International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 07, July 2026)

An Intelligent Clustering-Based Predictive Framework for Anomaly Detection and Data Quality Enhancement in System of Record (SOR) Systems

Dr. Amarasimha T.
Department of MCA, Aurora PG College, Nampally, Hyderabad, India

***Abstract*—System of Record (SOR) systems serve as the primary source of trusted enterprise data used for operational, analytical, and strategic decision-making. Poor data quality caused by duplicate records, missing values, invalid entries, and anomalous patterns can significantly affect organizational performance and governance. Existing approaches primarily focus on anomaly detection or data cleansing independently, resulting in reactive data quality management. This paper proposes an Intelligent Clustering-Based Predictive Framework (ICPF) for anomaly detection and data quality enhancement in SOR systems. The framework integrates K-Means clustering, DBSCAN-based outlier detection, predictive machine learning models, and a Data Quality Risk Score (DQRS) to identify, prioritize, and predict data quality issues. A simulated enterprise dataset containing customer, asset, and transaction records is used for an illustrative evaluation. The reported numerical results are therefore experimental example results and must be replaced by measured results from an executable implementation before journal submission. The framework is designed to provide proactive support for enterprise data governance and decision-making.**

***Keywords*—Anomaly Detection, Clustering, Data Governance, Data Quality, Machine Learning, Predictive Analytics, System of Record.**

I. Introduction

Modern organizations depend on System of Record (SOR) systems to maintain trusted information related to customers, assets, transactions, employees, and business operations. The quality of data stored in these systems directly influences business intelligence, regulatory compliance, operational efficiency, and customer satisfaction.

As organizations grow, the volume and complexity of enterprise data increase significantly. Manual validation techniques and rule-based approaches often fail to detect hidden anomalies and emerging data quality issues. Consequently, organizations experience duplicate records, incomplete information, invalid values, and inconsistent data across systems.

Machine learning techniques provide an opportunity to automate anomaly detection and improve data quality. Clustering algorithms can identify unusual patterns without requiring labeled datasets, making them suitable for enterprise environments where anomaly labels are rarely available. This research proposes an Intelligent Clustering-Based Predictive Framework that combines anomaly detection, predictive analytics, and data quality governance within a unified architecture.

II. Research Problem

Despite significant investments in enterprise data management, organizations continue to face data quality challenges. Common issues include duplicate customer records, missing contact information, invalid asset identifiers, inconsistent transaction records, and data synchronization failures.

Traditional data quality tools focus mainly on identifying existing issues rather than predicting future quality degradation. Furthermore, many anomaly detection solutions are not specifically designed for SOR environments. The lack of predictive capabilities can result in increased operational risk and delayed corrective actions.

III. Research Objectives

- Detect anomalies in SOR datasets using clustering techniques.
- Predict future data quality violations using machine learning.
- Develop a Data Quality Risk Score (DQRS) for prioritization.
- Improve enterprise data quality dimensions.
- Support proactive data governance initiatives.

IV. Literature Review And Research Gap

Existing studies have investigated anomaly detection using clustering algorithms, density-based methods, and machine learning techniques.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 07, July 2026)

K-Means clustering effectively groups similar records, while DBSCAN identifies sparse and anomalous observations. Predictive models such as Random Forest and Gradient Boosting can be used to forecast data-related events. However, the current study identifies four practical gaps: limited SOR-specific anomaly detection, insufficient integration of anomaly detection with future-risk prediction, limited support for data governance, and inadequate prioritization mechanisms for remediation.

The proposed ICPF addresses these gaps by combining clustering-based detection, predictive analytics, a risk scoring mechanism, and a recommendation layer within a single SOR-oriented architecture.

TABLE I. RESEARCH GAP AND PROPOSED RESPONSE

Existing Limitation	Impact	Proposed Response
General anomaly detection	Weak SOR context	SOR-oriented framework
Reactive quality checks	Late remediation	Predictive risk model
Detection without prioritization	High manual effort	DQRS risk categories
Separate quality controls	Fragmented governance	Integrated recommendation engine

V. Proposed Framework

A. Sample Enterprise Dataset

A simulated SOR dataset containing 100,000 records is used for the study design. The dataset represents customer, asset, and transaction information and includes Customer_ID, Customer_Name, Email, Phone, Asset_ID, Transaction_ID, Transaction_Amount, Transaction_Date, and Status attributes.

Artificial anomalies are introduced to represent common enterprise data quality problems: missing email addresses, duplicate customer records, invalid asset identifiers, extreme transaction values, and inconsistent status values.

B. Data Preprocessing

The preprocessing stage includes missing-value handling, duplicate detection, data normalization, feature engineering, and preparation of candidate outliers. Numerical variables are transformed to comparable scales before clustering.

C. Clustering-Based Anomaly Detection

K-Means clustering partitions records into groups based on feature similarity. For a record x and centroid y, the Euclidean distance is calculated as:

$$d(x,y) = \sqrt{\Sigma (x_i - y_i)^2}$$

Records with unusually large distances from their assigned centroids are treated as anomaly candidates. DBSCAN is then applied to identify dense regions and isolated noise points. DBSCAN is useful because it does not require a predefined number of clusters and can detect irregularly shaped clusters.

D. Data Quality Risk Score

The proposed Data Quality Risk Score (DQRS) combines multiple quality dimensions with the anomaly score to prioritize records for remediation:

$$DQRS = 0.25CE + 0.20CS + 0.20VE + 0.15DR + 0.20AS$$

Where CE is completeness error, CS is consistency error, VE is validity error, DR is duplication risk, and AS is the normalized anomaly score. The weighted score ranges from 0 to 100 after normalization.

TABLE II. DQRS RISK CATEGORIES

Score Range	Risk Category
0–25	Low Risk
26–50	Medium Risk
51–75	High Risk
76–100	Critical Risk

E. Predictive Analytics Module

Historical anomaly information is used to train predictive models such as Random Forest, XGBoost, and Gradient Boosting. The predictive layer estimates the probability of future anomaly occurrence, expected data quality degradation, and risk classification. The prediction output is combined with the DQRS to support proactive action.

F. Recommendation Engine

The recommendation layer maps detected quality problems to suggested actions, including record correction, duplicate removal, missing-value completion, standardization improvements, and governance review. Recommendations are intended to assist data stewards rather than automatically overwrite authoritative SOR records.

VI. System Architecture

Enterprise SOR Data
Data Collection Layer
Preprocessing and Cleansing
Feature Engineering
K-Means + DBSCAN Anomaly Detection
DQRS + Predictive Analytics Engine
Recommendation Engine + Data Governance Dashboard

Fig. 1. Proposed ICPF processing architecture.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 07, July 2026)

VII. Experimental Setup

The experimental design uses a simulated enterprise SOR dataset of 100,000 records: 50,000 customer records, 25,000 asset records, and 25,000 transaction records. Because the dataset is simulated for this manuscript, the results below are illustrative and should not be represented as measurements from a production SOR system.

The proposed approach is compared conceptually with Isolation Forest, Local Outlier Factor (LOF), One-Class SVM, and Autoencoder baselines. Evaluation metrics include accuracy, precision, recall, F1-score, ROC-AUC, and silhouette score.

TABLE III. ILLUSTRATIVE ANOMALY DETECTION RESULTS

Method	Precision	Recall	F1 Score
Isolation Forest	0.84	0.85	0.84
LOF	0.82	0.84	0.83
One-Class SVM	0.85	0.86	0.85
Proposed ICPF	0.91	0.93	0.92

VIII. Results And Discussion

Under the illustrative experimental scenario, the proposed ICPF records higher precision, recall, and F1-score than the selected baseline methods. The improvement is attributed to the complementary use of partition-based clustering, density-based detection, risk scoring, and predictive analytics. In an actual experiment, these values must be generated from a reproducible implementation using a fixed train/test protocol and clearly defined anomaly labels.

TABLE IV. ILLUSTRATIVE DATA QUALITY IMPROVEMENT

Metric	Before	After
Completeness	78%	96%
Consistency	74%	95%
Validity	82%	97%
Accuracy	80%	96%

The illustrative quality metrics indicate potential improvement after anomaly identification and remediation. However, these percentages are example values derived from the current manuscript and are not validated production measurements. For journal submission, the values should be replaced by results generated from the proposed algorithm and accompanied by statistical analysis.

IX. Research Contributions

- Development of an Intelligent Clustering-Based Predictive Framework for SOR data.
- Introduction of the Data Quality Risk Score (DQRS) for remediation prioritization.
- Integration of anomaly detection and future-risk prediction.
- Automated recommendation support for data governance.
- A scalable architecture for enterprise data quality management.

X. Conclusion

This paper presented an Intelligent Clustering-Based Predictive Framework for anomaly detection and data quality enhancement in System of Record systems. The integration of clustering techniques, predictive analytics, and Data Quality Risk Scoring provides a structured mechanism for proactive identification and management of enterprise data quality issues. The framework is designed to improve anomaly visibility, prioritize remediation, and support data governance. The current manuscript uses a simulated dataset and illustrative results; therefore, implementation-based validation is required before making definitive performance claims.

XI. Future Work

Future research will focus on real-time anomaly detection, explainable artificial intelligence, generative AI-assisted data correction, concept-drift handling, and integration with enterprise platforms such as ServiceNow, SAP, and Salesforce. Future experiments should also evaluate computational cost, scalability, robustness to class imbalance, and statistical significance across multiple real-world SOR datasets.

REFERENCES

[1] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 4th ed. Morgan Kaufmann, 2023.

[2] C. C. Aggarwal, Outlier Analysis, 3rd ed. Springer, 2024.

[3] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly Detection: A Survey,” ACM Computing Surveys, vol. 41, no. 3.

[4] ISO 8000, Data Quality Management Standards, International Organization for Standardization.

[5] Z. Fang et al., “Towards a Unified Framework of Clustering-Based Anomaly Detection,” Proc. International Conference on Machine Learning (ICML), 2025.