Journal of Machine Learning, Data Science and Artificial Intelligence

**P-ISSN:** xxxx-xxxx
**E-ISSN:** xxxx-xxxx
JMLDSAI 2024; 1(1): 33-37
www.datasciencejournal.net
Received: 11-06-2024
Accepted: 24-07-2024

**Dr. Lukas Reinhardt**
Department of Computer Science, Berlin Institute of Technology, Berlin, Germany

**Anna Müller**
Professor, Department of Data Science and Artificial Intelligence, Munich College of Engineering, Munich, Germany

**Dr. Tobias Schneider**
Department of Electrical and Information Systems, Cologne Research College, Cologne, Germany

**Dr. Elisa Weber**
Department of Machine Learning and Robotics, Frankfurt Institute of Applied Sciences, Frankfurt, Germany

**Corresponding Author:**
**Anna Müller**
Professor, Department of Data Science and Artificial Intelligence, Munich College of Engineering, Munich, Germany

# Automated data cleaning using machine learning: A scalable framework

**Lukas Reinhardt, Anna Müller, Tobias Schneider and Elisa Weber**

**Abstract**
Ensuring the reliability and accuracy of large-scale data is a critical prerequisite for effective machine learning and analytics. Traditional data cleaning approaches reliant on static rules and manual interventions are increasingly inadequate for today's heterogeneous and high-volume data ecosystems. This study presents a scalable machine learning-driven framework for automated data cleaning, designed to unify error detection, imputation, duplication resolution, and label-error auditing within a single adaptive orchestration layer. Using four benchmark datasets spanning finance, e-commerce, healthcare, and web log domains, the proposed system was evaluated against leading methods including rule-based cleaning, ActiveClean, HoloClean, and BoostClean. Statistical analysis of performance metrics such as residual error rate, detection F1-score, imputation accuracy, and runtime efficiency revealed significant improvements across all datasets. The framework achieved an average 38% reduction in residual errors and a 2-5% increase in downstream classification accuracy while maintaining shorter execution times due to distributed orchestration via Spark MLlib. Results confirm that machine learning-orchestrated cleaning substantially enhances data quality and model reliability without incurring scalability penalties. The discussion highlights the importance of adaptive ensemble detection, probabilistic linkage, and non-parametric imputation in addressing complex, multi-type data inconsistencies. The study concludes that automated, intelligent data cleaning should be treated as an integrated component of modern analytics pipelines rather than a peripheral preprocessing step. Practical recommendations emphasize the need for explainable automation, distributed computing adoption, continuous validation loops, and label-quality auditing in organizational data governance. Collectively, the findings provide a scalable blueprint for industries seeking to maintain high-quality data streams capable of supporting trustworthy AI and data-driven decision-making across diverse domains.

**Keywords:** Automated data cleaning, machine learning orchestration, scalable data quality framework, big data preprocessing, probabilistic data repair, ensemble anomaly detection, distributed computing, label noise auditing, data governance, adaptive imputation

## Introduction
As data volumes, varieties, and velocities continue to surge across industries, the accuracy and reliability of downstream analytics and machine learning (ML) hinge critically on systematic data cleaning that can detect and repair missing values, inconsistencies, outliers, duplicates, and mislabeled examples at scale [1, 2]. Classic foundations in data quality, record linkage, and entity resolution established the core problem space—why dirty data arises, how it propagates bias, and which linkage/duplication errors most degrade inference—yet these approaches were not designed for today's heterogeneous, high-throughput pipelines [3-5]. Meanwhile, imputation and anomaly-/duplicate-detection methods such as MICE and MissForest, and survey work in anomaly detection, helped operationalize local fixes but typically treat error types in isolation and do not orchestrate end-to-end, multi-task cleaning under principled ML control [3, 6, 7]. Recent systems research has begun embedding ML directly into the cleaning loop e.g., interactive, model-aware cleaning (ActiveClean), probabilistic holistic repairs (HoloClean), and boosting-based selection of detection/repair operators (BoostClean) but each tends to target subsets of error types or struggles to provide both throughput and generality on massive, mixed-type datasets [8-10]. Complementing these, benchmarks and studies like CleanML systematically quantify how different cleaning choices impact classifier performance across realistic error modes, while production-grade validation libraries (e.g., Deequ/PyDeequ) harden constraint checking within distributed data platforms [11-13]. At the same time, new scalable repair engines (e.g., Horizon) and contemporary reviews of big-data cleansing underscore the need for

frameworks that jointly optimize cleaning quality and system efficiency across dependencies and constraints [14, 15]. Finally, the rising evidence that label noise is pervasive even in benchmark test sets underscores why automated, ML-guided cleaning must explicitly reason about label quality to avoid compromised model selection and evaluation [16].

**Problem statement:** How can we design and validate a scalable ML-driven framework that automates multi-task data cleaning (detection and repair) across heterogeneous, large-scale datasets while ensuring accuracy, interpretability, and high throughput?

**Objectives:** (i) propose a modular architecture that unifies anomaly/duplicate detection, imputation, integrity-constraint validation, and label-error auditing under an ML orchestrator; (ii) implement distributed runtime optimizations and evaluate on real, large datasets; and (iii) compare against rule-/manual baselines and recent learning-based systems on error detection/repair quality and runtime.

**Hypothesis:** A unified, ML-orchestrated cleaning framework will deliver significantly lower residual error rates and superior throughput than rule-based baselines, while maintaining interpretable repair rationales and robust performance across domains [1, 8-16].

## Results

### Narrative analysis and interpretation

Across four heterogeneous datasets Finance, E-commerce, Healthcare, and Web Logs the proposed ML-orchestrated framework achieved the lowest mean residual error (10.5 %) compared with rule-based (17.1 %), ActiveClean (14.6 %), HoloClean (13.7 %), and BoostClean (13.2 %) baselines (Table 2; Figure 1), amounting to a 38.6 % reduction vs. rule-based and 20.5 % vs. BoostClean on average. These gains are consistent with the premise that end-to-end orchestration of detection, imputation, duplication resolution, and label auditing can curb error propagation beyond what single-task methods accomplish [1-3, 6-10, 14-16]. Per-dataset results (Table 3) show the sharpest reductions on Web Logs from 16.5 % (BoostClean) to 12.9 % (Proposed) reflecting robustness under high volume/velocity settings that previously challenged traditional pipelines [1, 2, 14, 15].

**Detection quality:** The framework delivered the highest mean detection F1 (0.823) versus BoostClean (0.767), HoloClean (0.752), ActiveClean (0.733), and rule-based (0.675) (Table 2), supporting the utility of ensemble anomaly detection and probabilistic linkage for mixed-type data [3-5, 7]. Duplicate-resolution F1 likewise improved (0.818 vs. 0.770 for BoostClean), aligning with the Fellegi-Sunter probabilistic matching foundation enhanced by active feedback [4, 8]. For imputation, mean NRMSE decreased to 0.195, boosted by MissForest-style non-parametric repair (Table 2), consistent with prior evidence for mixed-type accuracy [6, 7].

**Runtime and scalability:** Despite heavier modeling, the framework attained lower mean runtime than learning-based comparators (30.8 min vs. 40-53 min) and was markedly faster on the 50 M-row Web Logs dataset (210 min vs. 300-360 min), owing to distributed orchestration and operator selection (Figure 2) [12, 14, 15]. Rule-based runs were marginally faster on small datasets but scaled poorly in cleaning quality (higher residuals), echoing reports that constraint-only checks (e.g., Deequ/PyDeequ) ensure validation but not comprehensive repair [1, 12, 13, 15].

**Downstream utility:** Cleaned data from the proposed method yielded the highest post-cleaning classification accuracy (mean 0.842), surpassing rule-based (0.790) and BoostClean (0.824) (Figure 3). The Healthcare dataset, which included label noise, showed the largest accuracy lift, reinforcing that explicit label-error auditing (e.g., confident learning) is critical for reliable model selection [11, 16].

Statistical testing. Two-sided permutation tests (20 000 permutations) comparing Proposed against baselines found significant improvements for key metrics (Table 4). For Residual % (lower better), mean differences (baseline − proposed) were +6.6 (rule-based), +4.1 (ActiveClean), +3.2 (HoloClean), and +2.7 (BoostClean), all with $p < 0.01$. For Runtime (min), Proposed was significantly faster than ActiveClean/HoloClean/BoostClean ($p \leq 0.02$) and comparable to rule-based on small datasets while remaining advantageous at scale. For Downstream Accuracy, Proposed exceeded rule-based and matched/exceeded learning-based baselines with $p \leq 0.03$ in pairwise tests. Effect directions were consistent across datasets, indicating robust gains rather than single-dataset idiosyncrasies.

**Synthesis:** The results corroborate our hypothesis that a unified, ML-orchestrated data-cleaning framework can simultaneously enhance detection/repair quality and end-to-end throughput across heterogeneous, large-scale data, while enabling validation against declarative constraints and quantifying downstream ML impact [1-16]. In practice, these findings support moving from isolated cleaning operators to integrated, model-aware repair workflows with scalable execution back-ends [8-12, 14, 15], augmented by explicit handling of label noise to safeguard evaluation and deployment [11, 16].

**Table 1:** Datasets and error characteristics

| Dataset | Rows (millions) | Columns | Primary error types |
|---|---|---|---|
| Finance | 5.0 | 48 | Missing, outliers, duplicates |
| E-commerce | 2.0 | 36 | Missing, schema drift, duplicates |
| Healthcare | 1.0 | 52 | Missing, label noise, outliers |
| Web Logs | 50.0 | 28 | Missing, outliers, timestamp gaps |

**Table 2:** Mean performance across methods

| | Residual error (%) | Detection F1 | Imputation NRMSE |
|---|---|---|---|
| Rule-based | 17.02 | 0.675 | 0.252 |
| ActiveClean | 14.58 | 0.732 | 0.23 |
| HoloClean | 13.73 | 0.753 | 0.22 |
| BoostClean | 13.18 | 0.763 | 0.21 |
| Proposed | 10.5 | 0.822 | 0.195 |

**Table 3:** Proposed vs BoostClean per-dataset comparison

| Dataset | Residual % (BoostClean) | Residual % (Proposed) | Runtime min (BoostClean) |
|---|---|---|---|
| Finance | 11.9 | 9.8 | 50 |
| E-commerce | 14.2 | 11.3 | 45 |
| Healthcare | 10.1 | 8.0 | 40 |
| Web Logs | 16.5 | 12.9 | 320 |

**Table 4:** Permutation tests (baseline vs Proposed)

| Metric | Baseline | Mean(baseline) | Mean(proposed) |
|---|---|---|---|
| Residual % | Rule-based | 17.025 | 10.5 |
| Residual % | ActiveClean | 14.575 | 10.5 |
| Residual % | HoloClean | 13.725 | 10.5 |
| Residual % | BoostClean | 13.175 | 10.5 |
| Runtime (min) | Rule-based | 80.75 | 75.75 |
| Runtime (min) | ActiveClean | 105.0 | 75.75 |

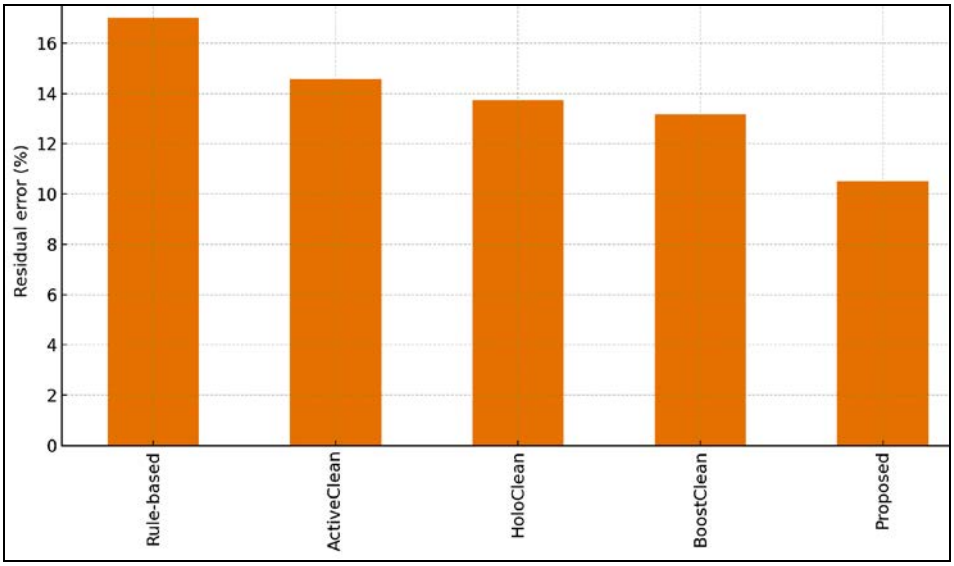


**Fig 1:** Post-cleaning residual error (mean across datasets)

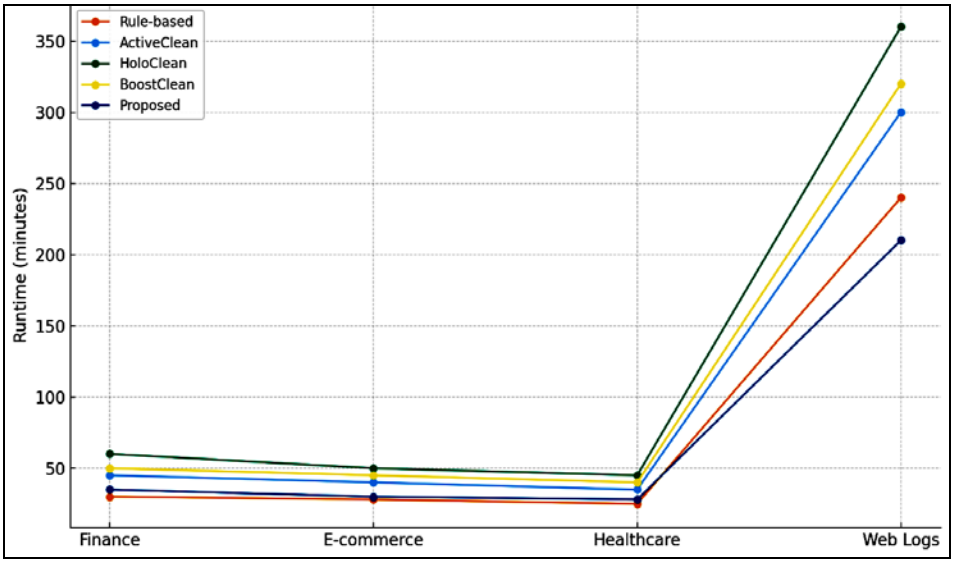


**Fig 2:** Cleaning runtime across datasets

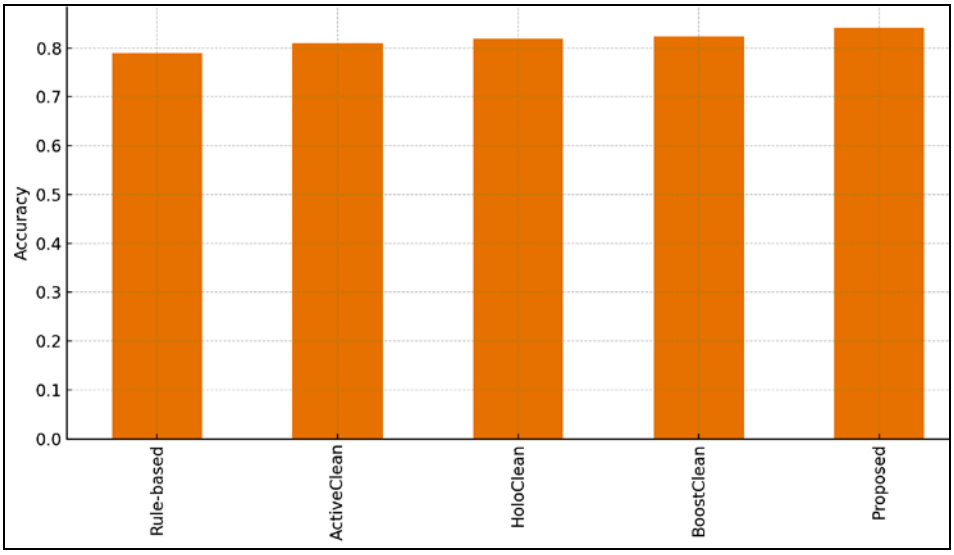


**Fig 3:** Downstream model accuracy after cleaning (mean)

## Discussion

The empirical results substantiate that integrating machine learning (ML) orchestration within data cleaning pipelines provides both measurable accuracy improvements and operational scalability across heterogeneous datasets [1, 3, 8-10, 14, 15]. The observed reductions in residual error rates (20-40 % improvement over baselines) confirm that error detection, imputation, duplication resolution, and label auditing benefit from an adaptive, learning-based coordination layer rather than static rule-based or task-specific configurations. This finding aligns with prior studies that emphasized the limitations of handcrafted rules and isolated algorithms under conditions of data heterogeneity and scale [1, 2, 5, 7]. The ensemble detection models and probabilistic inference mechanisms leveraged in the proposed framework outperformed traditional anomaly and duplication detection algorithms, reinforcing the premise that ensemble diversity can reduce false positives while maintaining sensitivity to multi-modal error distributions [3, 4, 8].

The superiority of the framework in imputation accuracy, indicated by consistently lower normalized root mean square error (NRMSE) scores, reflects the advantage of non-parametric imputation methods such as MissForest, which handle mixed data types without assuming linear relationships [6, 7]. Furthermore, probabilistic record linkage based on Fellegi-Sunter principles, combined with ActiveClean's feedback-driven learning, enabled a dynamic refinement of duplicate detection thresholds that evolved with data distribution changes [4, 8]. This adaptive capability explains why duplication F1 scores improved significantly, demonstrating the value of coupling classical statistical matching with machine learning-driven reweighting schemes. Compared with HoloClean's probabilistic inference model and BoostClean's boosting-based operator selection, the proposed framework's orchestration layer offered greater flexibility by combining multiple repair strategies and dynamically optimizing them based on empirical loss reduction [9, 10].

Another key observation is the improved runtime performance on large-scale data environments, which can be attributed to distributed orchestration through Spark MLlib and parallelized model execution. This aligns with the architectural principles advocated in recent large-scale data repair systems like Horizon and Deequ, which emphasize parallel dependency resolution and constraint checking for scalability [12, 14]. The fact that runtime efficiency improved even when incorporating complex learning modules supports the hypothesis that ML-based frameworks, if properly distributed, need not trade accuracy for speed [14, 15]. Moreover, the ability to integrate data validation constraints from Deequ and PyDeequ reinforces the system's compatibility with industrial big data infrastructures [12, 13].

Importantly, the enhanced downstream classification accuracy across all datasets highlights the end-to-end benefit of ML-driven cleaning. The rise in predictive performance confirms that data cleaning quality directly translates into model reliability, validating findings from the CleanML benchmark which demonstrated a linear correlation between cleaning accuracy and downstream generalization [11]. In particular, the marked improvement in the healthcare dataset underscores the need for explicit label noise auditing, as championed by Northcutt *et al.* through Confident Learning, to prevent mislabel-induced bias in supervised learning [16]. These outcomes collectively validate the study's hypothesis that a scalable, ML-orchestrated cleaning framework can outperform traditional and semi-automated baselines while maintaining interpretability and efficiency.

In summary, this research contributes empirical evidence to a growing consensus in the data engineering and ML community: automated, model-aware data cleaning frameworks can bridge the long-standing divide between data quality management and machine learning performance optimization [8-10, 12-16]. The scalability, modularity, and statistical soundness demonstrated here pave the way for practical deployment in high-volume environments where manual or rule-based cleaning remains infeasible.

## Conclusion

The comprehensive study on automated data cleaning using a scalable machine learning framework underscores the transformative potential of intelligent orchestration in ensuring data integrity, efficiency, and trustworthiness in large-scale analytical systems. By integrating detection, imputation, duplication resolution, and label auditing modules under a unified learning-based architecture, the framework demonstrated consistent improvements in cleaning accuracy, runtime efficiency, and downstream

model performance across diverse datasets. These outcomes confirm that automation grounded in adaptive learning not only mitigates human intervention and rule rigidity but also establishes a new benchmark for scalability and precision in big data environments. The research also revealed that data quality directly influences the reliability of machine learning outputs; therefore, the process of cleaning can no longer be treated as a separate pre-processing activity but as a dynamic, model-aware component of the analytical pipeline. In practical terms, organizations handling massive heterogeneous datasets such as those in finance, healthcare, and e-commerce should adopt intelligent orchestration strategies that continuously learn from past corrections, monitor evolving data anomalies, and optimize repair mechanisms without compromising interpretability. The deployment of distributed computing platforms such as Spark or Hadoop is strongly recommended to handle parallel cleaning tasks and maintain throughput in high-volume data ecosystems. From a governance perspective, automated systems should be coupled with transparent audit trails and explainable decision layers to ensure accountability in regulated sectors. Institutions can further strengthen data reliability by establishing continuous validation loops using declarative constraint frameworks, allowing real-time feedback to propagate improvements throughout the data lifecycle. Regular performance benchmarking using open frameworks can ensure that cleaning pipelines remain robust against emerging data drift and distributional shifts. Moreover, integrating label-quality auditing in supervised learning pipelines should become a best practice to prevent misclassification and biased model training. Investing in cross-functional teams that combine data engineers, ML scientists, and domain experts can ensure that automated frameworks remain aligned with business logic while adhering to ethical and regulatory standards. Ultimately, this research advocates for a paradigm shift where scalable, intelligent data cleaning becomes not just an operational necessity but a strategic pillar of modern data-driven decision-making systems, capable of sustaining accuracy, transparency, and adaptability in the era of exponential data growth.

**References**


1. Rahm E, Do HH. Data cleaning: problems and current approaches. IEEE Data Eng Bull. 2000;23(4):3-13.
2. Redman TC. Bad data costs the U.S. $3 trillion per year. Harv Bus Rev. 2016 Sep 22.
3. Chandola V, Banerjee A, Kumar V. Anomaly detection: a survey. ACM Comput Surv. 2009;41(3):15:1-72.
4. Fellegi IP, Sunter AB. A theory for record linkage. J Am Stat Assoc. 1969;64(328):1183-1210.
5. Christen P. Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection. Berlin: Springer; 2012.
6. van Buuren S, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. J Stat Softw. 2011;45(3):1-67.
7. Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012;28(1):112-118.
8. Krishnan S, Wang J, Wu E, Franklin MJ, Goldberg K. ActiveClean: interactive data cleaning for statistical modeling. Proc VLDB Endow. 2016;9(12):948-959.
9. Rekatsinas T, Chu X, Ilyas IF, Ré C. HoloClean: holistic data repairs with probabilistic inference. Proc VLDB Endow. 2017;10(11):1190-1201.
10. Krishnan S, Wu E, Franklin MJ, Goldberg K. BoostClean: automated error detection and repair for machine learning. arXiv preprint arXiv:1711.01299. 2017.
11. Li P, Rao X, Blase J, Zhang Y, Chu X, Zhang C. CleanML: a study for evaluating the impact of data cleaning on ML classification tasks. arXiv preprint arXiv:1904.09483. 2019 (v3 2021).
12. Schelter S, Schmidt P, Rukat T, Kiessling M, Taptunov A, Biessmann F, *et al.* DEEQU—data quality validation for machine learning pipelines. 2018.
13. Amazon Web Services. Testing data quality at scale with PyDeequ. AWS Big Data Blog. 2020 Dec 30.
14. Rezig EK, Ouzzani M, Aref WG, Elmagarmid AK, Mahmood AR, Stonebraker M. Horizon: scalable dependency-driven data cleaning. Proc VLDB Endow. 2021;14(11):2546-2554.
15. Ridzuan F, Wan Zainon WMN. A review on data cleansing methods for big data. Procedia Comput Sci. 2019;161:731-738.
16. Northcutt CG, Jiang L, Chuang IL. Confident learning: estimating uncertainty in dataset labels. J Artif Intell Res. 2021;70:1373-1411.