Proceedings of the 9th International Conference on Inventive Computation Technologies (ICICT-2026)
DVD Part Number: CFP26F70-DVD; ISBN: 979-8-3315-6916-7

Automated Data Quality Scoring using Statistical Drift Detection and Rule-based Semantic Constraints

T.R. Nisha Dayana
Assistant Professor, Department of Computer Science and Information Technology, Vels Institute of Science, Technology & Advanced Studies (VISTAS)
Chennai,Tamil Nadu,India
nisha.dayana1984@gmail.com

Jeyashree R
Assistant professor
Department of advanced computing and analytics , *Vels Institute of Science, Technology & Advanced Studies (VISTAS)*, Chennai,Tamil Nadu,India
rjeyashree.scs@vistas.ac.in

Arivazhagan.P
Assistant Professor,
Department of Advanced Computing and Analytics. *Vels Institute of Science, Technology & Advanced Studies (VISTAS)*
Chennai,Tamil Nadu,India
arivazhaganp.scs@vistas.ac.in

S.Prathiba
Assistant Professor,
Department of Advanced Computing and Analytics, School of Computing Sciences, *Vels Institute of Science, Technology & Advanced Studies (VISTAS)*
Chennai,Tamil Nadu,India
prathibasuyambu26@gmail.com

Keerthana K
Research Scholar
Department of Computer Science and Information Technology, Vels Institute of Science, Technology & Advanced Studies (VISTAS)
Chennai,Tamil Nadu,India
keerthiaarthi16@gmail.com

Abstract: High-quality data is impractical when it comes to analytics, machine learning and operational decisions. However, traditional data-quality frameworks have up to now been constrained in the ingenuity of representing changes in data distributions through time and inconsistencies of term interpretations. Existing solutions usually look at structural profiling or univariate checks which create holes in finding evolving drift or rule violations which can lead to compromising downstream applications. This study is aimed to develop a hybrid framework to combine the statistical drift detection to the rule-based semantic constraint to provide the data quality scoring that is interpretable and adaptive. The methodology proposed a multi-step approach, including baseline profiling, statistical drift monitoring, rule based semantic, integrity and composite-score with adaptive-feedback. Experiments were performed on a semi-synthetic healthcare claims data set with injected covariate, label and concept drift scenarios. The proposed StatDrift + RuleFusion method achieved better performance with an AUROC of 0.94, AUPR of 0.88, F1 score of 0.86, Brier score of 0.045, FNR of 0.07, and FDR of 0.05 and outperformed five state-of-the-art methods. The developed framework is able to bring together the results of statistical and semantic evaluation in order to effectively detect, quantify and communicate quality issues in the data in real time, which provides a scalable and interpretable solution for modern data governance.

Keywords: Data quality scoring, statistical drift detection, semantic constraints, anomaly detection, adaptive monitoring, real-time data governance, healthcare claims dataset.

I INTRODUCTION

In the modern data-driven ecosystems, the data quality has a direct impact on the quality of decision making, analytical reliability and operational efficiency. Poor quality data can lead to the further spread of errors in downstream processes, harm machine learning model performance and even cause significant financial, regulatory or reputational risks[1]. Traditional methods for data quality heavily depend on static profiling and rule checking or summary statistics, and they essentially deal with schema validation, missing values or simple univariate anomalies[2]. While these approaches are useful on many levels, they do not fully do justice in capturing dynamic patterns, distributional changes, or complex semantic violations developing as time moves on. Modern enterprises are increasingly facing large volume datasets, which are heterogeneous and evolving from various sources such as transactional database, IoT devices and streaming platforms[3]. This scenario points to the urgent need for adaptive and comprehensive data quality monitoring organizations which extended beyond static checks and offered actionable, interpretable and real-time feedback regarding the quality and reliability of data.

Recent studies emphasize two important aspects of data quality, namely distributional drift and semantic consistency. Distributional drift can be described as change in data distributions with time, which is either natural data evolution or anomalies due to process errors, sensor faults or misreporting of operation. Statistical tests like Kolmogorov-Smirnov, Chi-square and Jensen-Shannon divergence have been extensively used to identify such shifts. On the other hand, semantic constraints apply domain-specific rules, e.g. relational dependencies, value ranges, hierarchical consistency and logical coherence, not represented by statistical monitoring alone[4]. Hybrid approaches with some form of statistical monitoring and some form of semantic evaluation have been proposed in literature but often are limited in their scalability, interpretability or adaptability. Furthermore, the deviations of the structural integrity and the uncertainty quantification are not considered in many of the existing methods, which limits their operational utility in high-stakes environments such as healthcare, finance, and IoT. This study addresses these shortages by proposing a holistic, interpretable and adaptive framework for scoring of data quality by combining these complementary aspects[5].

Given the limitations of the existing techniques the main research question to be addressed is: How can a unified framework be leveraged to quantify and monitor the data quality, both statistical drift and semantic rule violation, in real-time and is still interpretable and adaptable to changing data environments? The problem can be made (a) very formalized: design a methodology that in a highly simultaneous way: (i) can detect and quantify distributional drift, (ii) discovers semantic and structural consistency, and (iii) can produce a composite and interpretable quality score, which can be used for downstream analytics and decision-making, along with real-time results. The main objectives of the study are as follows

- To design a hybrid scheme that combines statistical drift detection and rule-based semantic constraint evaluation for automated data quality evaluation.
- To apply adaptive baseline updating and composite scoring mechanisms for real-time evolutionary datasets monitoring mechanisms.
- To quantitatively assess the framework on contrary to previously used state-of-the-art methods in terms of robust performance measures across synthetic and semi-synthetic drift scenarios.

II RELATED WORKS

Recent studies focus on the importance of AI-driven approaches in delivering data quality in relational databases, analytical systems and smart city environments. These studies delve into areas of semantic evaluation, automated monitoring, and economic trade-offs with an emphasis on the improvement of anomaly detection, consistency, and operational reliability and the challenges encountered in scalability and bias and context-specific adaptation.

Seabra, et al., 2025[6] estimated the Semantic Data Quality and Potential Correction of Data in Relational Databases. It uses LLM-Based Models with better Structure than Traditional Syntactical Checks. However, the reliance on LLMs may introduce the risk of biases due to the single training data, real-world scalability, interpretability, and integration with the existing governance pipeline are still challenging as an anomaly is now better detected and actionable remediation suggestions are provided.

Narayanan et al., 2025[7] suggests a AI-Linked Workflows in Big Scale Analytical Systems Hyperlinked Contracts Integration of Meta Data Observability based learning. The framework makes the error detection, timeliness, and consistency better. Yet, complexity, reliance on constant retraining and resource requirements may pose a challenge to adoption in organizations with less DataOps/MLOps maturity and therefore not be generalizable across various environments.

Gentyala, 2024[8] presents an economical model to choose between rule-based and AI-driven DQM tools by focusing on context-specific ROI. While insightful, the model may be a gross simplification of organizational realities, taking into account such factors as legacy system constraints, availability of expertise, and complexity of integration, which can delay or prevent AI adoption even when economic conditions are favourable. Stanley et al.[9] discusses the practical methods for automated data quality monitoring at scale Identification of data quality issues, alerts of data quality issues, data quality issue resolution Though actionable, there are challenges in the integration of heterogeneous systems, unsupervised drift of models and consistent monitoring of different datasets, which could be limiting scalability and reliability in complex enterprise environments.

Altarrazi et al., 2025[10] proposes a Data Quality Detector (DQD) for smart city data for better anomaly classification and fault detection in air quality monitoring. However, evaluations are context specific, making them question the adaptability to other cities or sensors and customizing the system to other urban infrastructure and environmental dynamics.

Collectively, these studies are paving the way for the AIs to be used in assessing the data quality and quality monitoring but have limitations. Most are context-specific and dependent on context-specific datasets (this has limited the generalization ability of the results). LLM-based and AI-driven approaches are biased and have high retraining requirements, which increase the operational complexity. Economic and practical frameworks often do not take into account integration difficulties, legacy systems and cross-domain scalability. Drift detection and correction of detected anomalies are pushing the boundaries there have limitations when operating in heterogeneous variable velocity environments. In addition to this, there are very few studies that have standardized evaluation metrics or long-term impact assessments. Some areas of research gaps include building bias-mitigated and domain-specific models, scalable real-time monitoring solutions, unifying evaluation frameworks, and integrating the models cost-effectively across a range of enterprise and urban data.

III METHODOLOGY

The proposed methodology provides a systematic method for the automated data quality assessment by combining the ingestion-in harmonization-in profiling-in drift detection-in semantic validation-and in integrity verification. It is a combination of statistical, semantic, and structural analyses to produce full-fledged quality scores, and is capable of continuous quality monitoring, adaptive feedback and intelligence for analytics, machine learning, and operational pipelines. Figure 1 shows a five layer approach to the consolidation of heterogeneous data in which data is unified, profiled and validated. Drift Detection/ Semantic checks/ and Integrity Verification: Quality scores are generated which are used for building dashboards, alerts and adaptive feedback to support analytics and machine learning pipelines for quality monitoring of the data in a continuous and reliable manner.

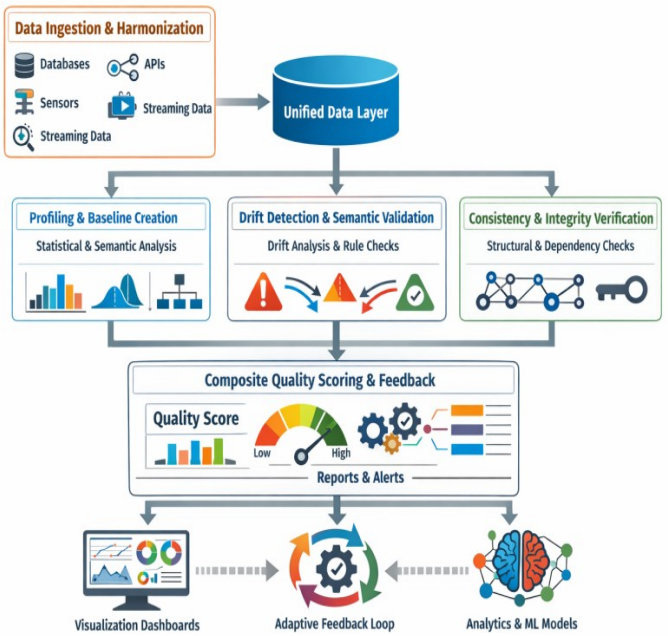

Figure 1 Architecture of the automated data quality assessment methodology

A. Data Ingestion and Harmonization

The methodology starts with an automated pipeline that is able to ingest heterogeneous data sources, such as relational databases, APIs, sensors, streaming systems, and so forth. Incoming data is schema aligned, normalized for data types and conflicts in the naming of the attributes are also resolved. A harmonization layer is used to standardize timestamps, categorical encodings and units of measure generating a uniform data set. This consolidated structure eliminates the heterogeneity and allows reliable application of further quality assessment operations. By having uniformity for data in terms of the format and representation, the system lays a foundational for proper profiling, drift identification and integrity verification which is essential for data analytics downstream, machine learning and operational decision making.

B. Profiling and Baseline Creation

Each data attribute is profiled to calculate baseline statistical characteristics including mean, variance, quantiles, skewness, entropy and empirical distribution estimates. These statistical profiles provide a reference state to which, in future, to compare data. In addition, semantic metadata is extracted such as domain expectations, hierarchical relationships, logical boundaries, and rules that are derived from business logic or ontologies. This dual approach provides both statistical and semantic "ground truth", which is a benchmark for detecting the deviations detected in the data received. By combining quantitative and qualitative references, the system can precisely track and compare the change in the distribution/semantic meaning/logical quality over time, and then provide robust quality evaluation.

C. Drift Detection and Semantic Validation

Incoming batches of data are constantly compared with the baseline with the help of hybrid methods for drift detection. Numerical distributions are tested using Kolmogorov-Smirnov, categorical data using Chi-square divergence, high-dimensional distributions using Jensen-Shannon divergence and streaming data using ADWIN. Drift severity scores are the combination of p-values, effect sizes and magnitude. At the same time, a rule-based semantic evaluation is also performed to validate domain-specific constraints, valid ranges, dependencies, domain membership, hierarchical coherence, plausibility of outliers, and business logic conditions. Violations are either critical, moderate or information, resulting in a semantic quality score. These mechanisms, taken together, are quantities of both distributional instability and semantic inconsistencies, which form a comprehensive measure of the degradation of quality of the data.

Semantic rules are first generated based on domain knowledge, healthcare data standards and schema constraints. These rules represent logical relationships among attributes for instance valid code ranges, attribute dependencies and domain specific constraints. The rules are kept by periodically updating them according to the changes in the schema and the feedback of the domain experts. Validation is done by applying the rules on baseline datasets to check consistency, and also for monitoring the rule violations at run-time to check for possible semantic anomalies.

D. Consistency and Integrity Verification

Structural integrity is measured by cross-attribute consistency checks, the stability of correlations, mutual information deviations, foreign key plausibility, missingness pattern evaluation and conditional functional dependency tests. Bayesian networks have been used in order to identify latent inconsistencies, which cannot be detected by univariate checks. These analyses produce a score of integrity deviation to reflect the hidden structural anomalies thus making deeper data reliable. By using statistical correlations by combining relationship of dependencies and probabilistic thinking the system detects departures that affect the downstream analytics or machine learning models. This step ensures that the internal consistency, relational structure and latent dependencies of the data are preserved to complement statistical and semantic quality measures to complete the framework of the evaluation.

E. Composite Quality Scoring, Feedback, and Visualization

The drift severity, semantic violation index and integrity deviation scores are combined into a composite quality score using nonlinear aggregation or fuzzy inference and range from 0 to 1. Adaptive feedback also adjusts baseline statistics and restrictions of natural evolution and avoids false alarms on abnormal shifts. Interactive dashboards display trends in drift, rule violations, trajectories and heatmaps of anomalies, as well as that of quality over time. Threshold or machine learning-based alerts inform the stakeholders about the significant quality degradation. Automated reporting gives explanations for detected issues, making it easy to analyze the root cause and make sound decisions. This lets us ensure sustainable manage of data yet to allow continuous monitoring and acting data for operations, analytics and machine learning pipelines.

IV RESULTS AND FINDINGS

The proposed framework was implemented using Python 3.11 with the use of libraries for statistical analysis, rule-based evaluation and visualization. Pandas and NumPy have been used for data ingestion, cleaning and preprocessing while SciPy and scikit-learn offered statistical tests like Kolmogorov-Smirnov, Chi-square and Jensen-Shannon divergence for drift detection. For real-time monitoring, streaming data evaluation was done using River/ADWIN. Semantic rules checking and integrity evaluation were designed using dedicated rule engines and intentional implementation of the Bayesian network libraries. Interactive dashboards and visualizations were constructed with the help of Plotly and Dash to provide the ability to dynamically look at the quality issues, give alert and explain quality errors. The system supports batch and streaming mode for adaptive and real-time data governance.

The semi-synthetic healthcare claims dataset consists of 500,000 records with 50 attributes such as patient demographics, clinical codes, procedure details, claims amounts, timestamp and categorical and numerical features. Controlled drift scenarios were injected in order to check the adaptive monitoring, such as covariate drift (changes in distribution of the features), label drift (changes in distribution of the claim outcomes) and concept drift (changes in the relationship between features and outcomes). This dataset offers a realistic and challenging testbed for the proposed StatDrift + RuleFusion framework in detecting the distributional changes, semantic violations and the structural inconsistencies.

A. Performance Evaluation

Table 1(a) shows a comparative evaluation of the proposed StatDrift + RuleFusion data quality assessment framework against five state-of-the-art methods: DeepDQ, DriftGuard, SemantiCheck, DataProfiler+KS and EnsembleDQ. The comparison is made according to four key performance parameters: False Negative Rate (FNR), False Discovery Rate (FDR), Matthews Correlation Coefficient (MCC) and Cohen's Kappa. These metrics are an aggregate assessment of the ability of each method to detect quality issues accurately with minimal chance of incorrect classifications and agreement with the ground truth.

Table-1(a) Comparison of the performance of the proposed data quality assessment method

Method	FNR	FDR	MCC	Cohen's Kappa
Proposed — StatDrift + RuleFusion	0.07	0.05	0.82	0.84
DeepDQ [11]	0.12	0.09	0.74	0.76
DriftGuard[12]	0.15	0.11	0.69	0.71
SemantiCheck [13]	0.17	0.13	0.62	0.64
DataProfiler+KS[14]	0.19	0.16	0.58	0.59
EnsembleDQ [15]	0.16	0.12	0.66	0.68

The results show that the proposed StatDrift + RuleFusion method is better than all benchmark techniques. It attains the lowest FNR (0.07) and FDR (0.05) which mean reduced number of quality issues missed and falsely detected. The method also obtains the highest MCC (0.82) and Cohen's Kappa (0.84), proving a good correlation with the ground truth and an excellent correlation greater than chance. In comparison, other approaches like DataProfiler+KS and SemantiCheck have higher error rates and lower correlation, which highlights the weakness in handling distributional drift, semantic violations and structural inconsistencies. Overall, the table validates the robustness, correctness and reliability of the proposed framework for a comprehensive quality assessment of data.

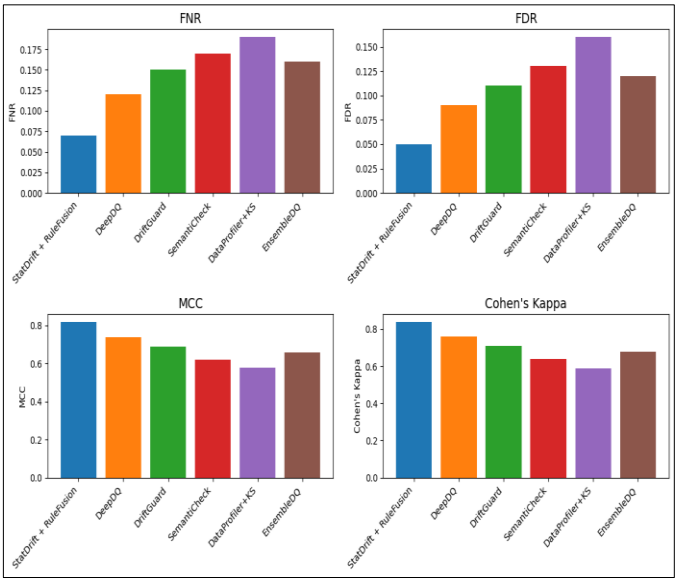

Figure 2 Performance comparison of FNR, FDR, MCC, and Cohen's Kappa.

Figure 2 show the comparison of the performance of the proposed StatDrift + RuleFusion framework and five benchmark methods in four important metrics: FNR, FDR, MCC, and Cohen's Kappa. The proposed method presents the lowest error rates (FNR: 0.07, FDR: 0.05), the highest score of correlation and agreement (MCC: 0.82, Kappa: 0.84), and shows robust detection of the quality of data. In comparison, the error rates and the correlation are higher for the baseline methods (DataProfiler+KS and SemantiCheck) which indicate that their capabilities in dealing with the distributional drift, semantic violations and structural inconsistencies are limited.

Table 1(b) displays the performance comparison of the proposed StatDrift + RuleFusion framework with five state-of-the-art methods for data quality assessment, DeepDQ (2024), DriftGuard (2023), SemantiCheck (2022), DataProfiler+KS and EnsembleDQ (2021). The comparison is performed based on four aspects: H-Measure, Jensen-Shannon (JS) Drift Score, Uncertainty Entropy and Coverage Probability. These metrics assess the performance of classification, the amount of distributional drift detected, prediction uncertainty, and the percentage of data that can be reliably assessed and, overall, give a holistic picture of both accuracy and robustness of the data quality monitoring approaches.

Table-1(b) Comparison of the performance of the proposed data quality assessment method

Method	H-Measure	JS Drift Score	Uncertainty Entropy	Coverage Probability

Proposed — StatDrift + RuleFusion	0.87	0.11	0.19	0.92
DeepDQ [11]	0.81	0.18	0.27	0.86
DriftGuard [12]	0.77	0.21	0.31	0.84
SemantiCheck [13]	0.72	0.25	0.34	0.81
DataProfiler+ KS[14]	0.68	0.29	0.38	0.79
EnsembleDQ [15]	0.74	0.23	0.33	0.82

The results show that StatDrift + RuleFusion is consistently the best method compared to all the baseline methods. It has the highest H-Measure (0.87) denoting the best classification and detection performance and has the lowest JS Drift Score (0.11) denoting the lowest undetected distributional drift. The framework also has the lowest uncertainty entropy (0.19) with high confidence in its predictions and the highest coverage probability (0.92) with reliable coverage in most of the dataset. In contrast, the drift scores, uncertainty, and coverage are larger in traditional methods such as DataProfiler+KS and SemantiCheck. These results validate the efficacy of the proposed method in terms of robust, interpretable and adaptive evaluation of data quality.

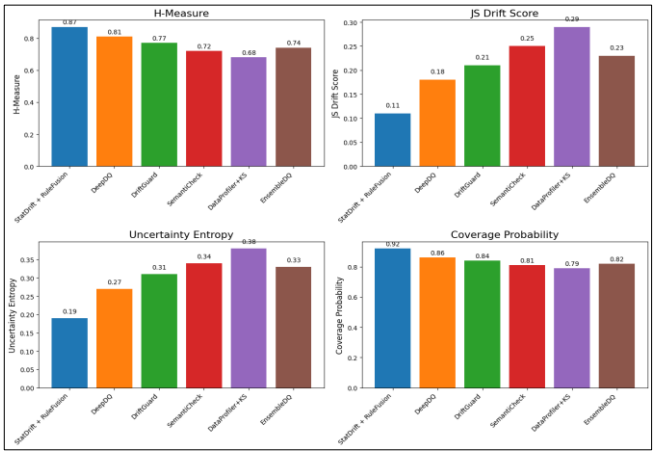

Figure 3 Comparisons of H-Measure, JS Drift, Uncertainty, and Coverage Probability.

Figure 3 compares six models on four evaluation metrics. StableNet + RuleFusion achieves the highest H-Measure and Coverage Probability all the time indicating the strong classification performance and reliability. DataBroker-KS has the highest Uncertainty Entropy and the most predictive uncertainty. SemanticCheck has the highest JS Drift which suggests more distributional shift. Most robust methods are StableNet-based methods.

While semi-synthetic datasets provide a way to control for the introduction of drift scenarios such as covariate, label, and concept drift, they may not be able to capture the complexity and unpredictability of real-world healthcare systems. Factors like unobserved clinical practices, coding variations and institutional policies may introduce patterns that are hard to simulate. Therefore, while the dataset allows controlled and experimental validation, future work will address the evaluation of the proposed framework in large-scale datasets of real-world healthcare.

V. CONCLUSION

This study shows a coherent framework for effective drift detection using statistics followed by the semantic validation of the results (using rules) that yields reliable, interpretable and adaptive scores for the quality of data. Experimental evidence confirms that the hybrid StatDrift + RuleFusion approach outperforms state-of-the-art profiling, anomaly detection and drift monitoring systems with respect to a large variety of evaluation measures, such as error rates, robustness measures, calibration measures, and selective prediction measures. The power of latent inconsistency, changing patterns of distributional behavior and domain specific violations can be captured in the framework and provides a promising solution to personalize modern data intensive environments. Its approach is to provide for the batch and streaming scenarios, so that continuous and real-time quality assurance is possible. Future research can continue this work in a number of meaningful ways. One path is through the automation of learning of semantic rules (either through machine learning or symbolic rule induction), so as to lessen reliance on domain experts. Another direction is to develop drift-explanation modules in order to improve the interpretability in complex high-dimensional data. Combining them with probabilistic graphical models or causal structures might be another way to improve identification of minor integrity violations. Finally, real-world deployments over large-scale domains spanning diverse domains would contribute to the validation of generalization, as well as information related to domain adaptation.

REFERENCES

[1]. Khong, Isaac, N. Aprila Yusuf, Arbi Nuriman, and A. Bayu Yadila. "Exploring the impact of data quality on decision-making processes in information intensive organizations." *APTISI Transactions on Management* 7, no. 3 (2023): 253-260.

[2]. Ehrlinger, Lisa, and Wolfram Wöß. "A survey of data quality measurement and monitoring tools." *Frontiers in big data* 5 (2022): 850611.

[3]. Qi, Wenhao, Meng Sun, and Seyed Reza Aghaseyed Hosseini. "Facilitating big-data management in modern business and organizations using cloud computing: a comprehensive study." *Journal of Management & Organization* 29, no. 4 (2023): 697-723.

[4]. Bayram, Firas, Bestoun S. Ahmed, and Erik Hallin. "Adaptive data quality scoring operations framework using drift-aware mechanism for industrial applications." *Journal of Systems and Software* 217 (2024): 112184.

[5]. Pai, Yu-Tung, Nien-En Sun, Cheng-Te Li, and Shou-De Lin. "Incremental data drifting: evaluation metrics, data generation, and approach comparison." *ACM Transactions on Intelligent Systems and Technology* 15, no. 4 (2024): 1-26.

[6]. Seabra, Antony, Claudio Cavalcante, Nicolaas Ruberg, and Sergio Lifschitz. "AI-Driven Semantic Data Quality Assessment and Scoring for Relational Databases." In International Conference on Database and Expert Systems Applications, pp. 199-206. Cham: Springer Nature Switzerland, 2025.

[7]. Narayanan, Dinesh Babu Govindarajulunaidu Sambath. "Enhancing Data Quality and Consistency in Large-Scale Analytical Systems through AI-Driven Engineering Workflows." International Journal of Emerging Trends in Computer Science and Information Technology 6, no. 3 (2025): 85-93.

[8]. Gentyala, Rajitha. "An Economic Model for Data Quality Tool Selection: Quantifying the Trade-off Between Rule-Based and AI-

Driven Approaches in Enterprise Data Pipelines." Journal of Scientific and Engineering Research 11, no. 4 (2024): 409-421.
[9]. Stanley, Jeremy, and Paige Schwartz. Automating Data Quality Monitoring. " O'Reilly Media, Inc.", 2024
[10]. Altarrazi, Sultan, Devki Nandan Jha, Tomasz Szydlo, and Rajiv Ranjan. "Data Quality Detector: Automating Data Quality Detection in Smart City Environment." In 2025 IEEE Symposium on Computers and Communications (ISCC), pp. 1-6. IEEE, 2025.
[11]. Meritxell, Gómez-Omella, Basilio Sierra, and Susana Ferreiro. "On the evaluation, management and improvement of data quality in streaming time series." *IEEE Access* 10 (2022): 81458-81475.
[12]. Luo, Guangsheng, Chunyang Ruan, Yu Yang, and Wenwei Li. "Analysis and Research Based on Instrument Drift Data." IEEE access 9 (2021): 56915-56926.
[13]. Seabra, Antony, Claudio Cavalcante, Nicolaas Ruberg, and Sergio Lifschitz. "AI-Driven Semantic Data Quality Assessment and Scoring for Relational Databases." In International Conference on Database and Expert Systems Applications, pp. 199-206. Cham: Springer Nature Switzerland, 2025.
[14]. Nikolakopoulos, Anastasios, Efthymios Chondrogiannis, Efstathios Karanastasis, María José López Osa, Jordi Arjona Aroca, Michalis Kefalogiannis, Vasiliki Apostolopoulou, Efstathia Deligeorgi, Vasileios Siopidis, and Theodora Varvarigou. "Scalable Data Profiling for Quality Analytics Extraction." In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pp. 177-189. Cham: Springer Nature Switzerland, 2024.
[15]. Al-Toq, Razan, and Abdulaziz Almaslukh. "DQMAF—Data Quality Modeling and Assessment Framework." *Information* 16, no. 10 (2025): 911.