International Journal of Scientific Research in Science, Engineering and Technology

Print ISSN - 2395-1990
Online ISSN : 2394-4099

OPEN ACCESS

Available Online at :www.ijsrset.com
doi : https://doi.org/10.32628/IJSRSET

Anomaly Detection in Cloud Computing: A Systematic Review of Machine Learning Approaches

Dhananjay Kumar, Jeetendra Singh Yadav

Department of Computer Science Engineering, Bhabha University, Bhopal, Madhya Pradesh, India

ARTICLEINFO

Article History:

Accepted : 25 May 2025
Published: 02 July 2025

Publication Issue :
Volume 12, Issue 4
July-August-2025
Page Number :
15-22

ABSTRACT

Cloud computing has become an essential pillar of digital infrastructure, offering on-demand services with high scalability and flexibility. However, ensuring consistent performance and reliability in such dynamic environments remains a significant challenge. Anomaly detection plays a critical role in identifying deviations from normal behavior that could indicate system faults, performance bottlenecks, or security breaches. Recently, machine learning (ML) techniques have shown remarkable success in detecting such anomalies due to their ability to analyze large volumes of heterogeneous data and adapt to evolving patterns. This systematic review explores various ML-based anomaly detection methods applied within cloud computing environments. The paper categorizes the approaches into supervised, unsupervised, and semi-supervised models, examining their performance, scalability, real-time capabilities, and implementation complexity. Additionally, it highlights the challenges related to data labeling, model generalization, and integration into live cloud systems. Key publicly available datasets and evaluation metrics used in the literature are also reviewed. The study concludes by identifying research gaps and proposing future directions to enhance the robustness, interpretability, and efficiency of ML-driven anomaly detection frameworks in cloud settings.

Keywords: Cloud Computing, Anomaly Detection, Machine Learning, Supervised Learning, Unsupervised Learning, Semi-Supervised Learning, Cloud Performance, System Reliability, Real-Time Monitoring, Cloud Security

INTRODUCTION

Cloud computing has revolutionized the delivery of IT services by providing on-demand access to shared computing resources such as servers, storage, databases, and applications over the internet. It enables businesses and individuals to scale resources dynamically, reduce infrastructure costs, and accelerate deployment cycles. As organizations

increasingly migrate their operations to cloud platforms, ensuring the stability, scalability, and efficiency of cloud environments has become critically important. Anomaly detection in cloud computing is the process of identifying unexpected patterns or behaviors in cloud systems that deviate from normal operations. These anomalies can indicate performance issues, system failures, or security threats, making timely detection critical for maintaining the reliability and efficiency of cloud services. As illustrated in **Figure 1**, the process begins with the cloud system generating vast amounts of operational data, such as logs, metrics, and user activity. This data is continuously collected and monitored using tools like AWS CloudWatch or Azure Monitor. The collected data is then analyzed by anomaly detection models—often based on machine learning techniques such as supervised, unsupervised, or semi-supervised learning. These models learn to differentiate between normal and abnormal behavior patterns and classify each incoming data point accordingly. If the data is flagged as anomalous, it can trigger alerts or automated responses to mitigate potential issues. By automating the detection of irregularities, cloud providers can ensure better performance, enhanced security, and improved user satisfaction.

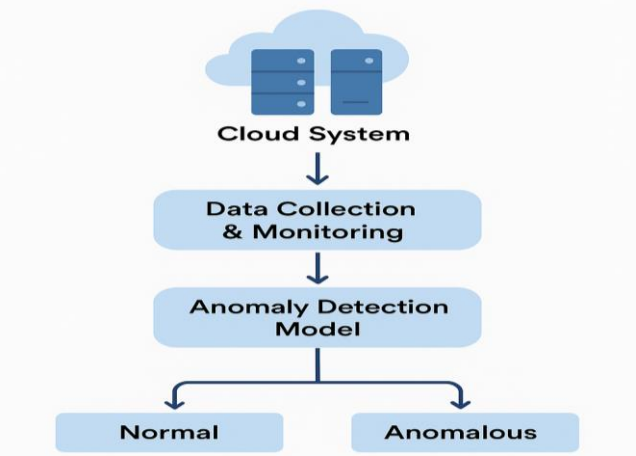

Fig 1: Anomaly Detection in Cloud Computing

A. Importance of Performance and Reliability in Cloud Services

The success of cloud services heavily depends on their performance and reliability. Users expect seamless access, minimal latency, and uninterrupted availability. However, cloud environments are inherently complex and dynamic, often involving multi-tenant architectures, virtualized infrastructure, and distributed components. Any anomaly, such as unexpected traffic surges, hardware failures, or security breaches, can severely degrade performance and lead to service-level agreement (SLA) violations, user dissatisfaction, and financial losses.

B. Role of Anomaly Detection

Anomaly detection plays a pivotal role in maintaining cloud service performance by identifying deviations from normal behavior. Detecting anomalies early allows service providers to proactively address potential issues before they escalate. These anomalies can manifest as performance bottlenecks, unauthorized access attempts, or abnormal usage patterns—making detection a cornerstone of cloud system monitoring, security, and optimization.

C. Motivation for Using Machine Learning in Anomaly Detection

Traditional rule-based anomaly detection techniques often struggle to cope with the scale, diversity, and dynamic nature of cloud data. Machine learning (ML) offers powerful solutions by learning patterns from historical data and generalizing to detect previously unseen anomalies. ML algorithms can adapt to changes in workload, detect subtle deviations, and operate effectively in real-time, making them ideal for anomaly detection in complex cloud environments.

D. Objectives and Scope of the Review

This review aims to systematically explore and evaluate the use of machine learning techniques for anomaly detection in cloud computing. It focuses on categorizing the ML approaches (supervised, unsupervised, and semi-supervised), analyzing their strengths and limitations, reviewing benchmark

datasets and tools, and identifying existing challenges and future directions in the field.

Structure of the Paper

The paper is organized as follows: Section II provides an overview of anomaly detection in cloud environments. Section III presents various machine learning techniques used for detecting anomalies. Section IV discusses the datasets and evaluation metrics used in the literature. Section V reviews tools and platforms for deploying ML-based solutions. Section VI outlines Conclusion.

Fundamentals of Anomaly Detection in Cloud Computing

Anomalies, also referred to as outliers or deviations, are data patterns that do not conform to expected behavior. In the context of cloud computing, anomalies can indicate faults, failures, or security incidents that impact system performance or reliability. Based on their nature, anomalies are typically classified into three categories:

- **Point Anomalies**: A single data instance significantly deviates from the rest (e.g., a sudden CPU spike).
- **Contextual Anomalies**: Data points that are anomalous only within a specific context (e.g., high memory usage at midnight might be unusual but normal during peak hours).
- **Collective Anomalies**: A group of related data instances that are anomalous when observed together, although individual points may appear normal (e.g., a sequence of low traffic on a normally high-traffic application).

Understanding these types is crucial for designing effective anomaly detection systems in cloud environments.

A. Sources of Anomalies in Cloud Systems

Cloud systems are highly dynamic and distributed, leading to multiple potential sources of anomalies, including:

- **Workload Spikes**: Sudden increases in user demand or traffic can overload services.
- **Hardware or Network Failures**: Physical component degradation, bandwidth bottlenecks, or configuration issues may cause unexpected behavior.
- **Virtual Machine (VM) or Container Issues**: Resource contention, migration delays, or misconfigurations can degrade performance.
- **Security Threats**: Attacks like Distributed Denial of Service (DDoS), data breaches, or unauthorized access generate irregular patterns.
- **Software Bugs or Deployment Errors**: Faulty updates or improper service integration can lead to inconsistent system states.

Identifying the source of anomalies is critical for root-cause analysis and effective remediation.

B. Challenges in Cloud-Based Anomaly Detection

Detecting anomalies in cloud systems presents several unique challenges:

- **High Volume and Velocity of Data**: Cloud environments generate massive data streams, making real-time processing difficult.
- **Data Variety and Heterogeneity**: Logs, metrics, traces, and events come from diverse sources and in varied formats.
- **Dynamic Behavior**: Rapid changes in workloads and infrastructure make it hard to define "normal" behavior.
- **Lack of Labeled Data**: Most real-world cloud datasets lack sufficient labels for training supervised models.
- **Noise and False Positives**: Anomaly detection systems may generate many false alarms, leading to alert fatigue.

These challenges demand more advanced, intelligent techniques for anomaly detection.

C. Need for Intelligent, Scalable Solutions

To effectively manage and monitor large-scale cloud systems, there is a growing need for intelligent and scalable anomaly detection solutions. Machine

learning techniques, particularly unsupervised and semi-supervised models, offer the ability to learn from raw or minimally labeled data and adapt to changes over time. Scalability is also essential, as the detection systems must process data across distributed cloud layers without introducing significant latency or overhead. Developing accurate, low-latency, and resource-efficient anomaly detection systems is key to ensuring high-quality cloud service delivery.

Overview of Machine Learning Approaches

Machine learning (ML) has emerged as a powerful tool for anomaly detection in cloud computing due to its capability to learn complex patterns from large-scale data and adapt to dynamic cloud environments. Depending on data availability, anomaly characteristics, and deployment context, different ML paradigms—supervised, unsupervised, semi-supervised, and reinforcement learning—are employed. This section presents a categorized overview of these approaches with practical examples from recent studies.

A. Supervised Learning Techniques

Supervised learning requires datasets labeled as either normal or anomalous. These models learn to classify new instances by recognizing patterns in historical data. Decision Trees, Support Vector Machines (SVM), and Neural Networks are among the most commonly applied algorithms in this category.

Decision Trees are interpretable and efficient for rule-based anomaly detection but may overfit in complex cloud scenarios.

SVMs are particularly effective for binary classification tasks in systems with well-defined boundaries between normal and abnormal behavior [3].

Neural Networks, especially deep models, have shown high accuracy in modeling non-linear relationships and are effective in analyzing high-dimensional cloud monitoring data [1].

Studies such as by **Tanam and Raja (2024)** [1] and **Yasarathna and Munasinghe (2020)** [3] have shown that supervised models can yield high detection accuracy when trained on sufficient labeled data. However, these approaches struggle with generalization to novel anomalies and require large volumes of labeled examples, which are often unavailable in real-world cloud systems.

B. Unsupervised Learning Techniques

Unsupervised learning is highly relevant in cloud environments where labeling is expensive or infeasible. These models learn to capture the normal behavior of a system and detect deviations.

Clustering algorithms, such as K-means and DBSCAN, are widely used to identify groups of similar behavior and flag outliers [4].

Autoencoders, as demonstrated by **Islam and Miranskyy (2020)** [5], are neural networks trained to reconstruct inputs; high reconstruction error indicates anomalies.

Zhang et al. (2018) [2] proposed **PerfInsight**, an unsupervised clustering-based anomaly detection system, highlighting the advantage of reduced dependency on labeled data and better generalization across diverse anomaly types. However, unsupervised models may generate higher false positives and are often less interpretable.

C. Semi-Supervised Learning Techniques

Semi-supervised learning is a hybrid approach that utilizes a large amount of normal data and a smaller set of labeled anomalies.

One-Class SVM, used by **Yasarathna and Munasinghe (2020)** [3], learns the boundary of normal data and detects deviations.

Self-training models iteratively use a small labeled dataset to train and then label additional data points, gradually improving detection accuracy.

These models reduce the reliance on fully labeled datasets while achieving reasonable performance, especially in cloud systems where anomaly labels are scarce. Nevertheless, they may suffer from error propagation and rely on high-quality initial labels.

D. Reinforcement Learning (RL)

Though less common, reinforcement learning (RL) is gaining attention for its potential in adaptive anomaly detection and response systems.

RL agents can **dynamically tune thresholds**, prioritize alerts, and implement **automated mitigation strategies** based on environmental feedback [15].

Ma et al. (2022) [15] introduced a **federated transformer-based framework** for cloud manufacturing that uses RL to preserve data privacy while optimizing detection.

The adaptability of RL is a major advantage, especially in rapidly evolving cloud environments. However, its practical deployment is complex, requiring careful design of the reward mechanism and exploration strategies.

Datasets and Benchmarking

A critical component of developing and evaluating machine learning-based anomaly detection systems is access to reliable datasets. The quality, volume, and nature of the dataset used directly influence the model's performance, generalizability, and applicability to real-world cloud environments. This section discusses commonly used public datasets, the distinction between synthetic and real-world data, and essential preprocessing and feature selection techniques[6].

A. Publicly Available Datasets

Several publicly available datasets are frequently used for benchmarking anomaly detection models in cloud computing. These datasets contain performance logs, system metrics, and event traces that reflect normal and abnormal system behaviors[7]:

- **NASA's Space Shuttle and Turbofan Engine Degradation Datasets** Widely used for detecting system failures and anomalies in time-series data; relevant for predictive maintenance in cloud data centers.
- **Yahoo Webscope S5 Dataset** Contains time-series data with labeled anomalies, simulating real-world traffic fluctuations. It is often used for evaluating streaming anomaly detection algorithms.
- **Microsoft Azure Telemetry Logs** Provides anonymized telemetry data including CPU usage, disk activity, and network performance, ideal for performance anomaly detection in cloud services.
- **KPI Datasets (e.g., AIOps Challenge)** Includes time-series key performance indicators (KPIs) with labeled anomalies from real cloud monitoring systems.

These datasets serve as benchmarks for testing model accuracy, speed, and scalability under different cloud scenarios.

B. Synthetic vs. Real-world Data

- **Synthetic Data**

Generated artificially using simulation tools or rule-based injection of anomalies. It provides controlled environments for model training and stress testing.

  - *Advantages*: Easy to generate, balanced class distributions, and known ground truth.
  - *Disadvantages*: May lack the complexity, noise, and unpredictability of real-world data.

- **Real-world Data**

Collected from live cloud systems, including logs, traces, and operational metrics.

  - *Advantages*: Offers realistic patterns, diverse anomalies, and operational relevance.
  - *Disadvantages*: Often unlabeled, imbalanced, and contains noisy or redundant features.

An ideal evaluation involves using a mix of synthetic data for model development and real-world data for validation[8].

C. Dataset Preprocessing and Feature Selection Techniques

Preprocessing and feature selection are essential to convert raw cloud data into a format suitable for machine learning models:

- **Data Cleaning**: Removing missing values, correcting inconsistencies, and handling outliers[9].

- **Normalization/Standardization**: Scaling numerical features to ensure uniformity and improve convergence in training.
- **Windowing for Time-Series Data**: Converting continuous data streams into overlapping/non-overlapping windows for temporal modeling.
- **Feature Extraction**: Deriving statistical (mean, variance), temporal (trends, spikes), and categorical features from logs and metrics.
- **Dimensionality Reduction**: Using techniques like PCA or t-SNE to reduce feature space while retaining critical patterns.
- **Encoding Categorical Data**: Transforming non-numerical data using one-hot encoding or label encoding.

Effective preprocessing improves model accuracy and reduces computational overhead, especially in large-scale cloud environments[10].

Tools, Frameworks, and Deployment Platforms.

Modern machine learning development benefits greatly from powerful and user-friendly open-source frameworks. These libraries provide scalable, modular components for data processing, model design, and performance evaluation:

- **TensorFlow**
  Developed by Google, TensorFlow supports deep learning and large-scale model training. It is widely used for building and deploying neural networks, including autoencoders and recurrent models for time-series anomaly detection.
- **PyTorch**
  A dynamic and flexible deep learning framework developed by Facebook. PyTorch is popular in research environments and increasingly in production, offering ease of use, GPU acceleration, and strong support for custom model design.
- **Scikit-learn**
  A comprehensive Python library for traditional machine learning algorithms like decision trees, SVM, and clustering. It is ideal for prototyping and integrating ML models with existing systems for anomaly detection.
- **Keras**
  A high-level neural networks API built on top of TensorFlow, offering a simplified interface for rapid model development and experimentation.
- **XGBoost & LightGBM**
  Gradient boosting frameworks known for high accuracy and efficiency, often used in supervised anomaly detection tasks with structured cloud telemetry data.

CONCLUSION

Anomaly detection plays a vital role in maintaining the performance, security, and reliability of cloud computing environments. With the increasing complexity and scale of cloud infrastructure, traditional detection methods are often inadequate. This literature review has highlighted how machine learning (ML) has emerged as a powerful solution, capable of identifying subtle and complex anomalies in large volumes of dynamic cloud data.

We explored various ML techniques, including supervised, unsupervised, semi-supervised, and reinforcement learning approaches, each offering unique strengths depending on data availability and operational requirements. The review also examined benchmark datasets, model evaluation metrics, deployment platforms, and key tools that support the development of intelligent anomaly detection systems. Despite their advantages, current ML-based methods face several challenges such as limited labeled data, scalability issues, and high false alarm rates.

To move forward, future research should focus on developing hybrid and adaptive models, integrating explainable AI (XAI) techniques, and leveraging edge-cloud collaboration for real-time detection. Emphasis should also be placed on creating standardized datasets and performance benchmarks to facilitate fair and meaningful comparisons across solutions.

In summary, machine learning-based anomaly detection holds great promise for enhancing cloud service performance, and continued research in this direction is essential for building more resilient and intelligent cloud ecosystems.

REFERENCES

[1]. A. Tanam and G. Raja, "A Systematic Analysis on Security and Anomaly Detection using Machine Learning in Cloud Computing," 2024 IEEE 6th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA), Hamburg, Germany, 2024, pp. 417–422, doi: 10.1109/ICCCMLA63077.2024.10871879.

[2]. X. Zhang, F. Meng and J. Xu, "PerfInsight: A Robust Clustering-Based Abnormal Behavior Detection System for Large-Scale Cloud," 2018 IEEE 11th International Conference on Cloud Computing (CLOUD), San Francisco, CA, USA, 2018, pp. 896–899, doi: 10.1109/CLOUD.2018.00130.

[3]. T. L. Yasarathna and L. Munasinghe, "Anomaly Detection in Cloud Network Data," 2020 International Research Conference on Smart Computing and Systems Engineering (SCSE), Colombo, Sri Lanka, 2020, pp. 62–67, doi: 10.1109/SCSE49731.2020.9313014.

[4]. X. Zhao and W. Zhang, "An Anomaly Intrusion Detection Method Based on Improved K-Means of Cloud Computing," 2016 Sixth International Conference on Instrumentation & Measurement, Computer, Communication and Control (IMCCC), Harbin, China, 2016, pp. 284–288, doi: 10.1109/IMCCC.2016.108.

[5]. M. S. Islam and A. Miranskyy, "Anomaly Detection in Cloud Components," 2020 IEEE 13th International Conference on Cloud Computing (CLOUD), Beijing, China, 2020, pp. 1–3, doi: 10.1109/CLOUD49709.2020.00008.

[6]. L. Jie, L. Xiangxiang and L. Haoxiang, "Anomaly Detection Method of Power Dispatching Data Based on Cloud Computing Platform," 2020 International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE), Bangkok, Thailand, 2020, pp. 23–26, doi: 10.1109/ICBASE51474.2020.00012.

[7]. R. Kumar and D. Sharma, "HyINT: Signature-Anomaly Intrusion Detection System," 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Bengaluru, India, 2018, pp. 1–7, doi: 10.1109/ICCCNT.2018.8494088.

[8]. D. Kadam, R. Patil and C. Modi, "An Enhanced Approach for Intrusion Detection in Virtual Network of Cloud Computing," 2018 Tenth International Conference on Advanced Computing (ICoAC), Chennai, India, 2018, pp. 80–87, doi: 10.1109/ICoAC44903.2018.8939107.

[9]. D. Fernando, M. A. Rodriguez and R. Buyya, "iAnomaly: A Toolkit for Generating Performance Anomaly Datasets in Edge-Cloud Integrated Computing Environments," 2024 IEEE/ACM 17th International Conference on Utility and Cloud Computing (UCC), Sharjah, United Arab Emirates, 2024, pp. 236–245, doi: 10.1109/UCC63386.2024.00041.

[10]. B. Cha and J. Kim, "Study of Multistage Anomaly Detection for Secured Cloud Computing Resources in Future Internet," 2011 IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing (DASC), Sydney, NSW, Australia, 2011, pp. 1046–1050, doi: 10.1109/DASC.2011.171.

[11]. L. Liang, "Simulation of Big Data Anomaly Detection Algorithm Based on Neural Network Under Cloud Computing Platform," 2024 International Conference on Electrical Drives, Power Electronics & Engineering (EDPEE), Athens, Greece, 2024, pp. 603–608, doi: 10.1109/EDPEE61724.2024.00118.

[12]. H. H. Bin Suhaimi and H. T. Zubair, "Data Leakage Detection in Cloud Computing Environment," in Proc. 2024 1st Int. Conf. Cyber Security and Computing (CyberComp), Melaka, Malaysia, 2024, pp. 7–12, doi: 10.1109/CyberComp60759.2024.10913619.

[13]. S. Ma et al., "Privacy-Preserving Anomaly Detection in Cloud Manufacturing Via Federated Transformer," IEEE Trans. Ind.

Informat., vol. 18, no. 12, pp. 8977–8987, Dec. 2022, doi: 10.1109/TII.2022.3167478.

[14]. K. R. Shreesha, S. Anjana and B. Padma, "Enhancing the Stadam SLA Trust Model with Machine Learning for Improved Anomaly Detection," in Proc. 2025 Int. Conf. Next Generation Communication & Information Processing (INCIP), Bangalore, India, 2025, pp. 727–731, doi: 10.1109/INCIP64058.2025.11019740.

[15]. W. Guo, L. Shi and Z. Wu, "Research on anomaly detection algorithm of time series data in cloud environment," in Proc. 2022 World Automation Congress (WAC), San Antonio, TX, USA, 2022, pp. 499–503, doi: 10.23919/WAC55640.2022.9934501.