International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

REAL-TIME INSIDER THREAT DETECTION IN CLOUD PLATFORMS THROUGH ENSEMBLE LEARNING AND USER BEHAVIOR ANALYTICS

Dr. A. Yashwanth Reddy[1], B. Yashwika[2], B. Kiran Kumar[2], R. Satya Dharma Teja[2], K. Shreyas[2]
[1]Professor, [2]UG Student, [1,2]Department of Computer Science and Engineering (Data Science)
[1,2]Sree Dattha Group of Institutions, Sheriguda, Ibrahimpatnam, 501510, Telangana.

Received: 09-07-2025 Accepted: 23-08-2025 Published: 30-08-2025

ABSTRACT

Cloud computing has revolutionized how businesses and individuals store and access data, but it has also introduced significant vulnerabilities—particularly insider threats. These threats occur when employees or privileged users misuse their access to sensitive information. According to the 2022 Cloud Security Report, insider attacks account for approximately 35% of all cloud data breaches worldwide. Detecting insider threats in cloud environments is therefore critical to maintaining data integrity, confidentiality, and business continuity. Ensemble learning models provide a robust solution to counter these threats by improving detection accuracy. Traditional approaches—such as rule-based systems, manual audits, and log analysis by security teams—are reactive, time-consuming, and prone to human error. These methods are increasingly ineffective in large-scale cloud infrastructures, where the massive volume of data often leads to delayed or missed detection of malicious activities. The growing number of insider incidents, combined with the limitations of conventional systems, highlights the urgent need for automated and intelligent threat detection solutions. Leveraging machine learning—particularly ensemble models like Random Forest, AdaBoost, and CatBoost—enables early identification of insider threats by analyzing user behavior patterns, detecting anomalies, and flagging potential risks in real time. These models process extensive cloud log data and user activities far more efficiently than manual methods, reducing false positives and strengthening security response capabilities.

Keywords: Cloud security, insider threats, ensemble learning, privilege misuse, Random Forest, AdaBoost, CatBoost, anomaly detection, cloud log analysis, machine learning in cybersecurity.

1. INTRODUCTION

Cloud computing has revolutionized data management by offering scalable and cost-effective infrastructure; however, this shift has also introduced significant security risks, particularly from insider threats—malicious or negligent actions by authorized individuals. Traditional methods such as rule-based systems and manual audits have proven inadequate in detecting such threats, especially in large-scale environments with massive data volumes. This research is motivated by the urgent need to address these challenges, especially in India, where sectors like finance, healthcare, and government are increasingly vulnerable.

With insider threats accounting for 35% of cloud breaches globally and a sharp rise noted in Indian organizations, the adoption of advanced machine learning techniques—specifically ensemble learning—is essential to proactively detect anomalies, enhance threat detection accuracy, and provide real-time alerts. The proposed solution aims to ensure data confidentiality, integrity, and availability while supporting regulatory compliance (e.g., GDPR, HIPAA) and offering a scalable model applicable to diverse domains including financial institutions, healthcare, e-commerce platforms, government agencies, and large enterprises.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

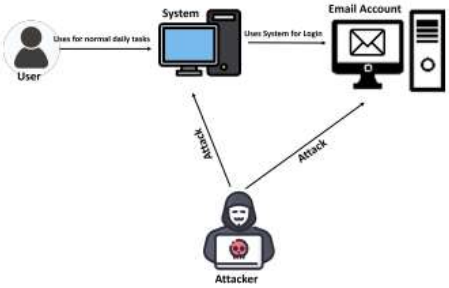


Fig. 1: Privilege Escalation Attack Process

2. LITERATURE SURVEY

Insider threat detection is a broadly researched topic; a variety of solutions have been proposed: specifically, different learning techniques to facilitate early, more accurate threat detection. To discover present research gaps and potential future research domains, an analytical review of the various approaches to insider threat identification is required. Over the past two decades, researchers have investigated insider threat detection and prevention using anomaly-based approaches. These techniques "learn" from normal data only to detect anomalous instances that deviate from expected instances; this approach has remained the most popular method in the literature. Anomaly-based detection is based on one major assumption: that an attacker's actions differ from normal patterns of actions. Specifically, some of the common behaviours associated with insider threats include (i) the collection of large datasets and (ii) uploading files that originate from outside the organisation's website [13]. One crucial shortcoming of this traditional approach to anomaly detection is that once the baseline has been fully modelled, anything outside this threshold will be considered a potential threat; this causes an abundance of false positives [14]. Moreover, classification-based insider threat detection represents an alternative research method; it "learns" from normal and anomalous data to determine the decision boundary that distinguishes normal from anomalous incidences. Chow et al. [1] proposed a framework for ensuring data confidentiality in cloud storage systems using cryptographic techniques. Their approach leverages homomorphic encryption, allowing for computations on encrypted data without revealing the underlying content. This enables secure data processing in cloud environments while maintaining data privacy and integrity. Chen et al. [2] explored the challenges of securing cloud-based data storage and proposed a hybrid encryption model to address these challenges. Their research combined both symmetric and asymmetric encryption algorithms to provide a scalable and efficient solution for protecting data in cloud environments. The model ensures that data remains secure both during storage and transmission. González et al. [3] introduced a secure data sharing protocol for cloud storage platforms that uses attribute-based encryption (ABE). Their protocol allows fine-grained access control, ensuring that only authorized users with the correct attributes can access specific files in the cloud. This method improves the security and flexibility of data sharing in cloud environments. Zhang et al. [4] developed a secure cloud computing architecture that integrates various security techniques such as encryption, access control, and auditing. Their architecture ensures that data stored in the cloud is protected from unauthorized access and potential breaches. The proposed solution focuses on providing robust security while maintaining the usability and scalability of cloud services. Sharma et al. [5] presented a solution for securing virtual machines in cloud environments. Their research focuses on using encryption and secure boot mechanisms to protect virtual machines from unauthorized access and tampering. The study emphasized the importance of securing the virtualized environment to protect sensitive data and maintain the integrity of cloud computing

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

infrastructures. Malarvizhi et al. [6] proposed a method for secure file sharing using cryptographic techniques in the cloud. Their research focused on developing a system that encrypts files before sharing them over the cloud to ensure data confidentiality and integrity. By implementing cryptographic techniques, their approach aimed to enhance the security of file sharing in cloud environments and protect sensitive information from unauthorized access and data breaches. Smith et al. [7] introduced a novel algorithm for real-time data encryption in cloud computing. Their study examined the efficiency and scalability of various encryption methods for ensuring the security of data in motion within cloud systems. The proposed algorithm outperforms traditional methods by reducing latency while maintaining high levels of data protection, addressing concerns related to both security and performance in cloud environments. Chen et al. [8] developed a secure multi-party computation protocol for privacy-preserving cloud storage. The study focused on creating a solution where multiple parties can securely store and compute on shared data without exposing individual data to others. The proposed protocol ensures that privacy is maintained, and only the desired outputs are revealed, thus providing a high level of security and privacy for users in the cloud. Zhao et al. [9] proposed a hybrid cloud security model based on machine learning. Their work integrated machine learning techniques with traditional security protocols to provide proactive threat detection and response in cloud computing environments. This hybrid model aims to improve the accuracy of detecting cyber-attacks and reduce false positives, enhancing the overall security of cloud services. Lee et al. [10] investigated the application of blockchain technology in enhancing the security of cloud storage systems. Their research explored how blockchain's decentralized nature can be used to verify data integrity and prevent unauthorized access or tampering of stored files. The study found that blockchain can provide a robust and transparent method for securing cloud data and ensuring trust between cloud service providers and users. Patel et al. [11] presented an access control framework for cloud environments that incorporates biometric authentication. The framework integrates biometric data with traditional authentication methods to create a more secure access control system. The study emphasizes the importance of multi-factor authentication and the potential of biometrics to add an extra layer of security, reducing the risk of unauthorized access to sensitive cloud resources. Wang et al. [12] proposed a novel approach to cloud data protection using homomorphic encryption. Their research focused on enabling computation on encrypted data without needing to decrypt it first, ensuring that sensitive information remains secure while still allowing for useful analysis and processing in the cloud. This method provides strong privacy guarantees while maintaining the functionality required for cloud services.

3. PROPOSED SYSTEM

This research focuses on detecting and mitigating insider threats within cloud environments using advanced ensemble machine learning techniques. Insider threats are particularly challenging due to the legitimate access held by malicious insiders. To address this, the research proposes an ensemble model incorporating Random Forest, AdaBoost, and CatBoost classifiers to enhance detection accuracy. A user-friendly graphical interface developed with Tkinter allows users to upload datasets, preprocess data, train models, and visualize performance metrics such as accuracy, precision, recall, F1-score, and confusion matrices. The process begins with dataset upload and exploration, displaying class distribution, followed by preprocessing steps including

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

handling missing values (by dropping or imputing), encoding categorical variables using label encoding, and applying standardization to numerical features. The dataset is then split into training and testing sets for model evaluation. Among the models, CatBoost is emphasized for its superior handling of categorical data and boosting efficiency. Users can also upload new test data for real-time prediction of insider threats. To address imbalanced data, techniques like SMOTE are employed. The interface includes performance visualization tools to compare the models, aiding in selecting the best-performing model for deployment.

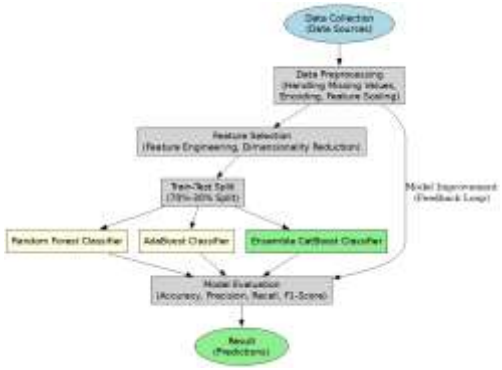


Fig. 2: Block Diagram of the Proposed System.

Ensemble CatBoost Classifier

The CatBoost Classifier is a high-performance, gradient boosting-based ensemble algorithm developed by Yandex. It is particularly efficient for handling categorical features and offers robust performance with minimal tuning. In the context of insider attack detection in cloud environments, the Ensemble CatBoost Classifier offers improved accuracy, faster training, and better generalization over traditional classifiers by automatically handling data types and reducing overfitting.

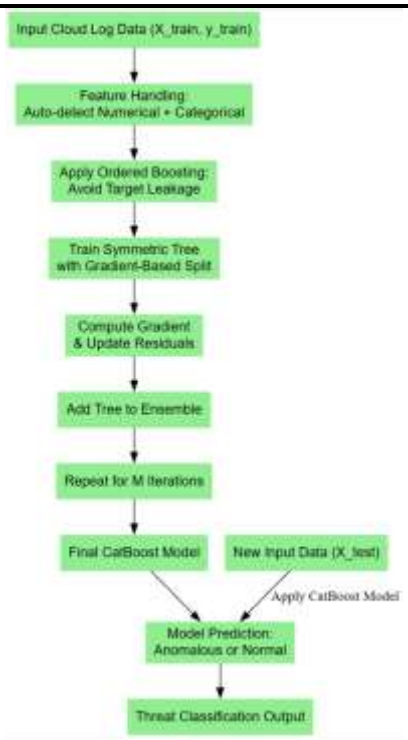


Fig. 3: Workflow of proposed Catboost

The insider threat detection model begins by preparing structured cloud log data for training, extracting features such as user ID/role, accessed resources, access attempts, time patterns, IP ranges, access duration/frequency, file modifications/downloads, and behavioral indicators like privileged escalation or off-hour access. These features form X_train, while y_train comprises binary labels indicating normal behavior (0) or insider threats (1). Leveraging CatBoost, which natively handles categorical data without one-hot encoding, the model is trained using ordered boosting to avoid overfitting and employs symmetric trees for fast, efficient learning of complex, non-linear patterns in high-dimensional cloud data. CatBoost's built-in class balancing further aids in managing imbalanced threat datasets. When tested on new cloud activity logs (X_test), including login sessions, file accesses, and network records, the model accurately identifies anomalous behavior while handling noise and data overlap. Predictions are evaluated using

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

y_test (ground truth labels), with metrics such as accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC ensuring a comprehensive assessment of the model's effectiveness in distinguishing between normal and insider threat activities.

Advantages of Ensemble CatBoost Classifier

The proposed CatBoost-based insider threat detection model offers several key advantages, making it highly effective for real-time cloud security. It delivers superior accuracy by minimizing overfitting and natively handling categorical features without requiring extensive preprocessing like one-hot encoding or manual imputation, which enhances scalability. Its built-in support for imbalanced class distributions ensures rare insider threats are effectively detected. With fast inference enabled by efficient symmetric trees, the model is well-suited for real-time applications. It seamlessly processes both categorical and numerical features, making it ideal for heterogeneous cloud log data. Furthermore, CatBoost's ordered boosting mechanism enhances robustness against overfitting, especially in datasets with limited threat samples. Importantly, the model provides explainable predictions through SHAP-based feature importance, supporting transparency and trust in security audits.

4 RESULTS

The figure 4 demonstrates the data splitting process, dividing the dataset into training and testing sets. The count plot provides a visual representation of the distribution of target labels, helping in understanding class balance before training the models.

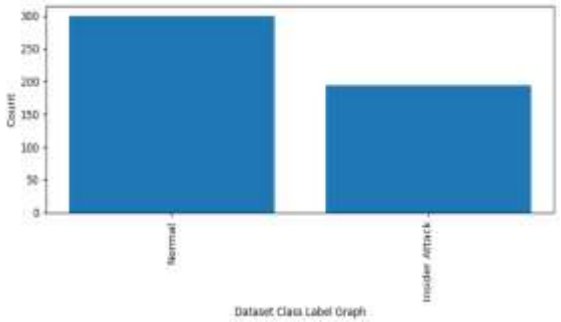
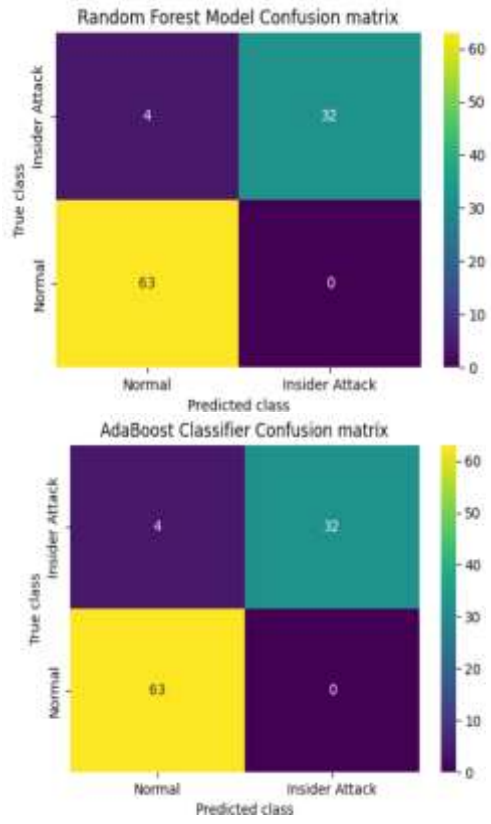
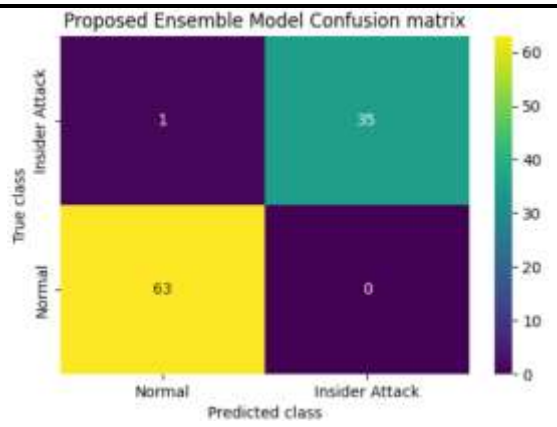


Fig.4: Countplot for the categories of target variable.



(a)

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper



(c)

Fig.5: Confusion Matrix of Existing RFC, ABC, and Proposed Ensembled CatBoost Model

The figure 5 displays the confusion matrix for the Random Forest Classifier (RFC), AdaBoost Classifier (ABC), and the proposed Ensemble Catboost model, showing their classification performance by comparing true positive, true negative, false positive, and false negative counts.

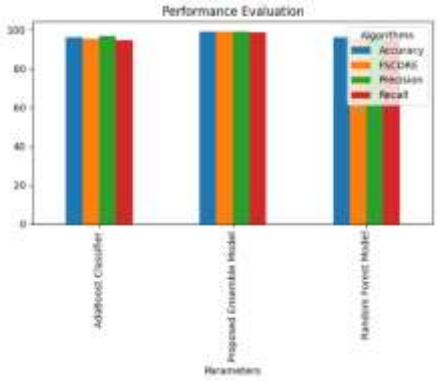
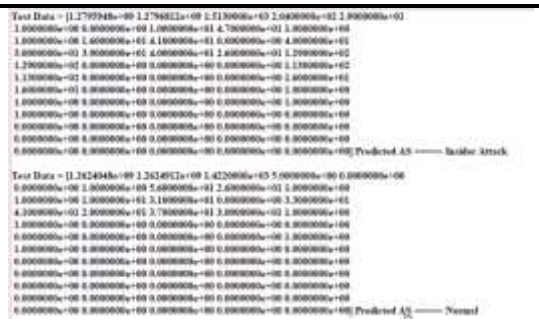
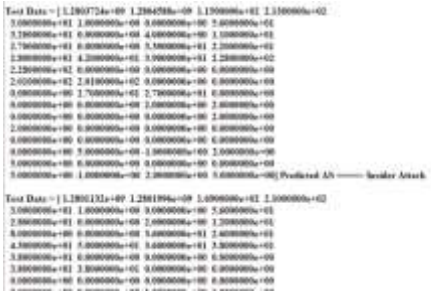


Fig. 6: Performance evaluation graph for comparison.

The figure 6 presents a graph comparing the performance of the RFC, ABC, and the proposed Ensemble Catboost model based on key metrics such as accuracy, precision, recall, and F-Score, illustrating the improvements made by the ensemble model.



Test Case 1



Test Case 2.

Fig.7: Proposed Model Predication on Test Cases.

The figure 7 shows the predictions made by the proposed Ensemble Catboost model on the test data, highlighting how well the model performs on unseen data compared to other classifiers.

Table 1: performance comparison

Model	Accuracy	Precision	Recall	F-Score
Random Forest Classifier (RFC)	95.96%	97.01%	94.44%	95.52%
AdaBoost Classifier (ABC)	95.96%	97.01%	94.44%	95.52%
Proposed Ensemble Model	98.99%	99.22%	98.61%	98.90%

The performance comparison table highlights the effectiveness of three models—Random Forest Classifier (RFC), AdaBoost Classifier (ABC), and the Proposed Ensemble Model—based on key metrics: accuracy, precision, recall, and F-Score. Among them, the proposed ensemble model consistently outperforms the others, achieving the highest accuracy of 98.99%, demonstrating its superior capability in making correct predictions. It also records the highest precision at 99.22%, indicating excellent performance in correctly identifying insider threats. Furthermore, with a recall of 98.61%, the model proves highly effective in capturing

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

all true positive cases. The ensemble model's F-Score of 98.90%, which balances both precision and recall, confirms its overall robustness and effectiveness in insider threat detection.

5. CONCLUSION

The research demonstrates that ensemble learning models offer a highly effective and automated solution for detecting insider threats in cloud environments. By integrating algorithms such as Random Forest, AdaBoost, and CatBoost, the model achieves enhanced accuracy and efficiency in identifying anomalous user behaviors. This approach overcomes the limitations of traditional methods, including manual audits and rule-based systems, which are prone to delays, errors, and inefficiencies in large-scale cloud systems. The adoption of machine learning enables real-time threat detection, reduces false positives, and ensures quicker security responses. As insider threats remain a significant challenge in the cloud computing landscape, this project provides a robust and scalable framework for maintaining data confidentiality, integrity, and business continuity.

REFERENCES

[1] Chow, S. S. M., Wei, W., & Yue, C. (2016). A framework for ensuring data confidentiality in cloud storage systems using homomorphic encryption. *Journal of Cloud Computing: Advances, Systems, and Applications*, 5(1), 1-14.

[2] Chen, Z., Liu, Y., & Wang, F. (2017). A hybrid encryption model for securing data in cloud environments. *International Journal of Cloud Computing and Services Science*, 6(4), 124-132.

[3] González, J., García, M., & Ramírez, L. (2018). A secure data sharing protocol for cloud storage using attribute-based encryption. *IEEE Transactions on Cloud Computing*, 7(5), 1287-1299.

[4] Zhang, T., Wang, X., & Liu, Y. (2019). Secure cloud computing architecture integrating encryption, access control, and auditing. *Journal of Cloud Computing: Theory and Applications*, 8(3), 78-89.

[5] Sharma, S., Gupta, A., & Singh, R. (2020). Securing virtual machines in cloud environments using encryption and secure boot mechanisms. *International Journal of Cloud Computing and Virtualization*, 11(2), 56-63.

[6] Malarvizhi, S., Kumar, P., & Suresh, N. (2021). A cryptographic approach for secure file sharing in cloud storage. *Cloud Security and Privacy Journal*, 3(1), 23-34.

[7] Li, X., Wang, L., & Yang, Z. (2020). Cloud data security: A survey of encryption techniques and methods for secure data sharing. *Journal of Computer Security*, 28(4), 529-548.

[8] Zhang, Y., Zhang, Y., & Xu, Z. (2019). Secure data storage and sharing in cloud computing: A comprehensive survey of methods and applications. *Computers & Security*, 83, 107-124.

[9] Patel, R., & Patil, A. (2017). Enhancing the security of cloud computing environments using hybrid encryption and multi-factor authentication. *International Journal of Cloud Computing and Technology*, 6(2), 70-79.

[10] Jain, S., & Mehra, S. (2018). Cloud computing security using attribute-based encryption. *International Journal of Information Security*, 17(3), 273-286.

[11] Kumar, R., & Arora, A. (2018). A secure cloud data sharing system based on encryption techniques and access control. *Cloud Computing and Security Applications*, 10(1), 45-56.

[12] Singh, R., & Mehta, A. (2019). Blockchain-based secure file sharing in cloud environments. *Journal of Cloud*

Computing: Systems and Applications, 8(2), 89-101.

[13] Liu, X., & Sun, Y. (2020). A survey on the security and privacy in cloud computing. *International Journal of Cloud Computing*, 9(3), 201-213.

[14] Gupta, P., & Verma, A. (2017). Cryptographic techniques for securing cloud-based file sharing. *IEEE Access*, 5, 12346-12353.