, or sabotage sensitive information. Ignorant insiders inadvertently cause harm due to non-compliance with security policies, lack of awareness, or human error. Compromised insiders are those whose credentials or systems have been taken over by external attackers, leading to unauthorised access to organizational assets[3].

2) Impact of Insider Threats on Organisations: Insider threats can have severe impacts on organisations, encompass ing a range of detrimental effects[3]. Theft of funds, fraud or cost to remedy the effects results in a large financial loss to an organisation. Another impact of insider threat is during a data breach which involves loss of sensitive information like intellectual property, private information about the customers, or any classified business or security information. Insider threats also lead to reputational damage, which in turn leads to a damaging impact on trust from customers, partners, and the public, causing a decline in business and market value. These can also cause major operational disruption due to unauthorised access of critical systems.

3) Challenges in Detecting Insider Threats: There are various challenges when it comes to detecting insider threats. One significant challenge is trust and access[3]; insiders often have legitimate access to systems and data, making it harder to flag potentially malicious activity as a threat. Insider threats can involve tactics like lateral movement or privilege escalation, which, the traditional security measures may struggle to identify. There may be a scenario where the large volume of data generated by user activities can flood the security monitoring tools, thereby making it difficult to identify the abnormal behaviour[3]. The use of multiple security tools can increase the complexity and make it harder to manage and analyse the data effectively.

B. Traditional Approaches to Insider Threat Detection

Traditional approaches to detect insider threats involve implementing strict access controls to limit the ability of employees to access sensitive data and systems. This is based on Role-Based Access Control (RBAC), where permissions are assigned according to an individual's role within the organisation. Another essential approach is network monitoring, which uses intrusion detection systems (IDS) and firewalls to monitor network traffic for suspicious activities and potential breaches[3]. There is usually a security team that analyses the logs generated by all the devices connected to the network on a regular basis to check for possible policy violations. Encryption of sensitive data is essential to protect against unauthorised access and potential data breaches. These traditional measures, while foundational, need to be supplemented with more advanced techniques to address the evolving nature of insider threats[3].

The methods mentioned above are reactive rather than proactive, which rely on detecting known threats or responding after an incident has occurred. They do not have the capability to pre-emptively identify and mitigate these risks. The evolving threat landscape poses the next level of challenge, as these insider threats employ new tactics to bypass the traditional security tools in place.

C. Machine Learning-Based Techniques

Machine Learning has become a potent weapon in the field of cybersecurity, offering the capability to detect and mitigate potential security risks posed by insiders[6]. By harnessing the potential of advanced algorithms and data analytics, machine learning empowers organisations to proactively identify and respond to insider threats, safeguarding critical systems and sensitive data[6]. Machine learning solutions offer several advantages in detecting insider threats, which are notoriously difficult to identify due to their subtle and complex nature[3]. The Machine Learning models, when deployed, run con tinuously and investigate the user behaviour to detect any anomalies or malicious intent. These models are trained with historical data of each user so that they establish a baseline of normal behaviour. Whenever there is a deviation from the normal behaviour, these models will report it as insider threats. The Machine Learning models can be trained using the new data as and when it is available to increase the accuracy of the result by reducing the false positives which will ensure that the security team can focus on genuine threats.

Early approaches relied primarily on classical ML models applied to statistical features extracted from logs because of their ability to model normal activity and detect deviations. These models are effective for coarse anomaly detection, and they are often limited in their ability to capture temporal correlations and fine-grained patterns.

To overcome the limitations of classical approaches, re searchers have explored time-series–based models and recur rent models to capture anomalous behavior in employee activity. However, employee network usage is sparse and

irregular, with long idle intervals and that are not continuous like regular time series data. Thus, these usage patterns often become weak or non-informative, reducing the effectiveness of modeling. Likewise, RNN-based methods also fail to accurately model the anomalous patterns as they struggle to capture long term employee behaviors, as their activity spans multiple days.

More recently, research has been shifted toward Convolutional Neural Networks (CNNs) and image-based representations of behavioural data. CNNs are effective in extracting local temporal bursts, periodic access patterns, and global structural anomalies that are difficult to identify in one dimensional formats. Using deep residual networks such as ResNets has emerged as a powerful mechanism for extracting features from such representations to uncover anomalies with higher accuracy, ensuring security teams can focus on genuine threats.

III. METHODOLOGY

In the proposed framework, the organisation employs multiple perimeter-level devices like firewalls with Intrusion Prevention/Detection Systems (IPS/IDS), Distributed Denial of Service (DDoS) protection systems and other security appliances to safeguard internal systems. In addition, employee access to the Internet is also strictly regulated by deploying a proxy server behind these perimeter devices that serves as the control point for all outbound Internet connections. Internet access is given only to the employees who are authorised by the administrator. Thus, the proxy server acts as both an enforcement and monitoring entity. If any suspicious activity is detected, then that anomalous connection is blocked at proxy level which prevents external communication, and this restriction can protect entire organisation's network from potential compromise.

By utilising advanced machine learning techniques, this pa per describes the framework that identifies insider threats with greater accuracy by analysing data from multiple dimensions of proxy server logs to identify abnormal or suspicious activity patterns. The methodology is detailed as follows:

A. Data Collection

We have collected employee logs from multiple proxy servers deployed at different campuses to gather relevant user activity, active timeline, and access patterns. For this study, the total size of the data acquired from all the proxy servers was found to be 113 GB for a period of 180 days.

B. Feature Selection

The data collected from these sources is processed into a structured format, and the following key features are extracted

1) Usernames: To capture the request initiator.
2) Timestamps: To analyse the temporal patterns of user activity, such as unusual login times or session durations.
3) Request Size: For evaluating the volume of data transfers that may indicate abnormal behaviour of an employee.

C. Feature Engineerin

We have applied advanced feature engineering techniques. Encoded each user's activity in one day into a single image, which will be a dataset for neural-network feature extraction. This 2D image representation enables the detection of fine-grained temporal anomalies such as minute-level spikes, short-term bursts, repeated short bursts, out-of-hours access, and prolonged abnormal behaviours. It also facilitates the analysis of complex time series patterns using machine learning models which are often challenging to capture effectively through tabular data alone.

1) Data Aggregation by Time Interval: Day-wise log data for each user is grouped and aggregated at one-minute intervals. Total data volume transferred (based on request size) in each interval is computed. We have captured user activity intensity over time and highlighted fluctuations in data usage patterns using the averaging process captures.
2) Image Generation through Pixel Coding: Aggregated data is transformed into colormap images where each pixel represents a one-minute time slice, and the pixel's intensity encodes the volume of data transferred per user at one-minute intervals throughout the day.

Fig. 1., Fig. 2., and Fig. 3. below shows the horizontal axis representing minutes within each hour, which captures fine-grained activity at a minute-level resolution. The vertical axis represents the 24 hours, divided into hourly segments. The intensity or color of each pixel encodes the volume of data transferred during that specific minute of the corresponding hour.

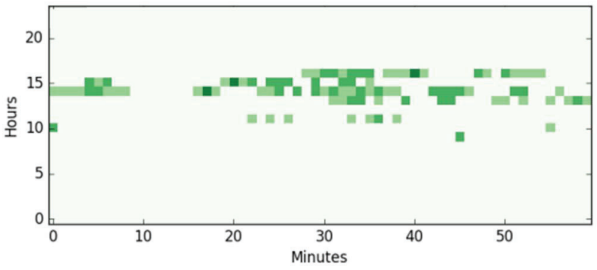

Fig. 1. Colormap Representation of One-Day Activity of an Employee.

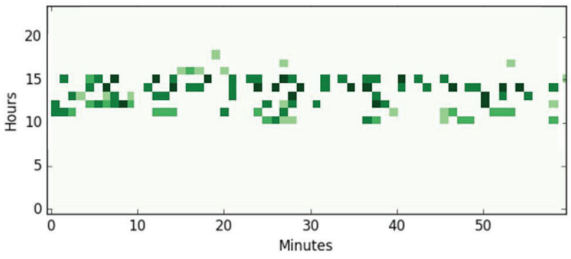

Fig. 2. Colormap Representation of the Same Employee's Activity on a

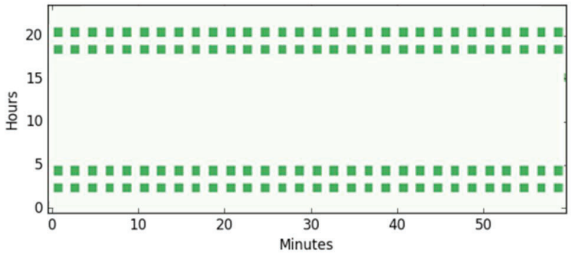

Fig. 3. Colormap Representation of a Test Anomalous Scenario

Number of Preprocessed rows for everyday activity for each employee	24*60 = 1440
Number of Preprocessed rows for everyday activity for 921 employees	1440*921 = 1326240
Total number of Preprocessed rows for 921 employees for 180 days	1326240*180 = 238 Million approx.
Total Number of processed datasets (PNG images)	180*921=165780

TABLE I: DESCRIPTION OF THE DATASET

D. Feature Extraction with ResNet50

Each colormap image is fed into ResNet50 network, which processes the image through multiple convolutional and residual blocks, and captures the features such as local patterns, global structures like repeated activity cycles. A fixed-length feature embedding vector will be generated for each image by encapsulating the user's activity pattern.

E. Feature Scaling and Dimensionality Reduction using PCA

Feature vectors extracted from ResNet50 model typically contain high dimensional and may contain features with varying magnitudes. To ensure each feature contributes equally to the learning process, we have applied the StandardScaler technique, which is sensitive to the scale of the data and particularly important for the machine learning models we have used. Principal Component analysis is employed to reduce the dimensionality of the standardised features.

F. Modeling for Anomaly Detection

To detect deviations from typical user behaviour, various machine learning algorithms are evaluated to determine the most suitable models that are capable of detecting deviations from established user behaviour. The core idea is to learn a baseline of normal activity for each user and flag instances that deviate significantly from the baseline as potential anomalies.

1) Model Selection: A combination of the following models are used for robust anomaly detection
   - One-Class Support Vector Machine (OC-SVM): OC-SVM learns boundaries around the majority of the training data and classifies new points falling outside the boundary as outliers.
   - Isolation Forest: It builds multiple binary trees to recursively partition data and assigns anomaly scores based on how early a data point is isolated.
   - Gaussian Mixture Model (GMM) with Isotonic Regression: GMM assumes that the data are generated from a mixture of several Gaussian distributions and assigns probabilities to each point, indicating how likely it belongs to the learned distributions.
2) Per-User Model Training: Each model is individually trained for each user on the data collected for six months, which uniquely captures the behaviour patterns of each user and increases the detection sensitivity.
3) Model Application and Prediction: For each user, new daily activity data are processed in real time, and user-specific trained models will evaluate whether the behaviour falls within the learned pattern.

IV. RESULTS AND EVALUATION

The study delved into comprehensive explorations of user behaviour in identifying insider threats. Through detailed data analysis and feature transformation, critical behaviour patterns were uncovered.

Several metrics are computed to assess machine learning models performance like recall, ROC AUC Score, accuracy and precision. These metrics are calculated on a per-user basis using anomalous test cases. We used an 80/20 train-test split for each algorithm and a small set of anomalous test cases was added to evaluate the models' ability to identify deviations from normal behaviour. The table below summarises the results of the experiments conducted using all three machine learning algorithms for a single employee.

Algorithm	Recall	Precision	ROC AUC Score	Accuracy
One-Class SVM	1	0.875	0.97	0.96
Isolation Forest	1	0.875	0.97	0.96
Gaussian Mixture Model	1	0.93	0.98	0.98

TABLE II: DESCRIPTION OF THE RESULT FROM THREE MODELS

Notably, the table's recall values confirm that all anomalous cases were correctly identified, which reflects the sensitivity of the models to the potential threats.

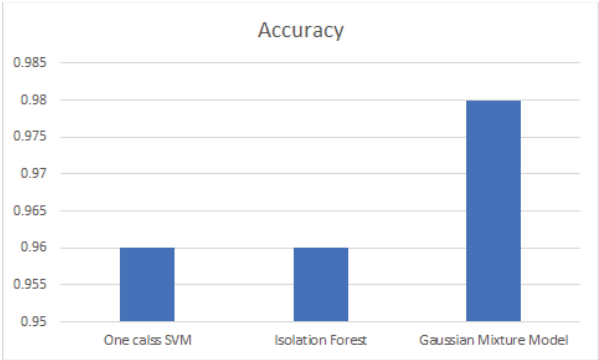

Fig. 4. Accuracy Chart

We have employed an ensemble-based decision strategy to enhance the robustness and reliability of anomaly detection by combining the outputs of all three models. The final decision logic is defined as follows:

- An event is classified as a confirmed anomaly (alert) only if all three models independently flag the activity as anomalous.
- If at least one of the models identifies the behaviour as anomalous, but the others do not, the event is categorised as a warning.
- If none of the models flags the event as anomalous, then it is treated as normal behaviour.

This multi-model consensus approach enables maintaining a balance between false positive reduction and early warning capability which enables the security team to prioritize high-confidence alerts for immediate investigation. This also helps to monitor warning-level events for patterns that may evolve into threats over time. The study's outcomes provide a solid foundation for tailored insider threat identification strategies and emphasize the need for continuous model optimisation to adapt to dynamic user behaviour over time.

As part of real-time monitoring, we have automated the system to generate daily user activity reports, which summarizes alerts and warnings triggered for each employee. It also includes a snapshot of the colormap usage pattern image used for detection. We are also generating monthly reports by aggregating the daily reports results to analyse employee usage trend. This reporting mechanism helps security teams track long-term behavioural trends of employees, identify them with recurring suspicious activity, and proactively investigate potential insider threats.

V. Conclusion

In this study, our approach was based on machine learning for detecting insider threats using user behaviour analytics. Created datasets using organisation proxy server logs and generated visual representations of user activity. We also applied advanced feature extraction techniques using ResNet50, followed by dimensionality reduction through PCA. We have used unsupervised machine learning models training on user-specific data to learn baseline behaviours and for detecting deviations effectively.

While the current system shows promising results, several enhancements are planned for future research to incorporate data from additional sources, such as file access logs and endpoint monitoring tools, to improve context and detection accuracy.

VI. References

[1] CISA Insider Threat Mitigation Guide. https://www.cisa.gov/sites/default/files/2022-11/Insider%20Threat%20Mitigation%20Guide Final 508.pdf (accessed Apr. 10, 2025).

[2] The Ultimate Guide to Building an Insider Threat Program. https://www.lapcom.com.hk/attachment/upload/ObserveIT%20-%20The%20Ultimate%20Guide%20to%20Building%20an%20Insider%20Threat%20Program%20eBook.pdf (accessed Apr. 10, 2025).

[3] Arfaoui, S. (2024). Enhancing Insider Threat Detection: A Literature Review on AI-Driven Solutions Leveraging Wearable Technology. New York University. https://doi.org/10.58153/r6w87-k8r78 https://www.proofpoint.com/us/blog/insider-threat-management/securing-your-weakest-link-announcing-our-brand-new-guide-building

[4] Fortinet EUBA https://www.fortinet.com/kr/resources/cyberglossary/what-is-ueba (accessed Apr. 08, 2025).

[5] ProofPoint. https://www.proofpoint.com/us/blog/insider-threat-management/securing-your-weakest-link-announcing-our-brand-new-guide-building (accessed Apr. 08, 2025).

Mohammed Nasser Al-Mhiqani, Tariq Alsboui, Taher Al-Shehari, Karrar hameed Abdulkareem, Rabiah Ahmad, Mazin Abed Mohammed, Insider threat detection in cyber-physical systems: a systematic literature review, Computers and Electrical Engineering, Volume 119, Part A, 2024, 109489, ISSN 0045-7906, https://doi.org/10.1016/j.compeleceng.2024.109489. https://www.sciencedirect.com/science/article/pii/S0045790624004166(accessed Sep. 09, 2025).