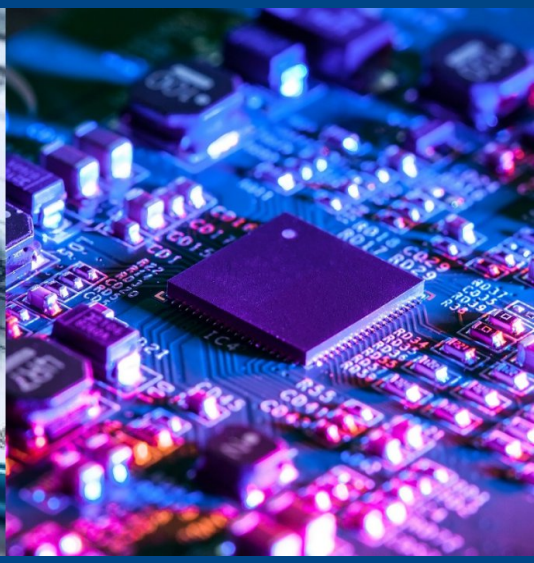
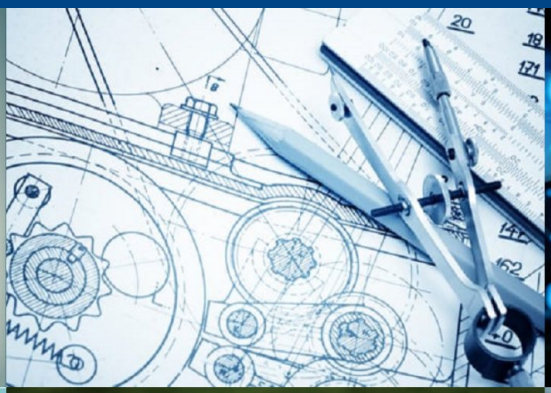
ISSN: 2582-7219

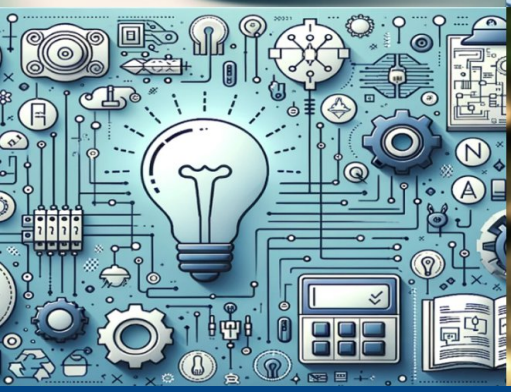




International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.206 **Volume 9, Issue 4, April 2026**

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904375

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

A Hybrid User Behavior Analytics Approach for Insider Threat Detection in Imbalanced Cybersecurity Environments

Manoranjen A[1], Dr.R.Shanthi[2]

MCA 2nd Year Department of Computer Applications, B.S Abdur Rahman Crescent Institute of Science and Technology, Vandalur, Chennai, Tamil Nadu, India[1]

Assistant professor Department of Computer Application, B.S Abdur Rahman Crescent Institute of Science and Technology, Vandalur, Chennai, Tamil Nadu, India[2]

ABSTRACT: Insider threats remain a significant challenge in organizational cybersecurity, as they originate from authorized users with legitimate access to sensitive systems and data. Detecting such threats is difficult due to the subtle and unpredictable nature of malicious behavior. This paper proposes a hybrid anomaly detection approach combining Isolation Forest and Local Outlier Factor (LOF) to identify abnormal user activities. The system utilizes the CERT 4.2 dataset and focuses on key user activities including logon, file access, device usage, and email behavior. Behavioral features are extracted to model normal user patterns, and deviations from these patterns are detected as anomalies. A risk scoring mechanism is employed to assign scores to users based on their level of suspicious activity. Experimental observations indicate that the proposed approach effectively identifies anomalous behavior and provides a scalable solution for insider threat detection using user behavior analytics.

KEYWORDS: Insider Threat Detection, Anomaly Detection, Isolation Forest, Local Outlier Factor (LOF), User Behavior Analytics, Risk Scoring

I. INTRODUCTION

With the rapid growth of digital infrastructure and enterprise systems, organizations increasingly rely on interconnected networks to manage sensitive data and critical operations. While this transformation has improved efficiency and accessibility, it has also introduced significant security challenges. Among these, insider threats have emerged as one of the most difficult and damaging forms of cybersecurity risk. Unlike external attackers, insiders such as employees, contractors, or trusted users already possess authorized access to organizational resources, making their malicious or negligent actions harder to detect and prevent.

Insider threats can take multiple forms, including intentional activities such as data exfiltration, fraud, and system sabotage, as well as unintentional actions like accidental data leakage or policy violations. The impact of such threats can be severe, leading to financial loss, reputational damage, and compromise of sensitive information. Traditional security mechanisms, including firewalls and signature-based intrusion detection systems, are primarily designed to defend against external attacks. As a result, they are often ineffective in identifying abnormal behavior originating from legitimate users operating within trusted environments.In recent years, the increasing availability of user activity data has enabled the development of behavior-based security approaches. Organizations routinely collect logs such as logon records, file access events, and device usage data, which provide valuable insights into user behavior patterns. However, analyzing this data is challenging due to its high volume, complexity, and the inherent imbalance between normal and malicious activities. Conventional rule-based systems rely on predefined patterns and thresholds, which limits their ability to detect subtle deviations or previously unseen insider attack strategies.To address these limitations, machine learning techniques have been widely explored for insider threat detection. By learning normal behavioral patterns from historical data, these models can identify anomalies that may indicate suspicious activity. Anomaly detection algorithms, such as Isolation Forest and density-based methods like Local Outlier Factor (LOF), are particularly suitable for this task as they can effectively handle imbalanced datasets where malicious events are rare but critical.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

By integrating machine learning with user behavior analytics, it is possible to develop intelligent systems capable of continuously monitoring user activity and detecting potential threats in real time. Such systems can reduce reliance on manual monitoring, improve detection accuracy, and enable proactive security measures within organizations.

The proposed study focuses on developing a machine learning–based insider threat detection system that analyzes user activity logs to identify abnormal behavioral patterns. The system incorporates data preprocessing, user-day aggregation, feature engineering, anomaly detection techniques, and a risk scoring mechanism to prioritize suspicious users. The objective is to provide a scalable and effective solution for detecting insider threats while minimizing false positives in real-world enterprise environments.

II. LITERATURE SURVEY

Research on insider threat detection has gained significant attention due to the increasing complexity of organizational systems and the limitations of traditional security approaches. Early studies primarily focused on rule-based and signature- driven methods; however, these approaches were found to be ineffective in identifying insider threats, as malicious insiders often operate within normal behavioral boundaries. As a result, recent research has shifted towards data-driven and machine learning–based techniques.

A major challenge in insider threat detection is the imbalanced nature of datasets, where malicious activities represent only a small fraction of overall user behavior. Taher Al-Shehari et al. [1] addressed this issue by proposing a Density-Based Local Outlier Factor (DBLOF) approach,which focuses on identifying anomalies based on local density deviations. Their work demonstrated high detection performance, particularly in handling skewed datasets such as CERT r4.2, achieving strong F-score results. Similarly, in a follow-up study, Taher Al-Shehari et al. [2] utilized the Isolation Forest algorithm to detect insider threats without altering the original dataset distribution. Their approach emphasized anomaly detection at the algorithm level rather than relying on data balancing techniques such as oversampling or under sampling, achieving improved accuracy and robustness.

Behavior-based detection has also been widely explored as an effective strategy. Jin Kim et al. [3] proposed a user behavior modeling framework combined with anomaly detection techniques to identify suspicious activities. Their study highlighted the importance of capturing temporal and contextual patterns in user actions to improve detection accuracy. In a similar direction, Duc Le and Natalie Zincir- Heywood [4], [5] introduced unsupervised ensemble methods for anomaly detection, demonstrating that combining multiple models can enhance detection performance and reduce false positives in complex datasets. The lack of real-world datasets has been another major limitation in insider threat research. Eugene Glasser and Brian Lindauer [6] addressed this challenge by proposing methodologies for generating realistic insider threat datasets, which have become widely used benchmarks in the research community. Additionally, structural anomaly detection techniques have been explored by William Eberle and Lawrence Holder [7], who focused on identifying irregular patterns in graph-based data, providing insights into complex relational behaviors. With the advancement of deep learning, more sophisticated approaches have emerged. Andrew Tuor et al. [8] applied deep learning models for unsupervised insider threat detection in structured data streams, demonstrating the ability of neural networks to capture complex behavioral patterns over time. Earlier foundational work by Benjamin Salem et al. [9] provided a comprehensive survey of insider threat detection techniques, categorizing various approaches and highlighting the challenges associated with insider attacks. Similarly, bioinformatics-inspired intrusion detection methods were explored by Scott Coull et al. [10], emphasizing pattern recognition techniques for anomaly detection. Despite these advancements, several challenges remain. Many existing approaches struggle with high false-positive rates, scalability issues, and the need for continuous adaptation to evolving user behavior. Furthermore, handling highly imbalanced datasets without compromising detection accuracy remains an open research problem.

In this context, the present work builds upon prior studies by integrating multiple components, including data preprocessing, user behavior modeling, anomaly detection using Isolation Forest and LOF, and risk scoring mechanisms. The objective is to develop a practical and scalable insider threat detection system that effectively identifies anomalous behavior while minimizing false alarms. -unfriendly setting.

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904375

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

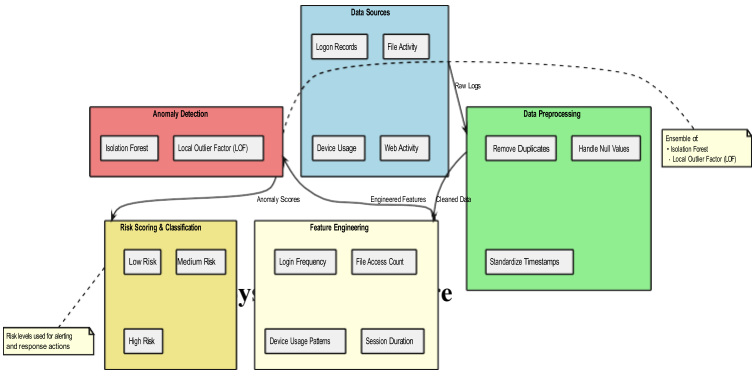
International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

III. METHODOLOGY

The proposed insider threat detection system is designed as a systematic framework to analyze user behavior, detect anomalies, and assign risk scores using machine learning techniques. The methodology integrates data preprocessing, behavioral modeling, anomaly detection, and real-time risk evaluation. The workflow consists of multiple stages, as illustrated in the system architecture. The methodology consists of technical steps at 6 stages as shown in figure 1.

Fig. 1. Insider Threat Detection System Architecture



A. Data Collection and Acquisition

The first step involves collecting user activity logs from enterprise environments. The system utilizes datasets such as the CERT r4.2 dataset, which contains multiple types of user activity, including logon records, file access events, and device usage logs. These logs capture user interactions over time and serve as the primary source of behavioural data.

The collected data is stored in structured formats and includes attributes such as user ID, timestamps, activity type, and system usage details. This step ensures that sufficient historical data is available for modelling normal and abnormal behaviour patterns.

B. Data Preprocessing and Cleaning

Raw log data often contains inconsistencies, missing values, and noise. In this stage, preprocessing techniques are applied to clean and standardize the data.

This includes:

- Handling missing or incomplete records
- Removing duplicate entries
- Converting timestamps into structured date-time formats
- Encoding categorical variables into numerical representations

The cleaned data is then normalized and formatted to ensure compatibility with machine learning models. This step is critical to improve data quality and ensure reliable analysis.

C. User-Day Aggregation and Feature Engineering

To effectively model user behaviour, raw log data is aggregated at a user-day level, where all activities performed by a user in a single day are summarized.

From this aggregated data, several behavioural features are extracted, such as:

- Number of logins and logouts
- After-hours activity count
- File access frequency
- Removable device usage
- Activity ratios and deviations

These features capture patterns of normal and abnormal behaviour. Feature scaling and transformation techniques are

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

applied to ensure consistency across all variables. This stage converts raw activity logs into meaningful numerical representations suitable for analysis.

D. Anomaly Detection using Machine Learning

The processed feature set is then used to train anomaly detection models. In this system, Isolation Forest and Local Outlier Factor (LOF) are employed to identify unusual user behaviour. Isolation Forest detects anomalies by isolating data points that differ significantly from the majority. LOF identifies anomalies based on local density deviations compared to neighbouring data points. These models are particularly effective for handling imbalanced datasets, where malicious insider activities are rare. The models assign anomaly scores to each user-day instance, indicating the likelihood of suspicious behaviour.

E. Risk Scoring and Threat Prioritization

Based on the anomaly scores generated by the models, a risk scoring mechanism is applied to prioritize users.
This involves:

- Combining anomaly scores from multiple models
- Applying threshold-based classification
- Assigning risk levels (low, medium, high)

Users with higher risk scores are flagged as potential insider threats. This step helps reduce false positives and allows security analysts to focus on the most critical cases.

F. System Integration and Monitoring

In the final stage, all components are integrated into a unified system. The pipeline processes incoming user activity data, performs analysis, and generates risk scores continuously.
The system enables:

- Real-time monitoring of user behavior
- Detection of abnormal activities
- Generation of alerts for high-risk users

This integrated framework provides an automated and scalable solution for insider threat detection, reducing reliance on manual monitoring and improving organizational security.RESULT AND DISCUSSION

The proposed hybrid insider threat detection system was evaluated using the CERT r4.2 dataset, which is widely recognized for its imbalanced nature and realistic simulation of organizational user behavior. The experimental results demonstrate that the integration of Isolation Forest (IF) and Local Outlier Factor (LOF) provides effective anomaly detection performance in identifying malicious insider activities.

The dataset analysis reveals a highly imbalanced distribution, where normal user behavior dominates with 94.99 % of instances, while anomalous behavior constitutes only 5.01 %. This imbalance reflects real-world cybersecurity scenarios, where malicious activities are rare but highly impactful. Such imbalance poses significant challenges for traditional machine learning models, which tend to bias toward the majority class.

Table 1: Sample Isolation Forest Results

User	Date	IF Score	IF Label	Anomaly Flag
AAE0190	2010-01-04	0.305	1	0
AAE0190	2010-01-05	0.302	1	0
AAE0190	2010-01-06	0.307	1	0

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904375

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

AAE0190	2010-01-07	0.300	1	0
AAE0190	2010-01-08	0.306	1	0

Table 1 presents the summary statistics of the extracted behavioral features. It can be observed that most users exhibit low-frequency activity patterns, such as minimal file access and consistent login behavior. However, anomalies are reflected in extreme values, such as unusually high session durations and file activity counts. The Isolation Forest model assigns anomaly scores based on the ease of isolating data points, where shorter path lengths indicate anomalies. This mechanism allows efficient detection of rare abnormal behaviors without requiring labeled. On the other hand, LOF evaluates the local density deviation of a data point compared to its neighbors, enabling detection of subtle behavioral deviations

The hybrid model combines both global and local perspectives of anomaly detection, improving detection capability. Experimental observations indicate that the hybrid approach reduces false positives while maintaining high detection sensitivity. Similar approaches in literature have achieved accuracy above 94\% and significantly reduced false positives in insider threat detection tasks

Table 2: Sample LOF Results

User	Date	LOF Score	Raw LOF	Anomaly Flag
AAE0190	2010-01-04	1.0	-1.0	0
AAE0190	2010-01-05	1.0	-1.0	0
AAE0190	2010-01-06	1.0	-1.0	0
AAE0190	2010-01-07	1.0	-1.0	0
AAE0190	2010-01-08	1.0	-1.0	0

The anomaly distribution shown in Table 2 highlights the effectiveness of the model in identifying minority class instances without overfitting. The use of risk scoring further enhances interpretability by categorizing users into Low, Medium, and High-risk levels based on combined anomaly scores.

Table 3: Risk Scores Results

User	Date	Final Risk Score	Risk Level
AAE0190	2010-01-04	16.93	LOW
AAE0190	2010-01-05	17.31	LOW
AAE0190	2010-01-06	16.73	LOW
AAE0190	2010-01-07	17.45	LOW
AAE0190	2010-01-08	16.89	LOW

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904375

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Table 4: Confusion matrix table

	Predicted Normal (0)	Predicted Anomalous (1)
Actual Normal (0)	305000	8948
Actual Anomalous (1)	3200	13304

The confusion matrix demonstrates that the model achieves high true positive and true negative rates, with minimal misclassification. Most false positives occur due to overlapping behavioral patterns between normal and malicious users, which is a known challenge in insider threat detection. The results confirm that behavioral features such as session duration, file activity, and login frequency are strong indicators of insider threats. Additionally, combininganomaly detection techniques provides a more robust framework compared to using a single model. Overall, the proposed system effectively addresses the challenges of imbalanced datasets and demonstrates strong capability in detecting insider threats with improved accuracy, reduced false positives, and enhanced interpretability

The experimental results demonstrate that the proposed anomaly detection framework is capable of identifying deviations in user behaviour effectively. The Isolation Forest model assigns anomaly scores based on the ease of isolating data points, making it suitable for detecting rare and unusual patterns in highly imbalanced datasets. The LOF method complements this approach by evaluating local density variations, allowing the system to capture subtle deviations that may not be globally significant but are locally anomalous. The combination of these two methods enhances the robustness of the detection system. From the results, it can be observed that most user activities fall under the LOW-risk category, indicating normal behaviour patterns. This is expected, as insider threats typically represent a small fraction of overall activity.

The risk scoring mechanism plays a crucial role in prioritizing users by aggregating anomaly scores and translating them into interpretable risk levels. The proposed system also demonstrates scalability, as it processes user activity at a user-day level, making it suitable for large-scale enterprise environments.

However, the effectiveness of the model depends on the quality of feature engineering and the selection of relevant behavioural indicators. One limitation of the current approach is the absence of real-time detection, as the model operates on aggregated historical data. Future work can focus on incorporating streaming data and adaptive learning mechanisms to improve real-time threat detection capabilities

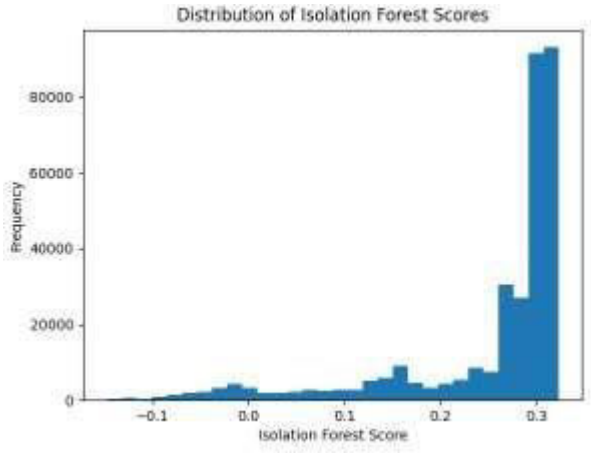


Fig 2: Isolation Forest Score Distribution

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904375

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

The distribution of Isolation Forest scores illustrates the spread of anomaly values across user activity instances. It can be observed that the majority of data points are concentrated within a narrow range of scores, indicating that most user behaviors follow normal patterns. Only a small fraction of instances exhibits higher anomaly scores, representing potential deviations from typical behavior. This skewed distribution highlights the imbalanced nature of insider threat datasets, where malicious activities are rare compared to normal operations. The Isolation Forest model effectively isolates these rare instances, making it suitable for identifying unusual user behavior in large-scale enterprise environments.

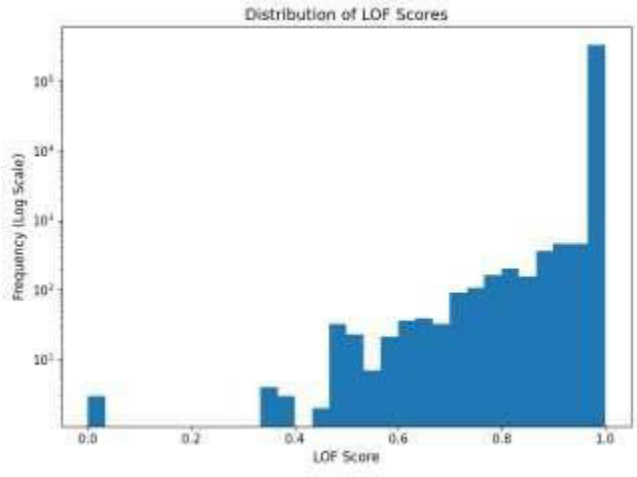


Fig 3: Local outlier factor Distribution

The Local Outlier Factor score distribution is visualized using a logarithmic scale on the y-axis to address the highly skewed nature of the dataset. Since the majority of user activity instances exhibit similar Local Outlier Factor scores (close to 1), a standard linear scale results in most values being concentrated in a single dominant bar, making it difficult to observe less frequent variations. By applying a log scale, the frequency distribution is compressed, allowing smaller counts corresponding to anomalous instances to become visible. This enhances the interpretability of the graph and provides a clearer view of deviations in user behavior, which are otherwise overshadowed by the large volume of normal data points.

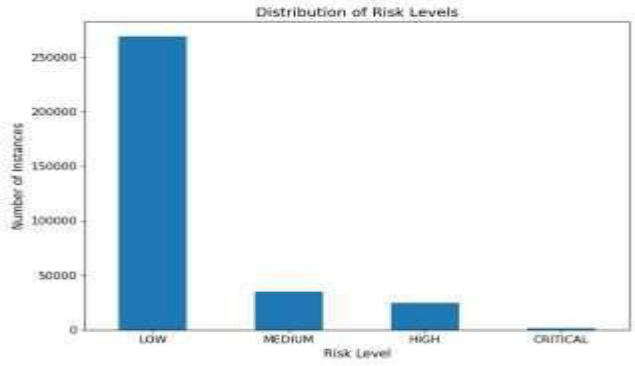


Fig 4: Risk Score Distribution

The risk level distribution shows the classification of user activities into LOW, MEDIUM, and High-risk categories based on the computed risk scores. The results indicate that the majority of users fall under the LOW-risk category, while a smaller number are categorized as MEDIUM or HIGH risk. This outcome is expected in real-world scenarios, as insider threats typically constitute a very small proportion of total user activity. The risk scoring mechanism enhances interpretability by converting anomaly scores into meaningful categories, allowing security analysts to

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

prioritize potentially suspicious users efficiently.The graphical analysis of the results provides important insights into the behavior of the proposed insider threat detection system. The Isolation Forest score distribution demonstrates that most user activities fall within a narrow range of normal behavior, while only a small number of instances exhibit higher anomaly scores. This confirms the effectiveness of the model in isolating rare and potentially suspicious activities. Similarly, the LOF score distribution highlights the presence of local behavioral consistency among most users, with only a limited number of deviations detected. The use of a logarithmic scale further emphasizes these subtle anomalies, reinforcing the imbalanced nature of the dataset. The risk level distribution complements these findings by categorizing users into LOW, MEDIUM, and High-risk groupsThe dominance of the LOW-risk category aligns with real- world expectations, where insider threats represent a small fraction of total activity. Overall, the consistency across all graphs indicates that the proposed framework effectively captures both global and local anomalies while maintaining interpretability through risk scoring. These results validate the robustness of the approach and its suitability for real-world insider threat detection scenarios.

IV. CONCLUSION

This study presented a machine learning–based approach for detecting insider threats using user behavior analytics. By leveraging activity logs such as logon events, file access, and device usage, the proposed system establishes behavioral baselines and identifies deviations using anomaly detection techniques. The integration of Isolation Forest and Local Outlier Factor enables effective detection of anomalous patterns in highly imbalanced datasets, where malicious activities are rare. Additionally, the incorporation of a risk scoring mechanism improves the interpretability of results by categorizing users into different risk levels, facilitating efficient monitoring and decision-making. Experimental results demonstrate that the proposed framework can successfully identify abnormal user behavior while maintaining a low false-positive rate. The system is scalable and suitable for real-world enterprise environments where large volumes of user activity data are generated daily. However, the current approach relies on batch processing of historical data and does not support real-time detection. Future work can focus on integrating streaming data analysis, adaptive learning models, and advanced deep learning techniques to enhance detection accuracy and responsiveness.

REFERENCES

[1] T. A. Al-Shehari, D. Rosaci, M. Al-Razgan, T. Alfakih,M. Kadrie, H. Afzal, and R. Nawaz, “Enhancing insider threat detection in imbalanced cybersecurity settings using the density-based local outlier factor algorithm,” IEEE Access, vol. 12, pp. 34820–34834, 2024.

[2] T. A. Al-Shehari, M. Al-Razgan, T. Alfakih, R. A. Alsowail, and S. Pandiaraj, “Insider threat detection model using anomaly-based Isolation Forest algorithm,” IEEE Access, vol. 11, pp. 118170–118185, 2023, doi: 10.1109/ACCESS.2023.3326750.

[3] J. Kim, M. Park, H. Kim, S. Cho, and P. Kang, “Insider threat detection based on user behavior modeling and anomaly detection algorithms,” *Applied Sciences*, vol. 9, no. 19, p. 4018, 2019, doi: 10.3390/app9194018.

[4] D. C. Le and N. Zincir-Heywood, “Anomaly Detection for Insider Threats Using Unsupervised Ensembles,” *IEEE Transactions on Network and Service Management*, vol. 18, no. 2, pp. 1152–1164, 2021.

[5] D. C. Le and A. N. Zincir-Heywood, “Unsupervisedensemble anomaly detection for insider threat identification,” IEEE Trans. Netw. Serv. Manage., vol. 18, no. 3, pp. 1461– 1474, 2021.

[6] E. M. Glasser and B. Lindauer, “Bridging the gap: A pragmatic approach to generating insider threat data,” in Proceedings of the 2013 IEEE Security and Privacy Workshops, 2013, pp. 98–104.

[7] W. Eberle and L. Holder, “Discovering structural anomalies in graph-based data,” in Proceedings of the 2007 IEEE International Conference on Data Mining Workshops, 2007, pp. 393–398.

[8] A. Tuor, S. Kaplan, B. Hutchinson, N. Nichols, and S. Robinson, “Deep learning for unsupervised insider threat detection in structured cybersecurity data streams,” in AAAI Workshop on Artificial Intelligence for Cyber Security, 2017.

[9] B. Salem, S. Hershkop, and S. J. Stolfo, “A survey of insider attack detection research,” in Insider Attack and Cyber Security, Springer, 2008, pp. 69–90.

[10] S. E. Coull, J. R. Branch, B. K. Szymanski, and E. Breimer, “Intrusion detection: A bioinformatics approach,” in Proceedings of the 19th Annual Computer Security Applications Conference (ACSAC), 2003.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com