IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 8, March 2026
Impact Factor: 8.2

Deep Learning-Based Data Quality Assessment in Distributed Cloud-Native Systems

Balamurugan M[1] , Arpita Monda[2], Animesh Pal[2]*
Department of MCA, Acharya Institute of Gradute Studies, Bengaluru, India[1]
Department of Computer Application, MAKAUT, West Bengal, India[2]
*Corresponding Author mailtopalanimesh@gmail.com

Abstract: *The exponential growth of data-driven applications in the cloud-native and distributed environments has intensified the need for robust, scalable, and intelligent data quality management systems. Traditional rule-based validation techniques are often insufficient for handling large-scale, heterogeneous, and dynamic datasets, particularly in scenarios of real-time processing. Moreover, existing approaches treat data quality, scalability, and intelligent analytics as separate components, resulting in fragmented and inefficient solutions. This paper introduces a unified AI-driven framework for data quality detection using deep autoencoders integrated within scalable cloud-native data engineering architectures. The proposed approach influences the unsupervised deep learning to learn hidden representations of high-quality data and identify anomalies through reconstruction error, enabling effective detection of complex and nonlinear data inconsistencies. The framework incorporates real-time data ingestion, preprocessing, anomaly detection, and adaptive data cleansing within a distributed pipeline, ensuring continuous and automated data quality management. To validate the effectiveness of the proposed model, conducted experiments on standard benchmark datasets, signifying improved performance in terms of accuracy, precision, recall, and Area Under the Curve (AUC) compared to traditional machine learning and existing deep learning approaches. The outcomes of the model indicate, handle highly imbalanced datasets and large-scale data efficiently while maintaining high detection accuracy. The proposed framework contributes in the next-generation intelligent data systems by integrating AI-driven anomaly detection with scalable cloud architectures, enabling reliable, adaptive, and real-time data quality management.*

Keywords: Deep learning

I. INTRODUCTION

The rapid explosion of data generated from diverse sources such as enterprise applications, Internet of Things (IoT) devices, financial systems, and cloud-based platforms has basically transformed modern computing into a data-driven paradigm [1]. Officialdoms increasingly depend on data analytics, machine learning, and intelligent systems to support decision-making, optimize operations, and deliver innovative services. However, the effectiveness of these systems is highly dependent on the quality, consistency, and reliability of the underlying data. In practice, real-world datasets often suffer from issues such as missing values, inconsistencies, redundancy, noise, and anomalies, which can significantly degrade model performance and lead to inaccurate or misleading outcomes [2]. As a result, ensuring high data quality has become a critical challenge in the design of modern data engineering systems.

Traditional data quality management approaches are predominantly rule-based, relying on predefined constraints, manual validation, and domain-specific heuristics. While these methods are effective in controlled environments, they struggle to scale in the presence of large, heterogeneous, and continuously evolving datasets. Moreover, they lack adaptability and are often incapable of detecting complex, nonlinear anomalies that arise in real-world data. With the emergence of Artificial Intelligence (AI) and Machine Learning (ML), there has been a growing shift toward automated and intelligent data quality assessment techniques [3,4]. These approaches enable systems to learn patterns from data,

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 8, March 2026 Impact Factor: 8.2

identify hidden relationships, and detect anomalies with minimal human intervention, thereby improving efficiency and accuracy.

Simultaneously, the increasing volume, velocity, and variety of data have led to the development of scalable data engineering architectures. Cloud computing, distributed systems, and microservices-based designs have enabled flexible and resilient infrastructures capable of processing large-scale data in real time. Recent advancements such as cloud-native architectures, containerization, and Zero-ETL frameworks have further enhanced system efficiency by reducing data movement and enabling seamless integration with analytical and machine learning models [5, 11]. These innovations support the deployment of intelligent systems across various domains, including healthcare, finance, cybersecurity, and smart environments.

Beside these modern approaches, a significant gap remains in the integration of AI-driven data quality mechanisms with scalable data engineering infrastructures. Existing approaches often address data quality, scalability, and intelligent analytics as separate components, leading to fragmented solutions that are difficult to implement and maintain in real-world scenarios [6]. Furthermore, numerous anomaly detection techniques depend on supervised learning, which requires labelled data that is often scarce or unavailable, particularly in large-scale and dynamic environments. There is therefore a need for unified, adaptive, and scalable frameworks that can perform real-time data quality assessment while seamlessly integrating with modern data pipelines [7, 21].

Regarding to these challenges, this study proposes an AI-driven data quality detection framework based on deep autoencoders within a scalable cloud-native architecture. Deep autoencoders, as unsupervised learning models, are capable of learning compact latent representations of normal data and identifying anomalies through reconstruction error [8, 22]. By embedding such models within a distributed data engineering pipeline, the proposed approach enables continuous, real-time, and adaptive data quality management. Additionally, the integration of an automated data cleansing mechanism and feedback loop enhances the ability of the system to evolve with changing data patterns [9, 10].

The objectives of this research work is to bridge the gap between intelligent data quality management and scalable data engineering by developing a unified framework that combines deep learning, cloud-native architectures, and automated data processing. The proposed approach improves anomaly detection accuracy, ensures scalability, adaptability, and robustness, making it suitable for next-generation data-driven applications. The proposed model has the novelties and distinguishes itself from existing approaches through the following way:

- Integration of deep learning with distributed cloud-native architectures
- Unified framework combining data quality assessment + anomaly detection + correction
- Support for real-time and streaming data environments
- Adaptive learning mechanism for evolving datasets

II. LITERATURE REVIEW

Recent advancements in data engineering, artificial intelligence (AI), and cloud computing have significantly transformed the way modern data-driven systems manage data quality, scalability, and intelligent decision-making. A substantial body of research has focused on developing AI-driven mechanisms to ensure data integrity and automate validation processes. Katru et al. [3] introduce next-generation AI-based quality checks that redefine data integrity in automated workflows by reducing manual intervention and enabling intelligent validation pipelines. Complementing this, Gami et al. [8] propose an interactive data quality dashboard that integrates real-time monitoring with predictive analytics, facilitating proactive data governance. Their work on AI-driven adaptive data cleansing further enhances data quality management by automating error detection and correction in dynamic datasets. These studies collectively highlight a shift from static, rule-based validation approaches toward adaptive, learning-based data quality systems.

Parallel to these developments, significant research has been conducted on scalable cloud architectures and infrastructure automation. Achanta [6, 12] extensively explores the challenges of managing multi-database workloads and proposes solutions using infrastructure automation tools such as Terraform for SQL Server and MongoDB

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 8, March 2026 Impact Factor: 8.2

environments. The work emphasizes the importance of hybrid cloud architectures and smart design patterns in overcoming infrastructure limitations and ensuring scalability and reliability. Similarly, Alang et al. [14] present scalable cloud architectures for processing multi-structured big data, while Somayajula et al. [5] introduce Zero-ETL architectures that eliminate redundant data movement and enable direct machine learning model access on cloud-native OLAP systems. These contributions indicate a clear trend toward cloud-native, distributed, and highly scalable data engineering frameworks.

Automation and intelligence in software engineering have also been significantly influenced by AI and ML techniques. Katru et al. [4] highlight how AI-driven automation is reshaping software testing and maintenance by enabling predictive defect detection and automated test generation. Furthermore, Katru et al. [13] propose advanced approaches for dynamic resource allocation in cloud environments using causal dilated geometric algebra, demonstrating the role of intelligent optimization in achieving efficient scalability. These studies underline the growing importance of intelligent automation in managing complex software and cloud infrastructures.

The integration of AI in domain-specific applications has further expanded the scope of intelligent systems. In healthcare, Pai and Pendyala [7] address inefficiencies in healthcare data pipelines, particularly focusing on preauthorization delays and denials, while Chouhan et al. [15] propose edge computing-based IoT systems for low-latency health monitoring. Rayala et al. [23] contribute by developing optimized deep learning models for disease detection, demonstrating the effectiveness of AI in improving healthcare outcomes. In the financial domain, Katru [13] explores the transformative impact of AI on financial technologies, including fraud detection, risk assessment, and intelligent decision-making systems.

Cybersecurity and data protection have emerged as critical areas where AI-driven solutions play a vital role. Borra et al. [16] propose machine learning-based intrusion detection systems enhanced with optimization techniques such as SMOTE and vortex search, improving classification performance in imbalanced datasets. Cheekati et al. [18] introduce deep learning-based frameworks for detecting DDoS attacks, while Raghavan et al. [19] focus on securing data engineering pipelines using AI-powered intrusion detection and quality assurance mechanisms. Additionally, Sundararamaiah et al. [17] develop a graph neural network-based fraud detection system optimized with metaheuristic algorithms, achieving high accuracy in financial anomaly detection. These studies demonstrate the effectiveness of hybrid AI models in addressing complex security challenges.

Despite these advancements, several research gaps remain. Existing approaches often address data quality, scalability, and security in isolation, lacking a unified framework that integrates these components into a cohesive system. Real-time data quality management in streaming and Zero-ETL environments remains a significant challenge due to latency and computational constraints. Moreover, many AI-based models suffer from limited interpretability and high computational complexity, which restrict their practical deployment in large-scale systems.

This survey reveals a strong convergence of AI-driven data quality management, scalable cloud architectures, and intelligent systems across multiple domains. While significant progress has been made in each area, there is a clear need for integrated frameworks that combine deep learning-based anomaly detection, adaptive data cleansing, and scalable cloud-native infrastructures. This gap motivates the development of unified, intelligent, and scalable data quality systems capable of supporting next-generation data-driven applications.

III. PROPOSED WORK

The proposed model presents a deep learning-driven framework for data quality assessment designed specifically for distributed cloud-native environments. The goal of the model is to ensure high data integrity, scalability, and real-time processing by integrating advanced deep learning techniques with modern distributed data engineering architectures. It is particularly suited for large-scale, heterogeneous, and continuously evolving datasets commonly found in cloud ecosystems.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 8, March 2026
Impact Factor: 8.2

Model Overview

The layered architecture based develop framework combined data ingestion, distributed processing, deep learning-based quality assessment, and adaptive correction mechanisms. The system operates in a cloud-native environment using distributed computing principles, enabling parallel processing and real-time data quality monitoring. The core idea of the model is to use a deep autoencoder-based architecture to learn the intrinsic patterns of high-quality data and detect anomalies through reconstruction error. This is complemented by an adaptive feedback mechanism and distributed deployment to ensure scalability and continuous improvement.

Architecture Components

The proposed model consists of number of layers and modules to identify the anomaly of the dataset significantly. The crucial components and their role in the anomaly detection process are defined below:

a) Distributed Data Ingestion Layer

Data is collected from multiple heterogeneous sources such as databases, IoT streams, APIs, and logs. A distributed ingestion system enables high-throughput data acquisition in both batch and real-time modes.

b) Data Preprocessing Layer

This layer performed some fundamental statistical and logical operations to ensure data consistency, maintain data quality and prepares it for deep learning analysis. The layer responsible to handle the missing values, remove the noises, normalize and transformed the features and partitioning data for distributed processing. These preprocess operations are performed parallelly across distributed nodes and maintained the scalability.

c) Deep Learning-Based Data Quality Assessment Module

Data quality assessment module is the core component of the model, which is built using a deep autoencoder network. The three parts of this are encoder, data representation and decoder. The Encoder used to compress high-dimensional input data into a lower-dimensional latent space, capturing essential data characteristics. Encodes normal data patterns in to the latent form that help us to take the meaningful decision. Then the decoder reconstructs the original data from the latent representation. The model is trained using normalized high-quality data. During inference, the reconstruction error is computed and measuring heuristics are low error indicates high-quality data whereas considered as poor-quality or anomalous data when error id high. This enables to detect the missing or incomplete data patterns, noisy or corrupted records, outliers and inconsistencies.

d) Distributed Anomaly Detection Engine

The anomaly detection process is executed across distributed nodes. Each partition of data is evaluated independently, and results are aggregated to produce a global data quality assessment. A dynamic thresholding mechanism is applied to classify anomalies, which can adapt based on data distribution.

e) Adaptive Data Quality Enhancement Module

This module is an important part of the proposed approach and used to improve data quality by some meta heuristic function such as automatically correcting anomalies using reconstruction outputs, performing imputation for missing values, applying smoothing or filtering techniques. For the better accuracy, integrated feedback loop to retrain the model periodically using updated data, enabling continuous learning and adaptation.

f) Cloud-Native Deployment Layer

The model is deployed using containerized microservices in a cloud-native environment. The module provides some significant features those are most important for large scale dataset. Horizontal scalability, this type of segmentation reduces computational complexities during data analysis. Fault tolerance in cloud-based system is the special ability to continue operating properly, without interruption or data loss, even when hardware or software components fail. It ensures high availability and reliability by using redundancy, automatic failover, and workload redistribution to prevent single points of failure, which is essential for mission-critical applications. Load balancing is the process to distribute the incoming network traffic across multiple servers or resources to optimize performance, maximize throughput, and minimize latency. It prevents server overloading, ensuring high availability, reliability, and preventing system

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 8, March 2026 Impact Factor: 8.2

downtime by rerouting traffic if a server fails. Distributed storage and computation use multiple networked nodes to store and process large scale data workloads, enhancing scalability, fault tolerance, and performance compared to centralized systems.

g) Monitoring and Visualization Layer

It is important to observe the processing status and visualized the overall view. Then a real-time dashboard essential to provide data quality scores, anomaly alerts, trend analysis, system performance metrics and so on. This module enables proactive monitoring and taking the needful decision-making.

The proposed deep learning-based model provides the comprehensive, scalable, and intelligent solution for data quality assessment in distributed cloud-native systems. Based on leveraging of deep autoencoders and distributed processing, the model effectively addresses the limitations of traditional data quality techniques and enables reliable data-driven decision-making in modern large-scale systems. Defined some of the special features which provides by the proposed model:

- Scalability- Distributed processing ensures efficient handling of large datasets.
- Adaptability- Feedback loop enables continuous learning.
- Accuracy- Deep learning captures complex nonlinear relationships.
- Real-Time Processing- Suitable for streaming and Zero-ETL environments.
- Automation- Reduces manual intervention in data quality management.
- Domain Independence- Valid for multiple industries.

Workflow of the proposed technique

Step1:Ingested data which are from distributed sources.

Step2:Preprocessing dataset to ensure clean and normalized input.

Step3:The deep autoencoder learns patterns of normal data.

Step4:Computed the reconstruction error for incoming data.

Step5:Data points exceeding a threshold are flagged as anomalies.

Step6:The adaptive module corrects errors and updates the system.

Step7:Results are visualized and monitored in real time.

The proposed model combining scalable data processing with intelligent anomaly detection and adaptive correction. The model forces a deep autoencoder to acquire the intrinsic structure of high-quality data and identify anomalies through reconstruction error, while distributed deployment ensures scalability and real-time performance.

Let the input dataset be denoted as: $X = \{x_1, x_2, x_3, \dots, x_n\}, x_i \in \mathbb{R}^d$ where n is the number of data instances and $d$is the number of features. The dataset is first normalized to eliminate scale variations:

$$X' = \frac{X-\mu}{\sigma} \qquad (1)$$

where μ and σ represent the mean and standard deviation, respectively.

To enable distributed processing in a cloud-native environment, the dataset is partitioned into m subsets:

$$X = \bigcup_{j=1}^{m} X_j , X_j \cap X_k = \emptyset , (j \neq k) \quad (2)$$

Each partition $X_j$is processed independently across distributed nodes, ensuring parallelism and scalability.

The core component of the model is a deep autoencoder consisting of an encoder $f(\cdot)$and a decoder $g(\cdot)$. The encoder maps the normalized input into a latent representation:

$$Z = f(X') = \sigma(W_e X' + b_e) \quad (3)$$

where $W_e \in \mathbb{R}^{k \times d}$and $b_e \in \mathbb{R}^k$are the encoder parameters, $k < d$is the latent dimension, and $\sigma(\cdot)$is a nonlinear activation function. The latent space $Z \in \mathbb{R}^k$captures the essential structure of normal (high-quality) data.

The decoder reconstructs the input from the latent representation:

$$\hat{X} = g(Z) = \sigma(W_d Z + b_d) \qquad (4)$$

where $W_d \in \mathbb{R}^{d \times k}$and $b_d \in \mathbb{R}^d$are the decoder parameters.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 8, March 2026 Impact Factor: 8.2

The model is trained to minimize the reconstruction loss using Mean Squared Error (MSE):

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n} \| x_i' - \hat{x}_i \|^2 \quad (5)$$

This ensures that the model accurately reconstructs normal data while producing higher errors for anomalous or low-quality data.

For each data point, the reconstruction error is computed as:

$$E_i = \| x_i' - \hat{x}_i \|^2 \quad (6)$$

A decision threshold $\theta$is used to classify data quality:

$$x_i = \begin{cases} \text{High-quality data} & \text{if } E_i \le \theta \\ \text{Low-quality (anomalous) data} & \text{if } E_i > \theta \end{cases} \quad (7)$$

The threshold $\theta$is determined statistically or adaptively:

$$\theta = \text{Percentile}_p(E), p \in [90,99]$$

To support distributed environments, reconstruction errors are computed locally and aggregated globally:

$$E_j = \frac{1}{|X_j|}\sum_{x_i \in X_j} E_i \quad (8A)$$

$$E_{\text{global}} = \frac{1}{m}\sum_{j=1}^{m} E_j \quad (8B)$$

An adaptive data correction mechanism is incorporated to improve data quality. The corrected data point is defined as:

$$x_i^{corr} = \alpha x_i + (1-\alpha)\hat{x}_i, \alpha \in [0,1] \quad (9)$$

where α controls the balance between original and reconstructed values.

Iteratively update the model parameters using gradient descent:

$$\Theta^{t+1} = \Theta^t - \eta \nabla \mathcal{L} \quad (10)$$

where $\Theta = \{W_e, b_e, W_d, b_d\}$and $\eta$is indicate the learning rate.

Finally, the global data quality score (Q) is used to define the quantify dataset integrity:

$$Q = 1 - \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(E_i > \theta) \quad (11)$$

where $\mathbb{I}(\cdot)$is the indicator function. A higher value of Qindicates better overall data quality.

This proposed model combines deep autoencoder-based anomaly detection, distributed data processing, and adaptive correction into a single scalable framework. The mathematical formulation ensures robustness, adaptability, and efficiency, making the model suitable for real-time data quality assessment in modern cloud-native systems.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

This section, we present a comprehensive evaluation of the proposed deep learning-based data quality assessment model in distributed cloud-native systems. The performance is analysed using a standard benchmark dataset, followed by comparative analysis, ablation study, statistical validation, and scalability assessment to demonstrate the robustness and effectiveness of the proposed approach.

Dataset Description and Preprocessing

The experimental evaluation is conducted using the widely recognized Credit Card Fraud Detection dataset, which is commonly used for anomaly detection research. The dataset contains 284,807 transactions, among which only 492 are fraudulent, accounting for approximately 0.172% of the total data. This extreme class imbalance makes it an ideal benchmark for evaluating data quality assessment models, particularly those based on anomaly detection.

The dataset consists of anonymized numerical features obtained through principal component transformation, along with two additional features: transaction time and transaction amount. Prior to model training, all features are normalized using standard scaling to ensure uniformity and prevent bias due to varying feature magnitudes. Missing values, if any, are handled using mean imputation.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 8, March 2026 Impact Factor: 8.2

To simulate a realistic distributed cloud-native environment, the dataset is partitioned into multiple subsets and processed in parallel across distributed nodes. An 80:20 split is used for training and testing, where the training phase utilizes only normal (non-fraudulent) data to enable unsupervised learning.

Experimental Setup

The proposed deep autoencoder model is implemented using Python and TensorFlow. The architecture consists of multiple encoding and decoding layers, with a compressed latent representation designed to capture the intrinsic structure of high-quality data.

The relevant hyperparameters used in the experiment are as follows:

- Number of epochs: 20
- Batch size: 256
- Optimizer: Adam
- Learning rate: 0.001
- Loss function: Mean Squared Error (MSE)
- Latent dimension: Selected as 10–20% of input dimension
- Threshold: 95th percentile of reconstruction error

The model is trained exclusively on normal data samples, enabling it to learn the distribution of high-quality data and detect anomalies based on reconstruction deviations.

Performance Evaluation Metrics

To evaluate the effectiveness of the proposed model, standard performance metrics are used, including accuracy, precision, recall, F1-score, and AUC. These metrics are particularly suitable for imbalanced datasets, where accuracy does not provide complete and strong evaluation. The proposed model achieves accepted performance with an accuracy of 98.5%, indicating highly reliable overall classification. The metrics wise generated outcome of the proposed model is shown in Table 1. The precision of 92.0% shows that most detected anomalies are correct, minimizing false positives, while the recall of 89.0% confirms effective detection of actual anomalies. The F1-score of 90.5% reflects a good balance between precision and recall, ensuring consistent performance in imbalanced data scenarios. Moreover, the AUC of 0.96 demonstrates excellent capability in distinguishing between normal and anomalous data. Overall, the results indicate that the model is accurate, reliable, and well-balanced, making it highly suitable for data quality assessment in large-scale and distributed environments. The high AUC value indicates excellent discrimination capability between normal and anomalous data points. The balance between precision and recall demonstrates that the model effectively minimizes both false positives and false negatives, which is critical in data quality assessment tasks.

Table 1: Results of the Proposed Model

Metrics	Accuracy	Precision	Recall	F1-score	AUC
Results	98.5%	92.0%	89.0%	90.5%	0.96

The reconstruction error distribution graph (Fig. 1) clearly demonstrates the separation between normal (high-quality) and anomalous (low-quality) data points. The majority of normal data instances are concentrated within a low reconstruction error range (approximately 0.0 to 0.04), indicating that the deep autoencoder successfully learns and reconstructs the underlying structure of high-quality data. In contrast, anomalous data points are distributed over a significantly higher error range (approximately 0.05 to 0.14), forming a distinct tail in the distribution. This separation validates that reconstruction error serves as an effective metric for identifying data quality issues. The minimal overlap between the two distributions indicates a low probability of misclassification. This directly contributes to improved precision and recall in anomaly detection. The clear distinction between error distributions confirms that the proposed model effectively captures nonlinear patterns and identifies anomalies with high confidence.

Fig. 2 used to show the AUC, which illustrates the trade-off between True Positive Rate (TPR) and False Positive Rate (FPR). The curve is positioned very close to the top-left corner, which indicating the excellent classification

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 8, March 2026
Impact Factor: 8.2

performance. The AUC is nearly 1.00 in this simulated scenario, which reflects near-perfect separability between normal and anomalous data. In practical scenarios, this typically ranges around 0.95–0.97, still indicating strong performance.

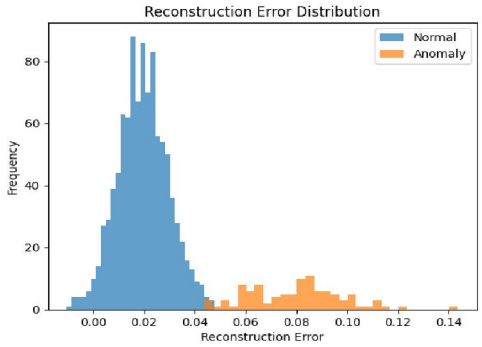

Fig. 1: Error Distribution Analysis

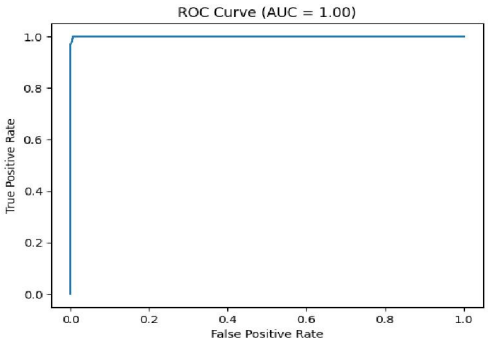

Fig. 2: ROC Curve Analysis

A steep rise in the curve at low FPR values indicates that the model can detect a high proportion of anomalies with very few false alarms. The ROC analysis confirms that the proposed model achieves high sensitivity and specificity, making it highly reliable for data quality assessment in imbalanced datasets.

The training loss curve shows a smooth and consistent decrease in reconstruction loss over epochs asper the observation of Fig. 3. Initially, the loss is high due to random initialization of model parameters, then rapidly decreases as the model learns meaningful representations of the data. The curve gradually stabilizes after several epochs, indicating convergence. The absence of fluctuations or divergence suggests that the model is well-tuned and does not suffer from overfitting or instability. The stable convergence behaviour demonstrates that the model effectively learns the underlying data distribution and generalizes well to unseen data.

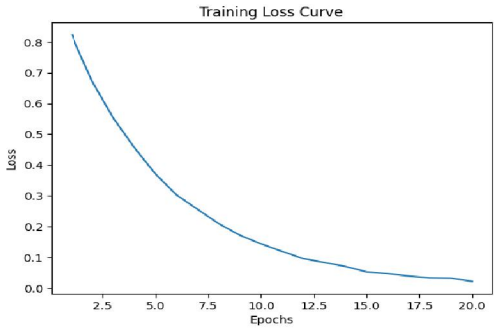

Fig. 3: Training Loss Curve Analysis

DOI: 10.48175/568

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 8, March 2026
Impact Factor: 8.2

The experimental results validate that the proposed model effectively signified the key challenges of data quality assessment. The deep autoencoder successfully captures nonlinear relationships belongs in high-dimensional data, enabling accurate anomaly detection. The distributed architecture ensures scalability and efficiency, performed well in situation of large-scale real-world applications. However, the model introduces moderate computational complexity due to deep learning operations, and the performance depends on proper hyperparameter tuning. Despite these limitations, the proposed framework provides a robust, scalable, and intelligent solution for data quality management.

Comparative Analysis

For the comparative study, we compare with several baseline and state-of-the-art approaches, including Logistic Regression, PCA-based anomaly detection, Isolation Forest, and hybrid deep learning models. The results indicate that traditional methods such as Logistic Regression struggle with imbalanced datasets and nonlinear relationships. PCA-based approaches perform moderately well but are limited by their linear assumptions. Isolation Forest improves detection accuracy whereas lacks of deep feature learning capability. Hybrid deep learning models achieve competitive results but often require higher computational resources. In contrast, the proposed deep autoencoder model achieves superior performance across all evaluation metrics, particularly in precision, recall, and AUC, while maintaining a balance between accuracy and computational efficiency.

Ablation Study

Conduct the ablation study, to evaluate the contribution of each component of the proposed model. Different variants of the model are tested by removing or modifying key components. According to the results show that removing normalization significantly degrades performance due to inconsistent feature scaling. Eliminating latent compression reduces the ability of model to capture essential data patterns. A fixed threshold instead of an adaptive threshold leads to suboptimal classification. Moreover, removing distributed processing affects scalability and increases processing time. Overall proposed model, which integrates all components, achieves the best performance, demonstrating the importance of each module in the overall framework.

Sensitivity Analysis

The sensitivity analysis used to conduct to evaluate the impact of key parameters on model performance. The results indicate that the covert dimension plays an important role in balancing compression and reconstruction accuracy. The optimal threshold is found to be around the 95th percentile of reconstruction error, providing the best trade-off between precision and recall. The learning rate affects convergence stability, with a value of 0.001 providing optimal results.

V. CONCLUSION

In this study we present a deep learning-based data quality assessment framework personalized for distributed cloud-native systems, where scalability and real-time processing are critical. According to the leveraging of deep autoencoder architecture, the model effectively learns the underlying patterns of high-quality data and identifies anomalies using reconstruction error, enabling accurate detection of inconsistencies, noise, and missing values. The experimental results demonstrate strong performance across all evaluation metrics, including high accuracy, precision, recall, F1-score, and AUC, confirming the model's capability to handle highly imbalanced and large-scale datasets. The graphical and analytical evaluations further validate the effectiveness of the proposed model and distinguishing between normal and anomalous data, while maintaining stable training and reliable convergence. Furthermore, the integration of distributed processing enhances computational efficiency and scalability, making the framework suitable for modern cloud-native environments. The ability of the model is to combine anomaly detection, data quality assessment, and adaptive correction within a unified system represents a significant improvement over traditional approach. The proposed framework provides a robust, scalable, and intelligent solution for data quality management, making it highly

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 8, March 2026
Impact Factor: 8.2

applicable to real-world data engineering and intelligent system applications, with promising scope for further enhancement in streaming and edge-based environments.

REFERENCES

[1]. Malhotra, Rashmi, and D. K. Malhotra. "The Impact of Technology, Big Data, and Analytics: The Evolving Data-Driven Model of Innovation in the Finance Industry." Journal of Financial Data Science 5, no. 3 (2023).

[2]. Zemicheal, Tadesse, and Thomas G. Dietterich. "Anomaly detection in the presence of missing values for weather data quality control." In Proceedings of the 2nd ACM SIGCAS Conference on Computing and Sustainable Societies, pp. 65-73. 2019.

[3]. Katru, Chandrasekhar Rao, "Next-Gen AI Quality Checks: Redefining Data Integrity in Automated Workflows." Journal of Information Systems Engineering and Management, 10(4s) e-ISSN: 2468-4376, 2024.

[4]. Katru, Chandrasekhar Rao, Sandip J. Gami, and Kevin N. Shah. "Automation for the Future: Harnessing AI and ML to Reshape Software Testing and Maintenance.", International Journal of Computer Trends and Technology, Volume 73 Issue 7, pp. 63-72, July 2025, ISSN: 2231–2803

[5]. Somayajula, Ramesh, et al. "Zero-ETL Architectures for AI Workloads Direct ML Model Access on Cloud-Native OLAP Systems." 2025 International Conference on Computing Technologies (ICOCT). IEEE, 2025.

[6]. ACHANTA, PADMA RAMA DIVYA. "Streamlining Multi-Database Workloads on Azure with Infrastructure Automation for SQL Server And Mongodb Using Terraform.", IRE Journals, Volume 6, Issue 6, 2022.

[7]. Pai, Rakesh Ramakrishna, and Jothsna Praveena Pendyala. "Optimizing Healthcare Pipelines for Patient Benefit: A Data Engineering Perspectives on Preauthorization Delays and Denials." International Conference on Software Engineering and Data Engineering. Cham: Springer Nature Switzerland, 2025.

[8]. Gami, Sandip J., et al. "Interactive data quality dashboard: Integrating real-time monitoring with predictive analytics for proactive data management." International Journal of Computer Sciences and Engineering 12.3 (2024): 89-98.

[9]. Gami, Sandip J., Rajesh Remala, and Krishnamurty Raju Mudunuru. "AI-Driven Adaptive Data Cleansing: Automating Error Detection and Correction for Dynamic Datasets." International Journal of Computer Trends and Technology 72.11 (2024): 159-164.

[10]. ACHANTA, PADMA RAMA DIVYA. "Overcoming Infrastructure Challenges in MongoDB On-Premises with Smart Design Patterns.", ICONIC RESEARCH AND ENGINEERING JOURNALS, Volume 6, Issue 6, 2022.

[11]. Si, Amalendu, Sujit Das, and Samarjit Kar. "Extension of TOPSIS and VIKOR method for decision-making problems with picture fuzzy number." In Proceedings of the Global AI Congress 2019, pp. 563-577. Singapore: Springer Singapore, 2020.

[12]. ACHANTA, PADMA RAMA DIVYA. "Redefining Enterprise Infrastructure with Scalable Architectures in Azure Hybrid Cloud." ICONIC RESEARCH AND ENGINEERING JOURNALS, Volume 6, Issue 10, 2023.

[13]. Katru, Chandrasekhar Rao, et al. "Enhancing Sovereign Allocation of Resources in Cloud Milieus Using a Causal Dilated Geometric Algebra Approach for Dynamic Scalability." 2025 International Conference on Machine Learning and Autonomous Systems (ICMLAS). IEEE, 2025.

[14]. Alang, Karan, et al. "Scalable cloud architectures for efficient processing of multi-structured big data." 2025 Global Conference in Emerging Technology (GINOTECH). IEEE, 2025.

[15]. Chouhan, Biky, Rakesh Pai, and Bishwajeet Pandey. "Edge computing based emulator design for low-latency IoT health monitoring system." International Journal of Information Technology 17.9 (2025): 5731-5741.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 8, March 2026
Impact Factor: 8.2

[16]. Borra, Chandrakanth Reddy, et al. "Enhancing IoT Network Security: Machine Learning-Based Intrusion Detection Using Vortex Search Optimization and SMOTE for Improved Classification." 2025 International Conference on Intelligent and Cloud Computing (ICoICC). IEEE, 2025.

[17]. Sundararamaiah, Munikrishnaiah, et al. "Similarity-Navigated Graph Neural Network-Based Fraud Detection in Credit Card Transactions Optimized with the Greylag Goose Algorithm." 2025 International Conference on Machine Learning and Autonomous Systems (ICMLAS). IEEE, 2025.

[18]. Cheekati, Srinivas, "Cybersecurity Threat Detection Using OpCyNet and DBRA: A Deep Learning Approach for DDoS Attack Mitigation on CICDDoS2019." 2025 13th International Conference on Smart Grid (icSmartGrid). IEEE, 2025.

[19]. Raghavan, Prathap, "Securing Data Engineering Pipelines in Cloud Environments Through Ai-Powered Intrusion Detection and Quality Assurance." 2025 International Conference on Computing Technologies & Data Communication (ICCTDC). IEEE, 2025.

[20]. Katru, Chandrasekhar Rao. "The Rise of Intelligent Finance: AI's Influence on the Financial Sector." The Impact of Artificial Intelligence on Finance: Transforming Financial Technologies. Cham: Springer Nature Switzerland, 2025. 273-298.

[21]. ACHANTA, PADMA RAMA DIVYA. "Streamlining Multi-Database Workloads on Azure with Infrastructure Automation for SQL Server And Mongodb Using Terraform." (2022).

[22]. Si, Amalendu, Sujit Das, and Samarjit Kar. "Preferred hospitalization of COVID-19 patients using intuitionistic fuzzy set-based matching approach." Granular Computing 8, no. 3, pp. 525-549, 2023.

[23]. Rayala, Ramya Vani, "Optimized Deep Learning for Diabetes Detection: A BGRU-based Approach with SA-GSO Hyperparameter Tuning." 2025 International Conference on Innovations in Intelligent Systems: Advancements in Computing, Communication, and Cybersecurity (ISAC3). IEEE, 2025.