Power System Technology
ISSN:1000-3673
Received: 16-11-2025 Revised: 05-12-2025 Accepted: 25-01-2026

Automated Data Quality Assessment Using Ensemble Machine Learning Techniques

Sanjay Kumar Mishra

Research Scholar (Computer Science), Mewar University, Chittorgarh

Email- skmjava@gmail.com

Prof. (Dr.) Chandikaditya Kumawat

Supervisor, Department of Computer Science, Mewar University, Chittorgarh

Email- chandikadityakumawat@mewaruniversity.co.in

ABSTRACT

In the era of big data, ensuring data quality has become paramount for organizations seeking to derive meaningful insights from their datasets. This research paper presents a comprehensive study on automated data quality assessment using ensemble machine learning techniques. We explore the integration of multiple ML algorithms to identify data quality issues including missing values, outliers, inconsistencies, and integrity violations. Our proposed framework combines Random Forest, Gradient Boosting, and Neural Network classifiers to achieve superior accuracy (94.7%) compared to individual models. The study demonstrates that ensemble methods significantly outperform traditional rule-based approaches and single ML algorithms in detecting complex data quality patterns. We evaluate our approach on five real-world datasets from healthcare, finance, and e-commerce domains, showing consistent improvements in precision, recall, and F1-scores. The findings suggest that ensemble ML techniques provide a robust, scalable solution for automated data quality management in diverse organizational contexts.

Keywords: *Data Quality, Machine Learning, Ensemble Methods, Automated Assessment, Random Forest, Gradient Boosting, Data Cleansing*

1. INTRODUCTION

1.1 Background and Motivation

Data quality is a critical factor that determines the success of data-driven decision-making processes in modern organizations. Poor data quality can lead to flawed analytics, incorrect business insights, and substantial financial losses. According to recent industry reports, organizations lose an average of $15 million annually due to poor data quality. Traditional approaches to data quality assessment rely heavily on manual inspection and rule-based

Power System Technology
ISSN:1000-3673
Received: 16-11-2025 Revised: 05-12-2025 Accepted: 25-01-2026

validation, which are time-consuming, error-prone, and fail to scale with the exponential growth of data volumes.

The emergence of machine learning (ML) technologies has opened new avenues for automating data quality assessment. ML algorithms can learn complex patterns in data and identify quality issues that may not be apparent through rule-based systems. However, individual ML models often have limitations in capturing the diverse nature of data quality problems. This research investigates the application of ensemble machine learning techniques, which combine multiple models to achieve superior performance and robustness.

1.2 Research Objectives

The primary objectives of this research are:

- To develop an automated data quality assessment framework using ensemble machine learning techniques
- To compare the performance of ensemble methods against individual ML algorithms and traditional rule-based approaches
- To evaluate the framework on real-world datasets from multiple domains
- To identify the most effective ensemble configurations for different types of data quality issues
- To provide practical recommendations for implementing automated data quality assessment in organizational settings

1.3 Research Contributions

This research makes several significant contributions to the field of automated data quality assessment. First, we propose a novel ensemble framework that integrates multiple machine learning algorithms specifically tailored for data quality detection. Second, we introduce a comprehensive taxonomy of data quality dimensions and their ML-based detection methods. Third, we provide extensive empirical evaluation across diverse domains, demonstrating the generalizability of our approach. Finally, we offer practical guidelines for practitioners seeking to implement automated data quality solutions.

2. LITERATURE REVIEW

2.1 Data Quality Dimensions

Data quality is a multi-dimensional concept that encompasses various aspects of data fitness for use. Wang and Strong (1996) identified 15 data quality dimensions grouped into four categories: intrinsic, contextual, representational, and accessibility. For the purpose of this research, we focus on the following six critical dimensions:

Power System Technology
ISSN:1000-3673
Received: 16-11-2025 Revised: 05-12-2025 Accepted: 25-01-2026

- Completeness: The extent to which data is not missing and is sufficient for the task at hand
- Accuracy: The degree to which data correctly represents the real-world entities or events
- Consistency: The absence of contradictions within the dataset or across multiple datasets
- Timeliness: The degree to which data is up-to-date and available when needed
- Validity: The extent to which data conforms to defined business rules and constraints
- Uniqueness: The absence of duplicate records representing the same real-world entity

2.2 Traditional Approaches to Data Quality Assessment

Traditional data quality assessment methods primarily rely on rule-based validation and manual inspection. These approaches involve defining explicit rules and constraints that data must satisfy. While effective for well-defined quality criteria, rule-based systems have significant limitations. They require extensive domain expertise to define comprehensive rules, struggle with evolving data patterns, and fail to detect subtle or complex quality issues. Moreover, maintaining rule sets becomes increasingly challenging as data sources and business requirements evolve.

2.3 Machine Learning for Data Quality

Recent research has explored various machine learning techniques for automated data quality assessment. Supervised learning approaches, including decision trees, support vector machines, and neural networks, have shown promise in detecting data quality issues when labeled training data is available. Unsupervised methods such as clustering and anomaly detection are useful for identifying outliers and unusual patterns without prior labeling. However, most existing studies focus on individual ML algorithms, with limited exploration of ensemble methods that combine multiple models for improved performance.

2.4 Ensemble Machine Learning Techniques

Ensemble learning is a machine learning paradigm that combines multiple models to achieve better predictive performance than any single model. The three main ensemble strategies are bagging, boosting, and stacking. Bagging methods like Random Forest reduce variance by training multiple models on different data subsets. Boosting algorithms such as Gradient Boosting and XGBoost sequentially train models to correct errors of previous models, reducing bias. Stacking combines predictions from multiple models using a meta-learner. While ensemble methods have proven highly effective in various domains, their application to data quality assessment remains relatively unexplored.

Power System Technology
ISSN:1000-3673
Received: 16-11-2025 Revised: 05-12-2025 Accepted: 25-01-2026

3. METHODOLOGY

3.1 Proposed Framework Architecture

Our proposed framework consists of five main components: Data Profiling Module, Feature Engineering Module, Base Learners, Ensemble Integration Layer, and Quality Assessment Output. The Data Profiling Module analyzes datasets to extract statistical properties and metadata. The Feature Engineering Module generates relevant features for quality detection, including statistical measures, pattern indicators, and domain-specific attributes. Base Learners comprise three primary algorithms: Random Forest, Gradient Boosting Machine (GBM), and Deep Neural Network (DNN). The Ensemble Integration Layer combines predictions from base learners using weighted voting and stacking techniques. Finally, the Quality Assessment Output provides comprehensive quality reports with issue identification and severity rankings.

3.2 Feature Engineering

Effective feature engineering is crucial for ML-based data quality assessment. We developed a comprehensive set of 47 features organized into five categories: Statistical Features (mean, median, standard deviation, skewness, kurtosis), Pattern Features (regex matches, format consistency, character distribution), Relational Features (foreign key violations, referential integrity), Temporal Features (timestamp consistency, sequence patterns), and Domain-Specific Features (business rule compliance, valid ranges). Features are automatically computed for each data attribute and normalized before model training.

3.3 Base Learning Algorithms

3.3.1 Random Forest Classifier

Random Forest is an ensemble of decision trees trained on random subsets of features and data samples. We configured our Random Forest with 200 trees, maximum depth of 15, and minimum samples per leaf of 5. This configuration balances model complexity with generalization capability. Random Forest excels at capturing non-linear relationships and provides feature importance rankings, which help identify the most critical factors for data quality assessment.

3.3.2 Gradient Boosting Machine

Gradient Boosting Machine sequentially builds an ensemble by training new models to correct errors of existing models. We implemented GBM with 150 estimators, learning rate of 0.1, and maximum depth of 6. The algorithm uses gradient descent to minimize a loss function, making it particularly effective for complex classification tasks. GBM demonstrates superior

Power System Technology
ISSN:1000-3673
Received: 16-11-2025 Revised: 05-12-2025 Accepted: 25-01-2026

performance on imbalanced datasets, which are common in data quality scenarios where quality issues may be rare events.

3.3.3 Deep Neural Network

Our DNN architecture consists of four hidden layers with 128, 64, 32, and 16 neurons respectively, using ReLU activation functions. We apply dropout regularization (rate=0.3) to prevent overfitting and use batch normalization for stable training. The output layer uses softmax activation for multi-class quality issue classification. DNNs capture complex non-linear patterns and interactions between features, complementing the strengths of tree-based methods.

3.4 Ensemble Integration Strategies

We evaluated two ensemble integration strategies:

- Weighted Voting: Each base learner provides a prediction with an associated confidence score. Final prediction is determined by weighted majority voting, where weights are proportional to each model's validation accuracy.
- Stacking: Base learner predictions are used as input features for a meta-learner (Logistic Regression). The meta-learner learns to optimally combine base model outputs, potentially capturing complementary information.

4. EXPERIMENTAL SETUP

4.1 Datasets

We evaluated our framework on five real-world datasets from diverse domains. Each dataset was selected to represent different data quality challenges and business contexts. The datasets vary in size, structure, and the types of quality issues they contain, providing a comprehensive test of our approach's generalizability.

Table 1: Dataset Characteristics

Dataset	Domain	Records	Attributes	Quality Issues	Size (MB)
Healthcare-EHR	Healthcare	125,000	34	Missing values, outliers, inconsistencies	45.2

Power System Technology
ISSN:1000-3673
Received: 16-11-2025 Revised: 05-12-2025 Accepted: 25-01-2026

Financial-Trans	Finance	500,000	28	Duplicates, format errors, invalid values	182.7
E-commerce-Orders	E-commerce	250,000	42	Missing data, referential integrity	95.3
Manufacturing-IoT	Manufacturing	1,000,000	18	Sensor errors, time gaps, outliers	215.6
Customer-CRM	Marketing	75,000	56	Duplicates, inconsistencies, missing values	38.9

Each dataset underwent preprocessing to create labeled training data. Domain experts manually annotated samples to identify data quality issues, creating ground truth labels for supervised learning. We employed stratified sampling to ensure balanced representation of different quality issue types.

4.2 Evaluation Metrics

We employed multiple evaluation metrics to comprehensively assess model performance:

- Accuracy: Overall correctness of predictions across all classes
- Precision: Proportion of true positive predictions among all positive predictions
- Recall: Proportion of true positive instances correctly identified
- F1-Score: Harmonic mean of precision and recall
- AUC-ROC: Area under the receiver operating characteristic curve
- Training Time: Computational efficiency measured in seconds

4.3 Experimental Protocol

We employed 10-fold cross-validation to ensure robust evaluation and minimize overfitting. Each dataset was randomly partitioned into 10 subsets, with 9 used for training and 1 for testing in each iteration. This process was repeated for all possible combinations. We calculated mean performance metrics across all folds along with standard deviations to assess result stability.

Power System Technology
ISSN:1000-3673
Received: 16-11-2025 Revised: 05-12-2025 Accepted: 25-01-2026

All experiments were conducted on identical hardware (Intel Xeon CPU, 64GB RAM, NVIDIA GPU) to ensure fair comparison of computational requirements.

5. RESULTS AND ANALYSIS

5.1 Overall Performance Comparison

Table 2 presents the overall performance comparison across different approaches. The results demonstrate that ensemble methods significantly outperform both traditional rule-based systems and individual machine learning algorithms. Our proposed ensemble framework achieves an average accuracy of 94.7%, representing a 12.3% improvement over rule-based methods and 4.2% improvement over the best individual ML algorithm (GBM).

Table 2: Performance Comparison of Different Approaches

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC
Rule-Based	82.4	79.8	81.2	80.5	0.823
Decision Tree	85.6	84.2	85.9	85.0	0.867
Random Forest	89.3	88.7	89.8	89.2	0.912
SVM	87.8	86.5	88.1	87.3	0.895
GBM	90.5	89.9	91.2	90.5	0.925
Neural Network	88.9	87.6	89.4	88.5	0.903
Ensemble (Voting)	93.2	92.8	93.6	93.2	0.951
Ensemble (Stacking)	94.7	94.3	95.1	94.7	0.964

Power System Technology
ISSN:1000-3673
Received: 16-11-2025 Revised: 05-12-2025 Accepted: 25-01-2026

The stacking ensemble consistently outperforms weighted voting across all metrics. This suggests that the meta-learner effectively captures complementary information from base models, optimizing the combination of their predictions. The high AUC-ROC score (0.964) indicates excellent discrimination capability across different quality issue severities.

5.2 Performance by Data Quality Dimension

We analyzed performance across different data quality dimensions to understand which types of issues are most effectively detected by our ensemble approach. Table 3 shows F1-scores for each quality dimension.

Table 3: F1-Scores by Data Quality Dimension

Quality Dimension	Rule-Based	Random Forest	GBM	Ensemble (Stacking)
Completeness	88.3	92.1	93.5	96.2
Accuracy	76.5	85.7	87.9	92.8
Consistency	79.2	88.4	90.1	94.5
Timeliness	82.7	89.6	91.3	95.1
Validity	84.1	90.3	92.2	95.7
Uniqueness	81.6	87.9	89.8	93.4

The ensemble approach shows substantial improvements across all quality dimensions, with the largest gains observed for accuracy and consistency detection. These dimensions often involve complex, context-dependent patterns that benefit most from the ensemble's ability to combine multiple perspectives. Completeness detection achieves the highest F1-score (96.2%), as missing value patterns are relatively straightforward to identify with appropriate features.

5.3 Domain-Specific Performance

Performance varies across different application domains due to differences in data characteristics and quality issue types. Table 4 presents accuracy results for each dataset.

Power System Technology
ISSN:1000-3673
Received: 16-11-2025 Revised: 05-12-2025 Accepted: 25-01-2026

Table 4: Accuracy by Dataset and Domain

Dataset	Domain	Rule-Based	Best Individual ML	Ensemble
Healthcare-EHR	Healthcare	80.3%	89.7% (GBM)	94.2%
Financial-Trans	Finance	85.1%	91.8% (GBM)	95.8%
E-commerce-Orders	E-commerce	81.7%	90.2% (RF)	93.9%
Manufacturing-IoT	Manufacturing	83.9%	90.4% (GBM)	94.3%
Customer-CRM	Marketing	82.4%	88.9% (RF)	95.1%

The Financial-Trans dataset achieves the highest accuracy (95.8%), likely due to its well-structured transactional data with clear validation rules. Healthcare-EHR shows slightly lower but still excellent performance (94.2%), reflecting the complexity and variability of electronic health records. Importantly, the ensemble approach demonstrates consistent superiority across all domains, validating its generalizability.

5.4 Computational Efficiency Analysis

While ensemble methods achieve superior accuracy, computational cost is an important practical consideration. Table 5 compares training and inference times across different approaches.

Table 5: Computational Performance (Average across all datasets)

Method	Training Time (seconds)	Inference Time (ms)	Memory Usage (MB)
Rule-Based	5.2	0.8	12
Random Forest	127.3	3.2	385

Power System Technology
ISSN:1000-3673
Received: 16-11-2025 Revised: 05-12-2025 Accepted: 25-01-2026

GBM	156.8	4.1	420
Neural Network	342.5	2.7	512
Ensemble (Voting)	298.4	9.5	1,217
Ensemble (Stacking)	315.7	11.3	1,289

The ensemble approach requires approximately 2.5 times the training time of the fastest individual ML method (Random Forest) and 3 times the memory. However, inference time remains under 12ms per record, which is acceptable for most real-time applications. For batch processing scenarios, these computational costs are negligible compared to the accuracy improvements gained. Organizations can trade computational resources for significantly improved data quality assessment.

6. DISCUSSION

6.1 Key Findings

Our research demonstrates that ensemble machine learning techniques provide a robust and effective solution for automated data quality assessment. The key findings can be summarized as follows: (1) Ensemble methods significantly outperform both traditional rule-based approaches and individual ML algorithms across all evaluation metrics and datasets. (2) Stacking ensembles achieve the best performance by learning optimal combinations of base model predictions. (3) The framework shows excellent generalizability across diverse domains and data quality dimensions. (4) While computationally more expensive than individual methods, the performance gains justify the additional resource requirements. (5) Feature engineering, particularly the inclusion of domain-specific features, plays a crucial role in model performance.

6.2 Practical Implications

For practitioners implementing automated data quality systems, we recommend:

- Start with comprehensive data profiling to understand quality issue patterns
- Invest in quality labeled training data through expert annotation
- Implement domain-specific feature engineering to capture business context
- Use stacking ensembles for maximum accuracy when computational resources permit
- Deploy models incrementally, starting with critical data quality dimensions

Power System Technology
ISSN:1000-3673
Received: 16-11-2025 Revised: 05-12-2025 Accepted: 25-01-2026

- Establish continuous monitoring and model retraining pipelines to maintain accuracy as data patterns evolve

6.3 Limitations and Future Work

Despite promising results, this research has several limitations that present opportunities for future work. First, our evaluation focused on structured tabular data; extending the framework to semi-structured and unstructured data (text, images, video) remains an important challenge. Second, the requirement for labeled training data may be prohibitive in some contexts; exploring semi-supervised and active learning approaches could reduce annotation burden. Third, we did not extensively investigate the interpretability of ensemble predictions; developing explainable AI techniques for data quality assessment would enhance trust and adoption. Finally, real-time processing of streaming data requires further optimization of the ensemble architecture. Future research should address these limitations and explore integration with automated data remediation systems.

7. CONCLUSION

This research presents a comprehensive investigation of automated data quality assessment using ensemble machine learning techniques. Through extensive empirical evaluation on five real-world datasets from diverse domains, we demonstrate that ensemble methods, particularly stacking approaches, significantly outperform traditional rule-based systems and individual ML algorithms. Our proposed framework achieves 94.7% accuracy, representing substantial improvements over existing approaches.

The results validate several important conclusions. First, combining multiple ML algorithms through ensemble techniques provides robustness against varied data quality issues and domain-specific challenges. Second, the investment in comprehensive feature engineering, including domain-specific attributes, is critical for achieving high performance. Third, while ensemble methods incur higher computational costs than individual models, these costs are justified by the significant accuracy improvements and remain practical for both batch and near-real-time applications.

As organizations increasingly rely on data-driven decision-making, automated data quality assessment becomes essential infrastructure. Our research provides both theoretical insights and practical guidance for implementing ML-based quality systems. The framework's generalizability across domains and quality dimensions suggests it can serve as a foundation for enterprise-wide data quality management.

Future work should focus on extending the framework to unstructured data types, reducing labeled data requirements through semi-supervised learning, enhancing model interpretability, and optimizing for real-time streaming data. Additionally, integration with automated data

Power System Technology
ISSN:1000-3673
Received: 16-11-2025 Revised: 05-12-2025 Accepted: 25-01-2026

remediation systems would complete the data quality management pipeline, enabling end-to-end automation from issue detection to resolution.

In conclusion, ensemble machine learning techniques represent a mature and effective approach for automated data quality assessment. Organizations seeking to improve their data quality management capabilities should seriously consider adopting these methods as part of their data governance strategy. The demonstrated benefits in accuracy, robustness, and generalizability make ensemble ML an attractive alternative to traditional manual and rule-based approaches.

REFERENCES

[1] Batini, C., & Scannapieco, M. (2016). Data and Information Quality: Dimensions, Principles and Techniques. Springer International Publishing.

[2] Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems, 12(4), 5-33.

[3] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

[4] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.

[5] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189-1232.

[6] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

[7] Dasu, T., & Johnson, T. (2003). Exploratory Data Mining and Data Cleaning. John Wiley & Sons.

[8] Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. IEEE Data Engineering Bulletin, 23(4), 3-13.

[9] Redman, T. C. (2001). Data Quality: The Field Guide. Digital Press.

[10] Dietterich, T. G. (2000). Ensemble methods in machine learning. Multiple Classifier Systems, 1-15.

[11] Wolpert, D. H. (1992). Stacked generalization. Neural Networks, 5(2), 241-259.

[12] Zhang, C., & Ma, Y. (2012). Ensemble Machine Learning: Methods and Applications. Springer Science & Business Media.

[13] Krishnan, S., Franklin, M. J., Goldberg, K., Wang, J., & Wu, E. (2016). ActiveClean: An interactive data cleaning framework for modern machine learning. Proceedings of the 2016 ACM SIGMOD International Conference on Management of Data, 2117-2120.

Power System Technology
ISSN:1000-3673
Received: 16-11-2025 Revised: 05-12-2025 Accepted: 25-01-2026

[14] Chu, X., Ilyas, I. F., Krishnan, S., & Wang, J. (2016). Data cleaning: Overview and emerging challenges. Proceedings of the 2016 ACM SIGMOD International Conference on Management of Data, 2201-2206.

[15] Rekatsinas, T., Chu, X., Ilyas, I. F., & Ré, C. (2017). HoloClean: Holistic data repairs with probabilistic inference. Proceedings of the VLDB Endowment, 10(11), 1190-1201.