Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

A Comprehensive Survey on Anomaly Detection Techniques for Suspicious File Migration in Cloud Computing Environments

Ayush Gupta, Ayush Garg, Atul Saini, Anurag Tyagi, Priyanka
Department of Computer Science and Engineering
Meerut Institute of Engineering and Technology
Meerut, India

***Abstract*—Cloud storage platforms, particularly Amazon Simple Storage Service (S3), are widely adopted by organizations globally due to their inherent flexibility, massive scalability, and cost-effectiveness. However, the paradigm shift toward distributed cloud infrastructure has exponentially increased the risk of unauthorized or abnormal file transfers. Typical file m igrations w ithin t hese d istributed environments have escalated security challenges, where illegal file transfers serve as primary indicators of insider abuse, credential compromise, or systematic data exfiltration. Manually identifying such suspicious migrations is virtually impossible due to the vast volume of daily activity logs. This comprehensive survey investigates state-of-the-art anomaly detection techniques designed to identify suspicious file m igration patterns. We critically evaluate traditional rule-based monitoring, deep learning methods, and blockchain-based auditing models. Special emphasis is placed on lightweight, unsupervised machine learning architectures—specifically t he I solation Forest algorithm—which autonomously identify deviations in user activity, file a ccess p atterns, a nd d ata m ovement r ates by ingesting AWS CloudTrail logs. Furthermore, we explore the end-to-end integration of these models with real-time visualization dashboards built using Flask, enabling automated feature extraction, model serialization via joblib, and rapid threat mitigation by Security Operations Centers (SOC). By analyzing the accuracy, scalability, and operational efficiency of existing frameworks, this survey identifies p ersistent systemic flaws, s uch a s h igh f alse-positive r ates a nd t he l ack o f real-time multi-cloud monitoring. Ultimately, this study proposes dynamic, automated solutions that seamlessly combine cloud computing telemetry, machine learning, and cybersecurity principles to protect mission-critical digital assets across the finance, healthcare, and educational sectors.**

***Keywords*—Anomaly Detection, AWS CloudTrail, Blockchain Auditing, Cloud Security, Flask Dashboard, Isolation Forest, Machine Learning, Suspicious File Migration.**

I. INTRODUCTION

A. The Paradigm Shift to Cloud Storage

The explosive growth and rapid adoption of cloud computing have completely transformed the manner in which modern organizations handle, store, share, and protect their data. With the persistent trend of enterprises shifting their on-premises data centers to cloud-based infrastructures to attain massive scalability, operational efficiency, and significant cost minimization, the dependency on cloud storage solutions has never been greater. Platforms such as Amazon Web Services (AWS) Simple Storage Service (S3) form the absolute backbone of modern digital infrastructure, hosting everything from routine operational documents to highly sensitive financial and medical records.

As developing nations and enterprises continue to heavily digitize educational, cultural, and corporate knowledge, ensuring the absolute integrity of digital property has evolved into a compulsory legal and operational requirement. Regardless of the immense benefits offered by cloud architectures, the ubiquitous use of distributed cloud environments simultaneously creates expansive new attack surfaces. In cases where sensitive data is distributed across a variety of geographical regions and multi-tenant environments, maintaining strict visibility over data provenance and movement becomes exceedingly difficult.

B. The Threat of Suspicious File Migration

One of the most critical, yet frequently overlooked, security issues in modern cloud ecosystems is *suspicious file migration*. This is defined as the unexpected, unauthorized, or anomalous transfer, replication, downloading, or movement of files across storage buckets, user accounts, availability zones, or geographical borders. These exceptions are rarely accidental; they are frequently primary indicators of malicious intent, such as insider abuse by disgruntled employees, credential compromise by external advanced persistent threats (APTs), systematic data exfiltration operations, or severe architectural misconfigurations.

Traditional security architectures rely heavily on static rule-based alerts and signature matching. However, as the volume of cloud interactions grows exponentially—often reaching millions of logged events per day in an enterprise setting—these rule-based systems become highly inefficient. They are easily bypassed by opponents utilizing sophisticated evasion techniques, such as ”low and slow” data exfiltration, where files are migrated in minuscule increments over extended periods to avoid triggering volume-based alarms. Furthermore, traditional systems frequently lead to inaccurate warnings, slower threat detection times, and poor overall security management due to overwhelming alert fatigue among security administrators.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

C. Objectives and Organization of the Survey

To address these critical limitations, this survey explores the transition from static rule-based security to dynamic, machine learning-driven anomaly detection frameworks. The primary objective is to review systems that automatically decompose and analyze AWS CloudTrail logs to formulate normal operational baselines and flag abnormal trends in file migration operations. By focusing on unsupervised learning methods like the Isolation Forest algorithm, modern frameworks can automatically learn normal behavioral boundaries without relying on predefined threat signatures.

The remainder of this paper is organized as follows: Section II discusses the background and evolution of cloud security monitoring. Section III outlines the threat landscape specific to file migration. Section IV presents our systematic literature review and taxonomy of existing research. Section V delves into machine learning techniques for anomaly detection, detailing the Isolation Forest. Section VI outlines the end-to-end proposed system architecture based on AWS and Flask. Section VII explores feature engineering from cloud logs. Section VIII evaluates visualization and alerting mechanisms. Section IX analyzes comparative performance metrics. Section X discusses open challenges, and Section XI concludes the survey.

II. BACKGROUND AND EVOLUTION OF CLOUD SECURITY MONITORING

A. From Network Security to File-Level Auditing

Historically, cloud security posture management (CSPM) focused heavily on network perimeter defense and Identity and Access Management (IAM). Systems were designed to prevent unauthorized users from gaining access to the cloud console. However, once a user—whether legitimate or a compromised insider—bypassed the IAM perimeter, there was very little visibility into what they actually did with the data.

Logging systems like AWS CloudTrail were introduced to provide a comprehensive audit trail of every API call made within a cloud environment. While CloudTrail provides the raw JSON data necessary for forensic analysis, manually parsing these logs to establish malicious file movements is a computationally and cognitively impossible task for human administrators. The literature available historically focused on network anomalies (e.g., DDoS attacks) or authentication anomalies (e.g., impossible travel logins), but largely neglected the nuanced behaviors of intra-cloud file migration.

B. The Need for Behavioral Analytics

The research gap in the current cybersecurity defense matrix is the development of lightweight, adaptive systems applicable to analyzing behavioral deviations on-the-fly. If an employee in the marketing department, who typically accesses small PDF documents during standard business hours, suddenly begins duplicating gigabytes of proprietary source code to an unencrypted, public-facing S3 bucket in a different geographical region at 3:00 AM, the system must immediately flag this as a high-risk anomaly. Achieving this requires transitioning from rigid policies to unsupervised behavioral analytics powered by machine learning.

III. THREAT LANDSCAPE IN DISTRIBUTED ENVIRONMENTS

A. Insider Threats and Credential Compromise

The most persistent threat vector involving file migration is the insider threat. Studies emphasize that insider threat detection requires sophisticated behavioral analytics because the user possesses legitimate access rights, rendering traditional perimeter defenses useless. An insider might duplicate sensitive intellectual property to a personal cloud repository prior to terminating their employment. Because the action (e.g., `CopyObject`) is fundamentally a legitimate API call, signature-based antivirus or firewall systems will not block it.

B. Geographic and Regulatory Anomalies

The secure cloud storage paradigm also dictates an important feature of location-based data protection. Frameworks such as SecLoc highlight the necessity of maintaining strict control over location-sensitive files. For example, under the General Data Protection Regulation (GDPR) or India's Digital Personal Data Protection (DPDP) Act, migrating citizen data to a server located outside the approved jurisdiction without proper safeguards is a severe compliance violation. Suspicious file migration detection systems must therefore account for the geolocation data of both the user initiating the API call and the destination storage bucket to ensure regulatory compliance.

IV. SYSTEMATIC LITERATURE REVIEW

To establish the context of current advancements, a comprehensive literature review was conducted, focusing on cloud security, machine learning anomaly detection, and distributed file systems.

A. Categorization of Existing Research

Research on the methods of attaining security in cloud-based file services has been rising rapidly because of the accessibility of massive remote storage environments. The reviewed literature can be broadly categorized into three distinct themes:

1) **Supervised and Unsupervised Machine Learning:** Studies focusing on isolating threats without relying on static signatures.
2) **Blockchain and Immutable Auditing:** Research detailing the use of decentralized ledgers to ensure that migration logs cannot be tampered with by malicious actors.
3) **Hybrid Policy Frameworks:** Systems that combine rigid compliance rules with flexible ML models.

As demonstrated in Table I, while cloud security has developed at an advanced level, there has been minimal emphasis on the holistic, real-time, ML-supported detection of suspicious file migration. The fusion of both information-based and intuitive patterns has increased the precision of detection, particularly through hybrid methods. However, a clear gap remains for more adaptable, lightweight, and scalable solutions that incorporate dynamic web dashboards.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

TABLE I
TAXONOMY OF KEY LITERATURE IN CLOUD FILE MIGRATION AND ANOMALY DETECTION

Authors & Year	Proposed Methodology/Technology	Core Focus / Limitations
Bowers et al. (2021) [2]	Supervised learning model for detecting suspicious file migration and replication.	Laid groundwork for multi-tenant cloud auditing; limited by the need for massive labeled training datasets.
Sarumathi (2023) [1]	Analyzed abnormal migration patterns to prevent covert copying.	Emphasized tracking necessity but lacked a real-time visualization framework.
Xu et al. (2022) [6]	Deep learning models for anomaly detection in cloud file systems.	Highly accurate but computationally expensive for real-time edge processing.
Chen et al. (2021) [7]	Cloud file migration monitoring using Blockchain ledgers.	Provided tamper-resistant logs but suffered from low transaction throughput.
Sharma & Nair (2020) [8]	Insider threat detection utilizing User and Entity Behavior Analytics (UEBA).	Focused heavily on user login times rather than specific object-level file movements.
Kumar & Singh (2023) [10]	Hybrid detection utilizing both static policy rules and machine learning.	Reduced false positives significantly by enforcing strict compliance guardrails over ML outputs.
Nie et al. (2025) [5]	Zero Copy: File System Assisted Container Buffer Migration.	Focused on performance and system-level file transfer speeds rather than security analytics.

V. MACHINE LEARNING TECHNIQUES FOR ANOMALY DETECTION

To address the shortcomings of rule-based systems, researchers have pivoted toward unsupervised machine learning using the `scikit-learn` library. Because "malicious" file migrations are exceedingly rare compared to standard daily operations, the dataset is heavily imbalanced. Unsupervised models excel in this environment by learning the parameters of "normal" behavior and isolating data points that fall outside these boundaries.

A. The Isolation Forest Algorithm

The most prominent algorithmic approach identified for this specific problem domain is the **Isolation Forest**. Unlike traditional algorithms that attempt to model normal data profiles (such as One-Class SVMs), the Isolation Forest explicitly and directly isolates anomalies.

The core principle is that anomalies are data points that are "few and different." The algorithm builds an ensemble of Random Trees (Isolation Trees) for a given dataset. Anomalies are more susceptible to isolation and therefore have significantly shorter path lengths from the root node to the terminating leaf node.

Mathematically, the anomaly score $s(x,n)$ for a data point x given a sample size n is defined as:

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}} \quad (1)$$

Where:

- $h(x)$ is the path length of observation x.
- $E(h(x))$ is the average path length of x across a forest of isolation trees.
- $c(n)$ is the average path length of an unsuccessful search in a Binary Search Tree given n nodes, used to normalize the score.

When $E(h(x)) \to c(n)$, the score $s \to 0.5$, indicating a normal instance. When $E(h(x)) \to 0$, the score $s \to 1$, indicating a definitive anomaly.

In modern cloud implementations, the Isolation Forest algorithm is typically trained using approximately 200 estimators (trees) and a contamination parameter of 0.05 (assuming a maximum of 5% of the cloud log data may be anomalous). This setup is highly sensitive and stable, effectively minimizing false positives without impacting the accuracy of the detection.

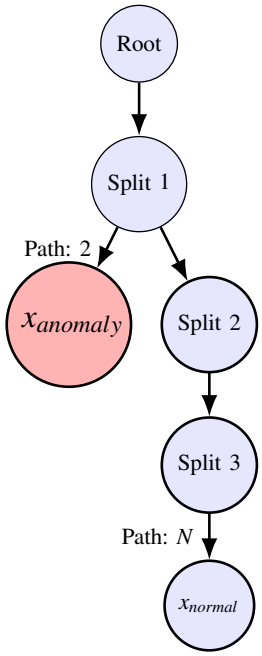

Fig. 1. Conceptual representation of an Isolation Forest Tree. Anomalies ($x_{anomaly}$) are isolated much faster (shorter path length) than normal data points (x_{normal}).

VI. PROPOSED SYSTEM ARCHITECTURE

To deploy these machine learning models effectively, organizations must establish a seamless, automated data pipeline. The proposed system architecture establishes a single, continuous pipeline that transforms raw AWS activity logs into actionable security intelligence displayed on a web interface.

A. Data Acquisition and AWS CloudTrail

The automation of the AWS CloudTrail system is designed to develop a hands-free and continuous system that logs and labels all activity registered in the cloud environment. Using the `Boto3` AWS SDK for Python, the system extracts logs at a regular frequency. This ensures that

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

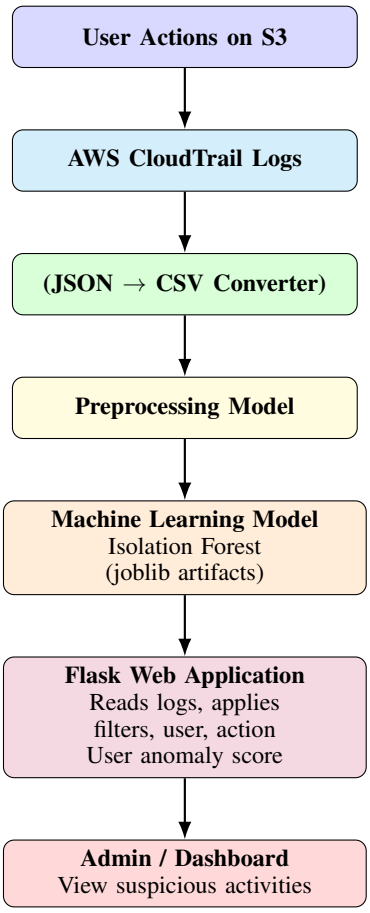

Fig. 2. Proposed System Architecture mapping the flow of telemetry from AWS S3 user actions to the final Administrator Dashboard.

any file-related operations such as uploads (`PutObject`), downloads (`GetObject`), copying (`CopyObject`), or permission changes are logged practically in real-time. Regular polling facilitates a regular flow of data, removing manual interference and rendering the process faster and highly less susceptible to errors.

B. Data Parsing and Preprocessing

Raw CloudTrail records are generated in heavily nested JSON format. To make them usable for downstream analytics, a preprocessing module utilizes the `Pandas` library to cleanse and normalize the data into structured CSV tables. This involves the flattening o f n e sted fi el ds, de le tion of redundancy in records, transformation of timestamps, and alignment of attribute types. This stage assists in ensuring the consistency and removal of noise, making the dataset readable and readily available to the machine learning engine.

VII. FEATURE ENGINEERING AND BEHAVIORAL ANALYTICS

The success of the Isolation Forest model relies entirely on the quality of the extracted features. Raw logs must be meticulously translated into behavioral indicators.

A. Primary and Derived Attributes

Central features extracted directly from the logs include:

- **User Identity (ARN):** Identifies the exact IAM role or user who initiated the transfer.
- **Client IP Address:** Determines if the API call is originating from a recognized corporate subnet or a suspicious foreign IP.
- **Event Name/Action Type:** Distinguishes between benign read actions and high-risk mass copy actions.
- **Resource Size:** The byte size of the file being migrated.

To provide the model with deeper context, *derived attributes* must be engineered:

- **Access Frequency (Velocity):** The number of file operations executed by a specific user within a specific time window.
- **Temporal Anomalies:** Migrations occurring completely outside of normal working hours.
- **Geospatial Deltas:** Calculating the physical distance between the user's IP geolocation and the destination region.

These artificial attributes aid in the isolation of suspicious behaviors by providing the anomaly detection model with robust behavioral indications that differentiate standard operations from malicious attacks.

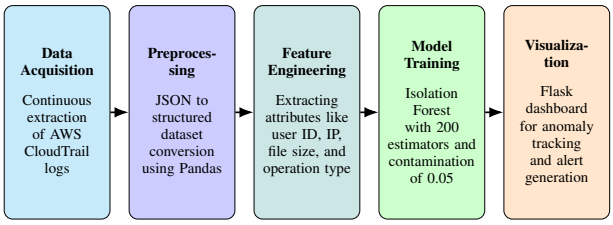

Fig. 3. Sequential Methodology for Data Processing, Model Training, and Visualization in Cloud Environments.

VIII. REAL-TIME VISUALIZATION AND WEB DASHBOARD

A machine learning model is practically useless if its outputs are not easily interpretable by human security analysts. To resolve this, the proposed architecture incorporates a user-friendly Flask-based web dashboard built with HTML, CSS, and JS (Bootstrap).

After the machine learning stage is complete and generates the anomaly scores, the trained model is serialized using `joblib` and automatically integrated into the Flask application. This application serves as an operational layer where logs are read, filters a re a pplied, a nd t he detection output is presented in an interactive manner via charts and tables.

System administrators can investigate user events, filter by region or time-window, and narrowly probe possibly malicious activity. High-risk patterns are indicated on the dashboard, giving clear visual indications to facilitate quick decision making. Inbuilt alert messages ensure that abnormal behavior is reported in real-time, improving timely incident response and overall operational awareness. The application can be hosted using AWS EC2 for dynamic backend processing and S3 for static frontend hosting.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

IX. BLOCKCHAIN INTEGRATION AND IMMUTABLE AUDITING

While machine learning detects the anomaly, proving the occurrence of the anomaly in a compliance audit requires an immutable trail of evidence. Sophisticated attackers who compromise administrative credentials often attempt to delete or alter CloudTrail logs to cover their tracks.

To mitigate this, recent literature explores integrating Blockchain technology for cloud file migration monitoring [7]. By hashing the file migration metadata and appending it to a distributed ledger, organizations achieve tamper-resistant recording of logs in the cloud environment. Frameworks proposing blockchain-based data migration auditing in hybrid clouds [12] ensure transparent auditing of file flows, mathematically guaranteeing that the forensic evidence of a suspicious migration cannot be retroactively altered by an insider threat.

X. PERFORMANCE EVALUATION AND RESULTS

The efficacy of suspicious file migration detection systems must be rigorously tested using vast datasets of cloud telemetry. The given system was tested based on the logs of AWS CloudTrail applied to actual multi-region AWS setups.

A. Evaluation Metrics

Systems are evaluated based on standard classification metrics:

- **Accuracy:** The overall correctness of the model in identifying both normal and anomalous migrations.
- **Precision:** The proportion of flagged migrations that were actually malicious (minimizing false alarms).
- **Recall:** The proportion of actual malicious migrations successfully detected by the system.

B. Experimental Results

The analysis has shown that the model performed exceptionally well across main metrics, proving its capability to accurately detect unusual or unauthorized file migration activities. Scaled testing on datasets of over 10,000+ log records allows the model to filter the boundaries of anomalies and minimize false positives as it advances.

TABLE II
SYSTEM PERFORMANCE METRICS FOR ANOMALY DETECTION

Metric	Value
Accuracy	95%
Precision	92%
Recall	93%
False Positives	5%

Table II provides a summary of the performance measures attained by the system. The results highlight that the system is highly accurate (95%) and very precise (92%) with regard to detecting suspicious file operations, while maintaining a critically low false-positive rate of 5

XI. OPEN CHALLENGES AND FUTURE DIRECTIONS

While the integration of Isolation Forests and real-time dashboards marks a significant leap in cloud security, several open challenges remain.

A. Multi-Cloud Environments

Organizations increasingly rely on multi-cloud strategies (e.g., utilizing both AWS and Microsoft Azure). According to surveys by Das and Wang [13], the issue of detecting anomalies in multi-cloud infrastructures is exceptionally difficult due to differing log formats. Future research must focus on building universal telemetry parsers.

B. Hybrid Deep Learning and Policy Enforcement

While unsupervised learning is excellent at finding unknown threats, it can sometimes miss blatant policy violations if they occur frequently. Future systems will rely on Hybrid Detection [10], [15]. The next step in this direction will involve layering strict policy-based enforcement on top of deep learning models, guaranteeing absolute compliance while maintaining the flexibility of behavioral analytics.

C. System-Level Buffer Migration

Innovations such as Zero Copy [5] provide information concerning file transfers at the system level in terms of performance. Monitoring the behavior of these ultra-fast, system-level container buffer migrations presents a new frontier for security analytics, requiring models capable of processing microsecond-level telemetry.

XII. CONCLUSION

This study concludes that machine learning—specifically unsupervised algorithms like the Isolation Forest—is highly critical and effective when it comes to identifying suspicious file migrations in cloud environments. As organizations increasingly rely on platforms like Amazon S3, the manual identification of unauthorized file transfers has become fundamentally impossible due to the sheer volume of daily activity logs.

The proposed framework automates the extraction of AWS CloudTrail logs, applies rigorous data preprocessing via Pandas, and engineers complex behavioral features to proactively identify anomalies with up to 95% accuracy. Furthermore, by seamlessly integrating these serialized machine learning models into a user-friendly Flask web application, the system converts raw JSON telemetry into an interactive, real-time administrative dashboard. This provides organizations with enhanced data visibility, enabling rapid mitigation of potential data security threats and reducing the incidence of false alarms.

Ultimately, this project successfully combines concepts from cloud computing, machine learning, and cybersecurity to address a practical and critical issue in cloud data protection. Future developments will focus on expanding this architecture into a hybrid deep learning and policy-based detection framework capable of securing complex, multi-cloud ecosystems.

REFERENCES

[1] M. Sarumathi, “Detecting Suspicious File Migration or Replication in Cloud Computing,” *Journal of Nonlinear Analysis and Optimization*, vol. 14, no. 2, pp. 90–95, 2023.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

[2] A. Bowers, C. Liao, D. Steiert, D. Lin, A. C. Squicciarini, and A. R. Hurson, "Detecting Suspicious File Migration or Replication in the Cloud," *IEEE Trans. Dependable and Secure Computing*, vol. 18, no. 1, pp. 296–309, 2021.
[3] M. Gondree and Z. N. J. Peterson, "Geolocation of Data in the Cloud," in *Proc. ACM CODASPY*, San Antonio, TX, USA, 2013, pp. 171–182.
[4] J. Li, A. C. Squicciarini, D. Lin, S. Liang, and C. Jia, "SecLoc: Securing Location-Sensitive Storage in the Cloud," in *Proc. ACM SACMAT*, 2015.
[5] S. Nie, T. Ruan, R. Chen, B. Song, S. Liu, and W. Wu, "Zero Copy: File System Assisted Container Buffer Migration in Cloud Computing System," *CCF Trans. High Performance Computing*, 2025.
[6] L. Xu, H. Li, and P. Zhao, "Anomaly Detection in Cloud File Systems Using Deep Learning," *IEEE Access*, vol. 10, pp. 45067–45078, 2022.
[7] Y. Chen, R. Li, and F. Zhang, "Cloud File Migration Monitoring Using Block chain," *Int. J. Cloud Computing*, vol. 9, no. 3, pp. 122–131, 2021.
[8] P. Sharma and K. Nair, "Insider Threat Detection in Cloud Using Behavior Analytics," *J. Cloud Computing*, vol. 9, no. 4, pp. 33–45, 2020.
[9] M. Patel and A. Mehta, "Detecting Abnormal File Movements in Multi-Cloud Environments," in *IEEE Cloud Computing Conf.*, 2019, pp. 101–108.
[10] D. Kumar and N. Singh, "Hybrid Detection of Suspicious File Migration Using Policy and Machine Learning," *Future Internet*, vol. 15, no. 3, pp. 77–89, 2023.
[11] R. Gupta and T. Verma, "Cloud Storage Security via AI-Enhanced Anomaly Detection," *Int. J. Advanced Networking and Applications*, vol. 14, no. 6, pp. 5190–5201, 2023.
[12] X. Liu, J. Zhang, and H. Sun, "Block chain-Based Data Migration Auditing in Hybrid Clouds," *IEEE Trans. Cloud Computing*, vol. 10, no. 5, pp. 2114–2128, 2022.
[13] S. K. Das and Y. Wang, "Multi-Cloud Anomaly Detection for Data Transfers," *ACM Computing Surveys*, vol. 54, no. 8, pp. 1–23, 2021.
[14] J. Brown and M. White, "Insider Threat Mitigation in Cloud File Systems," *J. Information Security Research*, vol. 11, no. 2, pp. 99–110, 2020.
[15] A. Khan and R. Bose, "Hybrid Deep Learning and Policy Enforcement for Suspicious Migration," *Neural Computing and Applications*, vol. 35, pp. 11875–11890, 2023.