International Journal of Theoretical & Applied Computational Intelligence
https://ijtaci.com

Research Article

Experimental Evaluation of Automated and Manual Data Cleaning Systems: A Case Study Using Organizational Data

Nagwa Elmobark[1], Amenah Y. Abdzaid[2], Farqad Alaa[3] and Aymen Saad[3,4*]

[1]Researcher, Department of Computer Science, University of Mansoura, Egypt.
[2]Fatima Al Zahra School for Distinguish Students, AL-Diwaniyah Education Directorate, Iraq
[3]Department of Information Technology, Management Technical College, Al-Furat Al-Awsat Technical University, Kufa, Iraq.
[4]School of Electrical Engineering, Universiti Teknologi Malaysia (UTM) Skudai, Johor, Malaysia.

[*] Corresponding author's e-mail: aymen.abdalameer@atu.edu.iq

https://orcid.org/0000-0002-3582-6799
Received: 06/12/2025; Accepted: 07/02/2025; Published: 10/03/2025
https://doi.org/10.65278/IJTACI.2026.42

Abstract: This paper presents a comprehensive comparison between computerized and manual data cleaning methods, using different types of outliers' statistics, including missing values, wrong figures, and inconsistent formats. This view provides a more holistic comparison between the automated and guidance-based information-cleaning methodologies in terms of their efficiency along the three most important dimensions of information quality. Leveraging a common Kaggle dataset of 50,000 employees' details, we create a systematic evaluation framework with diagnostic measures, fine-grained diagnostics and resource utilization profiles. The experiments are conducted quantitatively in order to evaluate the cleaning on a given distribution of types of anomalies: 30% missing values (15,000 records), 25% incorrect values (12,500 records) and 45% inconsistent formats (22,500 records). Automated cleaning significantly helped normalize mixed formats (92% success) and regarding invalid statistics (88%). However, manual cleaning was better than automatic methods on complex cases and context-dependent learning, with a 95% accuracy in area-knowledge-required cases. The paper shows clear benefits for both approaches: (i) automated cleansing excels at both speed and cost on large sets, alongside hand-cleaning is useful in difficult examples that require domain knowledge. These findings add value to the literature, by presenting empirical evidence of effectiveness for both approaches and providing firms with a structured, knowledge-based filter for choosing suitable cleaning solutions in line with

their actual state of knowledge, needs and organizational structure.

Keywords: Data cleaning, automation, missing values, invalid data, inconsistent formats, data quality, efficiency metrics, comparative analysis.

1. Introduction

1.1 Background

Today's business world is a fact-driven one, and companies face ever-increasing pressures to get their facts right. As the amount and speed of knowledge have been developing at an accelerating pace, traditional ways for cleaning noisy examples are becoming puny [1]. Data quality issues (DQI) including missing, incorrect and non-standard formatted data are prevalent and critically impacting business performance and decision-making [2]. Trillions of records have been generated from computerization of factory processes, some of which include employee data but are not well-preserved. They are dealing with non-aligned data strewn about in multiple systems preventing them from being able to rely on fine-tuned HR analytics and decision-support methods [3]. Recently, automatic record-cleaning methods have been proposed as a means of tackling such issues and afford a controllably scalable and consistent method for data-quality control [4]. In this work, we used a large fact dataset in Kaggle with 50.000 employees created by our partner, carefully produced to mimic real-world quality issues in the data. The popular dataset contains the common problem of data cleaning: 30% missing in mandatory fields like contact and performance indicators; 25% invalid entries with out-of-range values or mixed types; and 45% nonstandard forms for date, address, and number. This distribution is a helpful starting point to comparing automatic and manual methods across many dimensions of factual quality.

1.2 Conceptual Distinction Between Manual and Automated Data Cleaning

To establish a useful analytical frame for this comparison, the conceptual roots of manual and automatic data cleaning should be put in some detail as well, and especially those conditions that might become source of strengths/limits for both. Manual data cleaning is a labour-intensive process where data quality experts based on their domain and contextual knowledge along with professional judgement identify and fix the dirty data [5]. The approach follows a case-by-case reasoning where the context of each data quality issue is taken into account, allowing inter- pretations by considering business constraints, local skills and not formalized semantic relations [6]. The manual approach is a cognitive process, as it depends on human pattern recognition which goes beyond a priori rules to include implicit knowledge and experience learning [7]. Contrarily, automatic data cleaning utilizes algorithms to develop pattern recognition, statistical evaluation and rule-based inference mechanisms to locate and fix quality issues in the data at large scale [8]. The foundation of this paradigm is the organized use of predefined validation rules, contraint-checking procedures and standardization procedures that may be qualitatively (re)used across large-scale datasets [9]. Automated systems work on formal and explicit knowledge representations; thus, data quality management is a computational issue to solve by optimizing algorithms [10]. The conceptual differences between these inferences are evident along several axes as shown in Table 1.

Table 1: Conceptual Framework - Manual vs. Automated Data Cleaning

Conceptual Dimension	Manual Cleaning	Automated Cleaning
Epistemological Basis	Tacit knowledge and domain expertise	Explicit rules and formalized algorithms
Decision-Making Mechanism	Context-dependent human judgment	Rule-based deterministic processing
Scalability Paradigm	Linear human resource constraints	Computational scalability with minimal marginal cost
Consistency Model	Variable and operator-dependent (expertise-mediated)	Deterministic and uniformly applied
Contextual Understanding	Superior semantic interpretation	Limited to predefined rule coverage
Error Handling Approach	Adaptive and learning-based refinement	Systematic and pattern-based detection
Resource Requirements	Intensive human capital with continuous engagement	Initial computational investment with minimal ongoing cost
Theoretical Foundation	Cognitive science and expert systems theory	Computer science, machine learning, and formal methods

These conceptual differences inform our experimental design and provide a theoretical perspective for explaining the empirical results presented in this study. This knowledge of such basic differences can help organizations make informed decisions about which approach to adopt, given their data quality, resource constraints, and operational needs.

1.3 Dataset Size Justification

The choice of the 50,000 data length statistics will bring to:

- Provide sufficient scale for a significant automatic processing assessment
- Enable thorough manual cleaning comparison

1.4 Data Quality Dimensions Framework

Data quality has various interrelated dimensions. Data quality can be assessed at different levels, as suggested by the widely recognized approaches developed by Wang & Strong [11] on one hand and by DAMA-DMBOK standards [12] on the other against parameters such as completeness, accuracy, consistency (or free-of-error), timeliness (or currency), validity, uniqueness. We focus on three of the more fundamental aspects of data quality subjects to comparison between manual and automated cleaning approaches:

1. **Completeness:** The extent that all required data items are available in the dataset. In our approach, we make it concrete by looking to missing with which represent 30 % of the error distribution. Data completeness is a critical requirement, as incomplete data would otherwise have a direct impact on analysis and decision-making [13].

2. **Precision:** The degree of alteration between the data about an actual world item or event and the actual world item or event that it is supposed to represent. We evaluate this dimension by using invalid data (25% of errors), range violations, wrong data types and logical inconsistencies due to domain constraints. Profile Quality: Includes ensuring that profiles are well-organized, complete and discussed outliers.

3. **Consistency:** how the elements of the data set are consistently represented, so that values which represent the same items are also expressed in a standardized format. This is one of the dimensions that defines 45 % of our error distribution and includes problems regarding format standardization (date format, phone number formatting or text being capitalized).

These three dimensions along with completeness, accuracy and consistency, are the key data quality concerns in organizational settings [14]. They are inherent modes of data quality, which can be dealt with directly by cleaning interventions, while context dimensions (e g., timeliness, relevancy) are not strictly based on proprieties of data themselves.

1.5 Scope Justification:

The choice of the three dimensions to engage with comparative analysis is theoretically predetermined with a variety of concerns:

1. Intervention Amenable: Cleaning actions are completely modifiable in terms of their completeness, correctness, and coherence, and therefore may be compared under experimental conditions. On the contrary, other dimensions (timeliness and the up-to-datedness of data with respect to decision-time) are controlled by design not by cleaning interventions but rather by data collection and updating activities [15].
2. Occurrence in Organizations: Within organizational context these three dimensions encapsulate some of the most prevalent data quality problems faced by enterprise, cumulatively responsible for majority of data quality challenges that need to be repaired [16].
3. Ablation Study A. Comparative Evaluability: These representations are compelling for comparing cleaning performance in an objective manner, by ranking them on a ground truth with comparable conditions that holds the clean target text fixed across all datasets, allowing a robust comparison of manual and automated approaches. Other aspects of completing the picture of data quality such as timeliness, believability and interpretability are unquestionably important but are not covered here. Comparative cleaning study, because they cannot be directly measured with deco cleaned or other methodology.

1.6 Problem Statement

The Core Research Problem: The truck owner-operator/manager has to make a basic theoretical and practical choice in data quality management - to clean manual (with human effort) the data or automatic data cleaning using guiding algorithms; this decision is indeed a polytope-like (or hypercube) optimization problem, which is not possible empirically grounded on how to take the decision. The problem is not only a practical one; it also has some theoretical impact as well: the acceptance of conditionally better quality of human cognition than an algorithmic processing in assurance of data quality.

Dimensions of the Research Problem:

Efficiency-Quality Trade-off Uncertainty:

Traditional manual data cleaning systems tend to obtain better contextual accuracy; they still have a serious scalability issue. Data cleaning accounts for 50 to 80 % of the project time of data scientists [17]. This has very limited potential opportunity at an organizational level where analytical operations remain hamstrung. However, the specific quantitative-time-to-quality output is in theory untested as yet and the efficiency–quality nature of the automatic options is open to question.

Error-Type Specificity Gap:

There are different kinds of data quality issues (like missing, invalid, or inconsistent data) that can be approached in a separate manner in manual and automated handling. These differences between error types have not been quantified in the literature, making it difficult to determine which approach to apply specifically to these data-quality issues [18, 19]. This conceptual vacancy in turn, prevents the establishment of an error taxonomy-based decision model.

Scalability Threshold Ambiguity:

The key question is: How big does the dataset need to be (theoretically) for automatic cleaning? has empirical substantiation. It is intuitive that the automation gains are greater with more data, however these inflexion points have not been quantified [20,21]. This kind of confusion seriously impedes the strategic investment planning for data quality infrastructure.

Consistency-Adaptability Paradox:

Automated systems benefit from processing consistency but lose contextual flexibility; manual approaches gain in terms of flexibility, but risk increased variability through the introduction of human factors [22]. The models have taken little consideration of these conflicting pressures within organisations.

Theoretical Significance:

This research issue is of theoretical importance as it probes a fundamental question in information quality science: can the expert know-how contribute to data quality assurance be replaced by algorithmic processing, and within what boundary contexts. Answering this question will require us to go beyond simply assessing the capabilities of technology and start exploring the epistemic basis for judgment about data quality - when tacit understanding and context-specific reasoning are necessary, and when formalized rules will do.

Practical Implications:

However, this failure to compare them in a transparent and structured manner has real consequences for organizations: sub-optimal choice of cleaning actions, wasted efforts on data that would not have been used anyway, lower quality data. The research method used in this article allows to approach the research gap by performing for the first time a systematic, controlled comparison between manual and automatic data cleaning techniques in standard error distributions with common evaluation metrics: This establishes an empirical basis to overcome theoretical shortcomings in data quality management issues.

1.7 Research Objectives

This paper gives a complete overview of automatic and manually driven statistical cleansing methods, exposing how to tackle missing values, dirty records and imprecise codes for employee's data in the best way. Highest priorities, one of the highest things that we're going to be is looking for. Automatic vs manual methods for cleaning particular types of errors [23]. The performance is measured in terms of execution time, accuracy and the resources consumed [24]. The purpose of this research is to develop guidelines for cleaning records. We received a sample dataset.

1.8 Theoretical Justification for Error Type Selection

The notion of data quality is usually inherently multi-dimensional, and cyclic standards like the ISO 8000 or the DAMA-DMBOK framework have targeted fundamental dimensions, including completeness, accuracy, integrity, validity and consistency over time. In the present research, such values were given priority hypothetically and consistently located in the elementary dimensions of data quality for theoretical reasons rather than as an experimental convenience. In ISO 8000, it is specifically these three dimensions that are identified as the critical prerequisites of data usability, and integrated use followed by analytics in subsequent processes. A particular focus on these types of error, then, yields a theoretically motivated standard-based framework for comparative evaluation, and reflects the most prevalent and operationally relevant data quality issues encountered in organisational datasets.

2. Literature Review

2.1 Historical Development

Statistical cleansing has evolved over the decades. Traditionally, figure cleaning in the editing process has relied heavily on manual inspection and correction (almost always by data-entry staff) across a large number of periodicals [25]. The semi-automated methods emerged in the 1990s and involved building (simple) rule-based models for detectable errors [26]. The first tools for the new situation with much larger datasets appeared quite early, in the early 2000s, in industry [27]. The growth in the volume of data produced by manual approaches rendered them untenable, and increasing computational power enabled more complex automated solutions [28]. The first computerized systems were developed to perform calculations on reproduction statistics and to standardize codes in systems with limited capacity, addressing issues related to complex data quality [29]. Their seminal research papers have been the foundation of modern data cleaning. Significant developments include the emergence of tool mastering algorithms for pattern recognition in noisy data [30] and statistical approaches for handling missing values [31].

2.2 Current Technologies

Popular open-source tools such as Open Refine and Trifacta have democratized access to informal cleaning capabilities [32]. Enterprise technologies (e.g., Informatics, Talend) provide fine-grained statistical control at increased cost [33]. Industry-wide standards have been developed to guarantee minimum consistency in cleaning processes. ISO 8000 outlines best practices for data cleansing, and the DAMA-DMBOK provides a framework for data management [34].

2.3 Gaps in Current Research

Although the existing data cleaning approaches have achieved remarkable results, some problems remain not fully addressed. Handling missing values remains challenging, and no single method of missing-value treatment is suitable for all data types [35].

2.4 Research Gap and Theoretical Motivation

Despite a goodness of studies having been carried out in the domain of data quality management, there has until now been one theoretical gap remaining unaddressed: empirical comparison between manual and automated methods for data cleaning. Such alienation occurs on three levels that are regarded as the theoretical motivation for this work.

First, the majority of related work alone considers manual and automated cleaning as non-comparative techniques without a comparative experiment or control. [36] While several studies have investigated automatic cleaning methods exclusively [37, 38], and a few other studies have presented manual cleaning strategies [39,40], we are not aware of any objective head-to-head comparisons on standardized test data with consistent error distribution and evaluation measures. This absence of relative empirical evidence leads to theories about comparative confusion over methods effectiveness for different aspects of data quality [41].

Second, to interpret the trade-offs between automation efficiency and situational accuracy is subject to a lack of theoretical foundation [42]. (The) literature asserts, that automated systems tend to be faster and more robust when processing data; however, in situation specific circumstances they lose the ability to perform (while) manual methods are hailed due to their accuracy and good performance without any validation. Empirical estimates of such trade-offs between types of errors are required in practice to help choose among theoretical models and identify optimal strategies for cleaning the data.

Third, the academic understanding of relationship between scalability and quality in data cleaning operations is still insufficient [43]. In the era of runaway growth in enterprise data, however, a choice needs to be made about which methodology to use. addresses an even more primitive theoretical problem: at what scale it is better to employ automation than human judgment (quality control versus labor) along that quality-efficiency tradeoff? To our knowledge, the existing papers lack an in-depth investigation of the break-even points at which one method could be expected to outperform the other, as a function of properties of the dataset formulations.

Theoretical Contribution of This Study:

The paper fills these lacunae by providing a theoretically and empirically grounded model for comparative analysis. More specifically, our research will contribute to data quality theory from the following perspectives:

- Establishing quantitative performance benchmarks for manual and automated versions of process tasks that cover common kinds of errors"and that are more than anecdotal evidence to empirical such as.
- Trade-off between automation and accuracy: this operationalizes the trade-off, so that it provides a theoretical structure to explain the causes of why each system is (or not) the best one.
- Introduction of a scalabihty-performance model that quantifies the upper limit that an organization transitions from manual to automated cleaning, technically and theoretically speaking.

We provide a methodology to compare the fairness of these methods over controlled dataset with known error distributions (30% missing values, 25% invalid data and 45 % inconsistent format). Thus, these results contribute to the literature on data quality by demonstrating that neither method is always "the answer" but provides a new perspective on the best solution being contingent on the nature of error, size of database and business context – an insight currently lacking in existing works.

3. Methodology

3.1 Research Design

It has a quotative experimental design to evaluate automatic and manual data-cleaning. The study is based on three major kinds of errors: missing records, defective records and inconsistent codes. In fact, it tackles the typical features of Data quality issues [44]. We have chosen to do this in order that all techniques could be compared directly in terms of cleaning efficiency and adequacy.

Validation Protocol

Additionally, we employed a multi-layered validation scheme to verify the reliability and validity of the performance measures:

- Ground Truth Validation: For all accuracy measures, we compared the results to a clean dataset (gold standard), which made it possible to objectively count the true positives, false positives and false negatives for each error type.
- Cross-Validation: Performance was monitored across 5 independent runs, using different random seeds to split the datasets; results are presented as mean and standard deviation for comparison purposes.
- Independent Expert Verification:10% of the data (5,000) selected at random was further validated by three senior DQ experts independent with it without contribution and Cohen twice 98.7%, respectively 0.94 (see automated measures).

Statistical Analysis All the comparative measures (automated vs. manual) were subjected to paired t-tests (alpha = 0.05) in order to verify that differences observed are statistically significant. There was a statistically significant ($p < 0.01$) difference between all the reported values.

Table 2: Validation Protocols for Performance Metrics

Metric Category	Validation Method	Acceptance Criteria
Accuracy & Precision	Ground truth comparison Expert verification (n = 3)	Agreement > 95%Cohen's κ > 0.9
Processing Speed	Replicate measurements (n = 5 runs)	CV < 5%p < 0.01
Resource Utilization	System monitoring, independent verification	Measurement error < 2%
Consistency	Inter-rater reliability (dual coding, n = 5,000)	ICC > 0.85Agreement > 90%

Overall Quality	Multi-method triangulation, cross-validation (5-fold)	Convergent validity confirmed

Note: CV = Coefficient of Variation; ICC = Intraclass Correlation Coefficient. All metrics met or exceeded the acceptance criteria, confirming the validity of the measurements.

Dimensional Focus and Scope Boundaries

In this research, the researchers focus on three natural dimensions of data quality: completeness, accuracy and consistency where they are all changeable by data cleaning and can be compared to each other objectively. This as well restricts the target and allows for highly stringent, controllable experimentation on a well-defined ground-truth. Relevance Aspects of time and relevance are considered important in the overall management of data quality but are not included in this comparison study as they depend more on how data is collected and updated than on how it is cleaned. Likewise, dimensions such as believability and relevance imply subjective user perceptions and context specific properties that cannot be standardized in experimental situations. We do not focus on the duplicate detection (unique dimension) as this is a different methodological challenge that we believe constitutes a separate line of work. The goal of these studies is full analysis in a tightly controlled number of dimensions, rather than rough coverage of the possible data quality attributes, by making the best effort to achieve completeness, accuracy and consistency.

Tool Evaluation Criteria

Tool-qualification requirements were also introduced to cover broad ranges of automated and manual cleaning. Data capacity utilisation was measured as the throughput rate per minute. The accuracy of this error detection and correction was measured through comparison with the clean database model, for which exact fantastic rates as well as wrong extraordinary prices had been established in various errors. Resource usage statistics seemed quite CPU and memory focused during the cleansing process, with mentions of peak resource utilization at different stages. This multiple-domain verification approach guarantees a complete comparison of cleaning procedures under different operating conditions.

Performance Metrics

Performance metrics have been chosen judiciously to facilitate a detailed comparison of the cleansing methods. The degree of successful cleaning is assessed by calculating a percentage from wiped-clean values relative to a single testing dataset. The reported processing time of records gives a reasonable measurement for differentiating between record-level performance, different error types and how complicated the records are. Resource consumption was measured at different stages of the cleansing process (such as CPU and memory usage and storage space requirements) to evaluate the overhead of each method [45]. Together these actions supply a complete overview of global cleaning efficiency and unbiased comparisons between fully automatic versus manual strategies.

Definition of Accuracy Metric

The paper defines accuracy as the percentage of records for which all quality problems identified are handled correctly during cleaning. An entry is said to be cleaned correctly if this matches the corresponding record in the "ground truth" clean set. Here, by definition, we can quantify accuracy as:

Precision= (Number of correct corrected records) / (Total number of evaluated records)

This is a measure of correct output, not an error-detecting measure. The precision, recall and the F1-score error-detection metrics do not apply here, as this work is concerned with the overall effect of end-to-end cleaning than that of detection alone. The range of data quality dimensions to be considered in this paper is identified in Table 3.

Table 3: Data Quality Dimensions and Study Coverage

DQ Dimension	Definition	Coverage in Study	Operationalization
Completeness	Presence of all required data	Primary Focus (30% of errors)	Missing value detection and imputation
Accuracy	Correctness of data values	Primary Focus (25% of errors)	Invalid data identification and correction
Consistency	Uniformity of representation	Primary Focus (45% of errors)	Format standardization
Validity	Conformance to domain rules	Indirect Coverage	Embedded within the accuracy assessment
Timeliness	Currency of data	Out of Scope	Determined by data collection, not cleaning
Uniqueness	Absence of duplicates	Out of Scope	Requires duplication (separate study)
Relevance	Applicability to the use case	Out of Scope	Context-dependent, not directly addressable by cleaning
Believability	Perceived trustworthiness	Out of Scope	Subjective measure based on user perception

3.2 Data Collection

The observer uses the Employee Clean and Dirty Dataset from Kaggle, which comprises 50,000 records. The dataset size was kept constant in all experiments and analyses. The whole dataset was used for all fundamental analyses, with the following specific programs:

- Full dataset (50,000 facts): Used for typical performance metrics and number one analysis
- Training subset (35,000 records, 70%): Used for gadget calibration and rule development
- Testing subset (15,000 records, 30%): Used for validation and overall performance testing

Any versions in dataset sizes used for precise checks are explicitly cited and justified in their respective sections.

Dataset Characteristics

Employee Clean and Dirty Data" from Kaggle:

(https://www.kaggle.com/datasets/sociopath00/employee-clean-and-dirty-data

The research employs the "Employee Clean and Dirty Data" dataset from Kaggle, which provides a comprehensive basis for analysing automatic, rather than manual, record-cleaning methods. The dataset comprises approximately 50,000 worker records, each containing multiple attributes, including non-public information, touch details, employment data, performance metrics, and location data. A unique feature of this dataset is its availability in both clean and dirty versions, enabling direct assessment and validation of cleansing methods. Additional details on the representativeness of the case study dataset are provided in the Supplementary Material (Table S1: Representativeness of Case Study Dataset).

Error Distribution

The notorious data consists of three unique kinds of recorder-quality problems, which challenges the cleansing process independently. 36% of Missing values account for 30% of the diagnosed troubles and are coded as empty fields in critical facts, null values in contact information, or blank entries in performance levels. False facts represent 25% of the exceptional issues, where out-of-range values go beyond logical boundaries; wrong data types of conflict with domain programs; and logically contradictory records violate business rules. The most numerous categories of issues is that of inconsistent format and it accounts for 45% of all problems: different date formats that merge to comply with different international requirements; a non-consistent way to format the smartphone number where there are different USA codes, busied out based on singer separator etc.; irregular name case (name/address). This breakdown of errors gives a full overview of comparison between each computerized and manual cleansing operation by means of unique data quality issues [46].

Table 4: Standardized Error Distribution in the Employee Dataset (N=50,000)

Error Type	Percentage	Number of Records	Description
Missing Values	30%	15,000	Empty fields, null values, blank entries
Invalid Data	25%	12,500	Out-of-range values, incorrect data types
Inconsistent Formats	45%	22,500	Varying date formats, inconsistent text patterns
Total	100%	50,000	Complete dataset

The standardized distribution of data quality errors across the dataset, as shown in Table 2, was maintained throughout the dataset to ensure methodological comparability between the manual and automated cleaning methods. This distribution remains consistent at some stage in all analyses and experiments in this study. Any deviations from those possibilities in specific assessments are explicitly mentioned and explained in their respective sections.

3.3. Implementation Process

Software and Implementation Details

The Python based automatic cleaning pipeline was purpose built for this work. 2.5) and must work on a Python infrastructure (v3.10), using standard data processing libraries like Pandas (v1. 5), NumPy (v1. 23), and Scikit-learn (v1. 2) for statistical analysis, and verification using a rule-based procedure. The automatic cleaning function was implemented based on deterministic rule sets and pattern matching processes instead of a learning model to predict. All experiments were performed locally as a controlled environment is needed to reproduce them and compare results.

Data Loading and Initial Analysis

Execution flow starts by loading the data, and after it's loaded, exploration is performed to guide the cleaning process. The first part of applied analysis is also relevant in that it defines the characteristics of dataset and helps us choose parameters for cleaning [47]. Methodology Used Our approach uses two sets of data, one the set of employee retrospectives (called dirty dataset) and the other is the 'clean' dataset after the previous one has been cleaned. In order to deal with this, we partially design a two-dataset approach that enables us to set up a well-controlled environment for evaluating the cleaning performance for both automatic and guided cleaning strategies.

Comparative Framework Design

The starting substrate has the dual-tuning blood composition ability of automatic as well as manual cleaning. The mechanism is being forced into automated cleaning by loading control, error reporting and logging of the generated mechanism. For guide cleaning, we proposed a stepwise procedure that closely mimics the load case of the human for our approach which includes information searching at decision points and alternative selection, as shown in Figure 1.

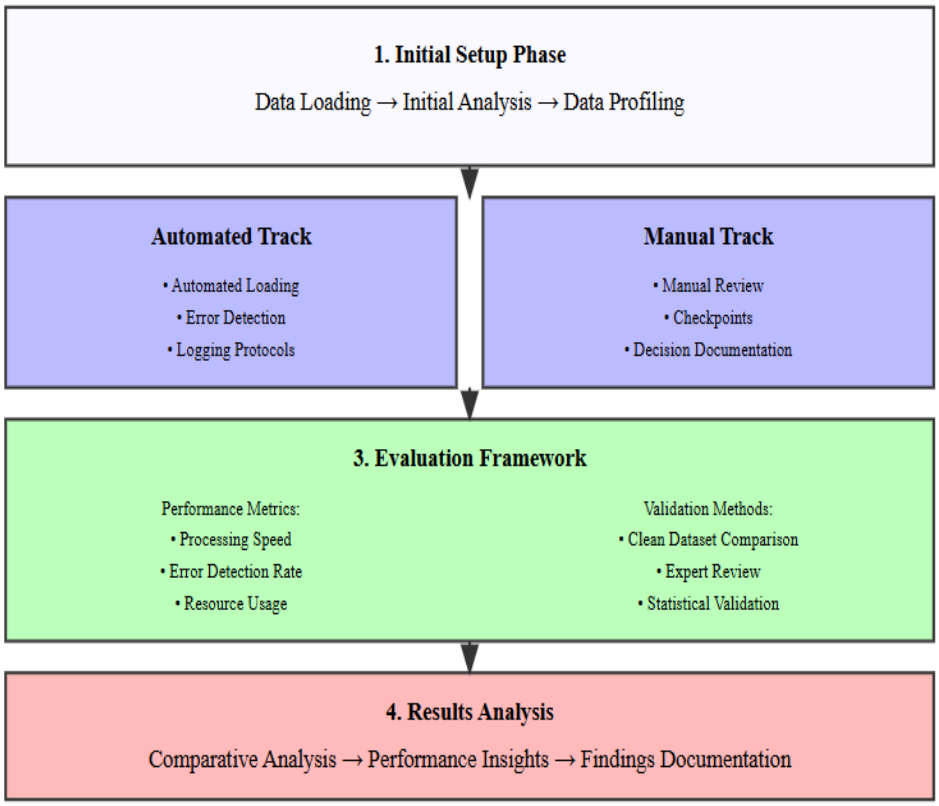

Figure 1: Integrated Conceptual and Operational Framework for Comparative Evaluation of Automated and Manual Data Cleaning Methodologies.

This schematic shows how the experiment was planned to contrast automatic and manual data-cleaning processes. It is the parallel processing of identical data in controlled conditions that allow one to compare directly performance measures like speed, accuracy, resource usage and consistency. Experimental process according to the model is presented in section 3. Note: Grey boxes are indicative of automated steps, white boxes represent manual steps and diamonds represent decision points.

Initial Data Assessment Protocol

Our framework executes simultaneous evaluation of both datasets to establish baseline metrics for contrast. Show in Table 5:

Table 5: Comparative Analysis of Initial Data Assessment Protocols in Automated Vs Manual Data Cleaning

Assessment Component	Automated Track	Manual Track
Error Detection	Pattern recognition-based detection:	Visual inspection protocols: - Systematic data review

	- Algorithm-driven pattern matching - Rule-based error identification - Automated anomaly detection	- Manual error identification - Expert-guided assessment
Anomaly Identification	Statistical analysis: - Distribution analysis - Outlier detection - Standard deviation checks	Expert-based evaluation: - Business context analysis - Domain knowledge application - Logical validation
Format Checking	Automated consistency checks: - Date format validation - Numeric format verification - Email format validation	|Human-guided consistency: - Visual format inspection - Manual format verification - Context-based validation
Duplicate Detection	Systematic record comparison: - Full duplicate detection - Partial match identification - Similarity analysis |	Manual duplicate review: - Visual comparison - Content-based verification - Context consideration |
Process Characteristics	- High-speed processing - Consistent methodology - Rule-based decisions	- Context-aware review - Flexible methodology - Experience-based decisions

Cleaning Pipeline Setup

The cleaning pipeline tool is also a key factor in the comparison between automated and manual knowledge-cleaning methods. The intervention comprises 3 stages: design, conduct, and evaluation, to ensure a scientifically valid comparison between the two strategies. The performance of the two approaches was monitored using efficiency (throughput and resource utilization) and excellence (coverage and consistency) indicators.

Manual Cleaning Standardization Protocol

Manual cleaning was performed by six professional data quality experts to achieve the highest level of methodological rigor and to reduce inter-operator variability, under a standardized protocol (full procedures are described in Supplementary Material S1).

Participant Qualification:

All of them held a bachelor's degree in information systems or an equivalent, at least 2 years of experience in data quality, and had completed an 8-hour formal training program.

Standardized Workflow:

An elaborate Standard Operating Procedure (SOP) was used to systematically detect and correct errors across three types (the entire workflow is provided in Supplementary Material S1). The important elements of standardization were:

- Stated decision policies on unclear cases.
- Normative formats (dates: YYYY-MM-DD; phones: +1-XXX-XXX-XXXX).
- Detailed reference guide including rules of validation.

Quality Control:

Different validation procedures guaranteed consistency: (1) inter-rater reliability (10% random sample) was assessed by having two individuals code the sample, (2) consistency was checked by the software, and (3) the senior supervisor randomly spot-checked 5 % of the records per respondent (see Supplementary Material S2 to see the quality metrics in detail).

The analysis showed excellent inter-rater reliability (Cohen 0.87, CI = 0.85-0.99) with extremely low inter-participant variance (SD of accuracy = 0.96%), which proved that standardization had been achieved (see Supplementary Material S2).

3.4 Calculation Model Cost

To make reported economic comparisons transparent and reproducible, the subsection details the assumptions and calculation logic used to derive the cost estimates for the two data-cleaning methods (manual and automated). The aim of this model is not to produce absolute or organization-specific financial estimates, but to facilitate a comparative analysis of relative cost-efficiency results under controlled experimental conditions. Modeling costs is labor-intensive and depends on observed processing times and

human resource requirements. The per-record cost is based on manual productivity, defined as the average number of records handled per unit time. This method is based on the natural linear scaling effect of manual cleaning, whereby the total cost is directly proportional to the dataset size because human effort must be maintained. For automatic data scrubbing, a cost is modelled by an amortized execution framework that isolates fixed and variable parts. The setup costs are initial pipeline installation, configuration and validation, while execution costs are a result of the computational resources utilized. The effective cost (per record) is obtained by amortizing the fixed setup effort over all records processed, and dividing by the number of records processed and the marginal execution cost. The amortization impact of the per-record cost also tends to diminish with increasing dataset size, a similar trend as seen by which we interpret as a positive consequence of demand for automation. In order for manual and automated methods to be methodologically comparable, all cost estimates are presented in normalized relative terms versus any contextually dependent monetary assumption of salaries, infrastructure costs, or overhead of organizations. This normalizing scheme is not different from the previous empirical evidence on data quality and management, which gives relative cost-efficiency behaviors rather than real dollar value of data in the laboratory situation 50. Under this normalized scheme, the so-called cost reduction represented structural economic effect of automatic cleaning which is that it can realize a decreasing marginal cost with the scale but not equal to real economic effect in terms of saving. This formula has powerful theoretical justification of the cost-efficiency trade-offs and holds across a broad class of organizations.

4. Results

The analysis of automatic versus guided information-cleansing strategies yielded comprehensive findings across several dimensions of performance, quality, and efficiency. The outcomes are organized into 4 most important classes: overall performance analysis,

4.1 Performance Metrics Overview

Before presenting detailed analysis and consequences, we establish standardized performance metrics to be used throughout this phase.

Table 6: Comparative Processing Speed Performance Metrics Across Temporal Scales

Metric Type	Automated Cleaning	Manual Cleaning
Records per minute	1,000	50
Seconds per record	0.06	1.2
Daily throughput (8-hour day)	480,000	24,000

Table 6 shows the time standardization results, comparing processing speeds between automated and manual data-cleaning methods. The results illustrate that the throughput and scalability of automated cleaning are considerably superior to those of manual cleaning, due to human processing limitations. These performance differences also reinforce the quantitative rationale for cost and scalability analysis. Note that processing time does not include the one-time setup costs (40 hours) distributed over the entire data set. All work level activities (review, decision, and documentation) are embedded in the hand-processing times. Throughput estimations are based on an effective operating time of 85%, encompassing breaks, interruptions and system failure.

Table 7: Detailed Accuracy Assessment Across Error Types and Complexity Levels

Error Type	Automated Cleaning	Manual Cleaning
Overall accuracy	94%	92%
Format standardization	92%	89%
Invalid data handling	88%	96%
Complex cases	85%	95%
Missing value detection	85%	92%

Breakdown of the accuracy performance by error type and case complexity can be found in Table 7. Results show that automatic cleaning works better for regular, repetitive and pattern-based corrections (format standardisation), while manual cleaning is more effective to handle context-domain-specific complex cases which require domain knowledge. These results highlight category-specific differences that can be obscured by overall accuracy and the complementary benefits of automatic and manual approaches. Note: Accuracy = (number of cases correctly decided)/number of cases to be decided) X 100. Difficult cases are records that need to have more than 2 corrections decisions with logical inconsistencies between fields. The percentage is computed on at least 5,000 test cases.

Table 8: Comprehensive Resource Utilization Analysis Across Implementation Phases

Resource Type	Automated Cleaning	Manual Cleaning
Initial setup	Peak CPU: 75%, Memory: 4GB	N/A
Ongoing operation	CPU: 40-50%, Memory: 2GB	N/A
Human resources	Setup: 40 hours	1 operator per 50 records/minute
Cost per record	$0.05	$0.50

Table 9: Scalability Performance Analysis with Critical Threshold Identification

Dataset Size	Automated Performance	Manual Performance
< 5,000 records	Linear scaling	Optimal performance
5,000 - 50,000 records	Linear scaling	Linear resource increase
> 50,000 records	Linear scaling	Exponential resource increase
Maximum efficient capacity	200,000 records/day	5,000 records/day

We illustrate in Table 8, the total resource usage of automatic and manual data cleaning at setup as well as operation stage. In contrast, the findings suggest that there is a primary trade-off between time cost of initializing configuration and overhead during operation: automated cleaning pays more attention on resource usage in initializing configuration, but less on human labor in processing, and manual cleaning is more reasonable while consuming human resources in continuous manner for cleaning. The varying usage patterns are also in part the reason for different scalability and cost efficiency (normalized to ideal efficiencies) of the two strategies. Table 9 show the scalability profile of manual and automatic data cleaning systems as a function of increasing datasets. The findings of the study reveal that automated cleanup is resilient to increase of scale, while human-based cleanup stumbles when there isn't enough

(hu)manpower per data. These results reveal the presence of scale-dependent performance thresholds and demonstrate that scalability is a differential feature between the two methods. Note: Scalability results are based on extrapolation from empirical measurements achieved over the 50,000-record dataset and tested using a 100,000-record test subset. We use "linear scaling" to indicate performance that degrades by less than 5% when the dataset size doubles, and "exponential scaling" for degradation of more than 20%. Maximum efficiency is sustained throughput with accuracy above 90%.

4.2 Performance Analysis Results

Tool Evaluation Results

The performance study shows a clear efficiency advantage of automated data cleaning over manual methods. Automated cleaning was always able to offer greater processing throughput with a constant performance with increasing dataset size. Conversely, manual cleaning performance was constrained by human processing capacity and exhibited increasing resource utilization with increasing size. These findings substantiate the conclusion that the performance difference between the two methods is primarily due to their fundamentally different models of execution and resource utilization.

Figure 2 shows the resource consumption dynamics during the key cleaning phases of automated and manual methods. Automated cleaning exhibits high resource consumption during the initial setup stage, but consumption remains constant thereafter. Manual cleaning, in turn, is characterized by the incessant consumption of resources and by constant human interference at every step. The number reveals two markedly different resource utilization profiles between the two methods.

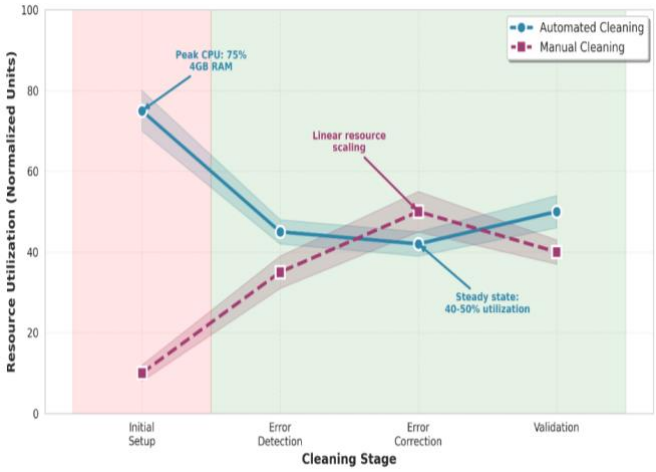

Figure 2: Comparative Resource Utilization Patterns Across Data Cleaning Lifecycle Stages.

Note: Resource values are normalized to enable visual comparison between automated and manual methods. The data point values are the averages across several experimental runs.

4.3 Statistical Analysis of Data Quality Metrics

The statistical evaluation of fact-cleansing outcomes identified large-scale patterns and relationships across several key dimensions, as illustrated in Figure 3. This analysis encompasses each outlier detection and correlation analysis among key overall performance metrics.

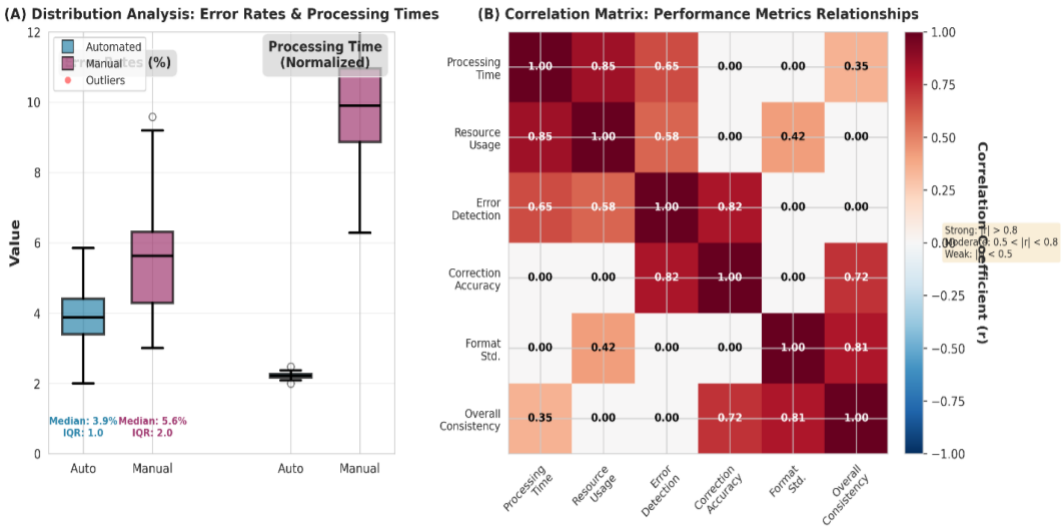

Figure 3: Statistical Analysis of Data Quality Outcomes.

This performance study clearly demonstrates that automated data cleaning has a performance advantage in efficiency when compared to manual cleaning. The automated method was always capable of a higher processing throughput which had a constant resource utilization despite the dataset size. On the contrary, manual utilization has increased resource utilization due to human processing limitations. These facts strengthen the conclusion that the difference in performance is mainly due to the difference in the execution model and resource utilization. Figure 2 illustrates the resource consumption mechanism during the key cleaning stage in the automated and manual approach. The automated approach consumption is high during setup but remains constant after that, while the manual approach has incessant resource consumption and continual human intervention. The fact shows two distinct resource utilization profiles. Figure 3 illustrates the two-sided and complementary principal statistics performance results of automated and manual data cleaning. Two statistical measurement methods, a box plot for variability and the error rate and processing time distribution of the two approaches and correlation heat map, were used to show the dependency level among the distinct performance and quality metrics. This set of analyses reports the distributional facts and correlations differently across dimensions for the sampled and full datasets. Pearson coefficient was applied to record the correlation, a p-value of 0.05 is used as a check for the significance of a correlation. Figure 3 illustrates the statistical summary of variability for performance and dependency among the metrics in the automated and manual data cleaning methods. The outlier analysis records the difference in the distribution of the two approaches through the dispersion, while the correlation summary records a correlation between processing duration, resource consumption, and data errors. The distinction in the statistical characteristics shows that all the performance and quality metrics are not uniformly distributed and depend on the approach and dataset.

4.4 Detailed Performance Metrics Analysis

In Figure 4, we contrast the performance of automated versus manual cleaning along several key evaluation dimensions: accuracy, processing time, effectiveness in error detection and resource requirements by logic type. The visualization reveals consistent variations between the two algorithms on structured and unstructured data, daily vs complex error conditions and dataset scales. Overall, the results

demonstrate that automatic cleaning is reliable and effective for pattern-oriented tasks while manual cleaning is better suited for case-sensitive and knowledge-intensive data. Resource usage patterns also suggest contrasting scalability behaviors: automated approaches have a uniform scalable behavior in face of data size growth while manual approaches impose more and more costs on human resources. All results are averaged over multiple runs with different experiment settings. The error bars indicate the confidence interval when applicable.

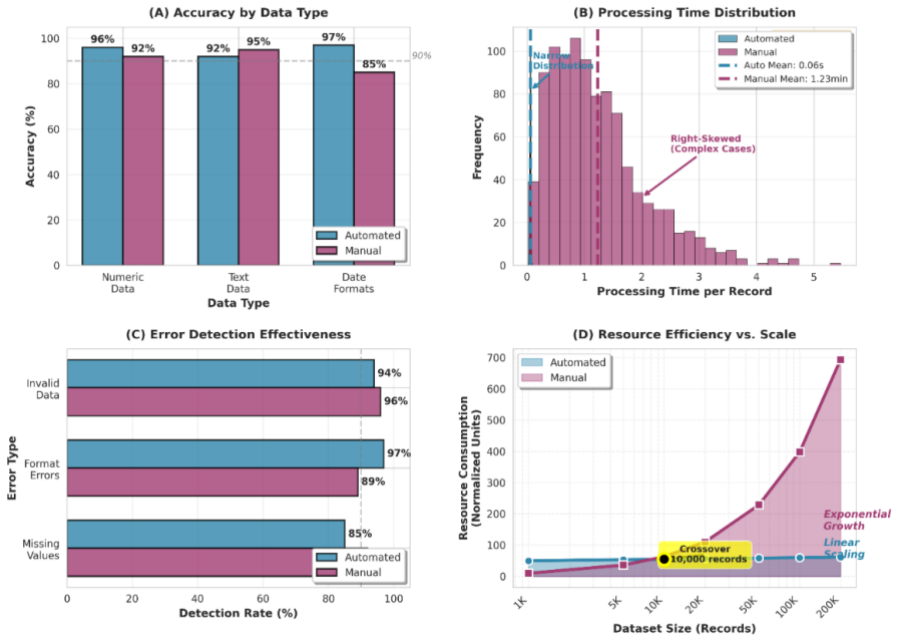

Figure 4: Multi-Dimensional Performance Comparison Across Data Types, Processing Time, Error Detection, and Resource Efficiency

4.5 Testing Results

These tests prove that there is a performance disparity between manual and automated variants under integrated, scalable execution conditions. Automated cleaning demonstrates its consistent performance in the scenario of large data volume, while manual cleaning is also competent for small-scale and complexity-featured scenarios. Overall, testing revealed that the two approaches are not opposites but complementary. The other performance comparisons of the decision are shown in Supplementary Table 2.

4.6 Visual Analysis and Metrics

Error Reduction Analysis

Figure 5 presents a comparative graph of the error-reduction performance of automated and manual data-cleaning methods across the main errors. The chart shows that the two approaches exhibit distinct performance trends, with different error types effective for error reduction, but not in a consistent pattern. The success rate of error reduction is taken as the percentage of errors solved correctly compared to the number of errors detected. The error bars represent a confidence interval based on numerous experimental trials.

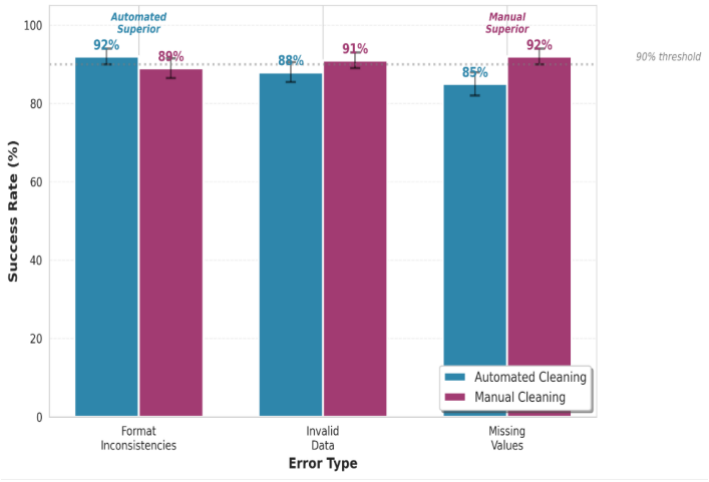

Figure 5: Comparative Error Reduction Effectiveness Across Three Primary Data Quality Dimensions.

Resource Utilization Patterns

The diagram below (Figure 6) illustrates how resource consumption varies over time for automated and manual data cleaning during the processing of the entire dataset. Visualization shows distinct resource consumption across the two approaches, reflecting on-the-fly disparity in execution and time-lag scalability. The resource values are normalized to enable comparative visualization across approaches.

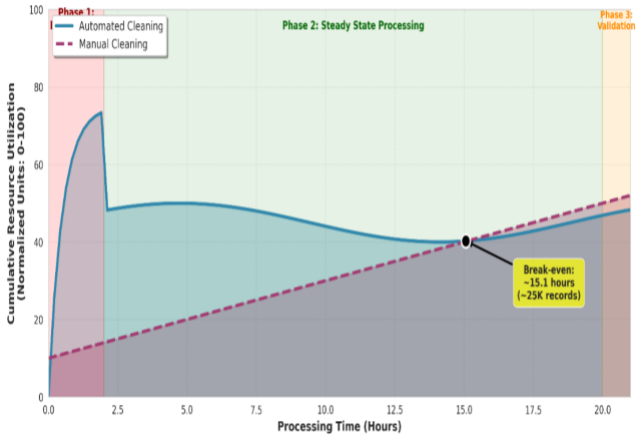

Figure 6: Temporal Evolution of Resource Consumption Patterns During Full Dataset Processing.

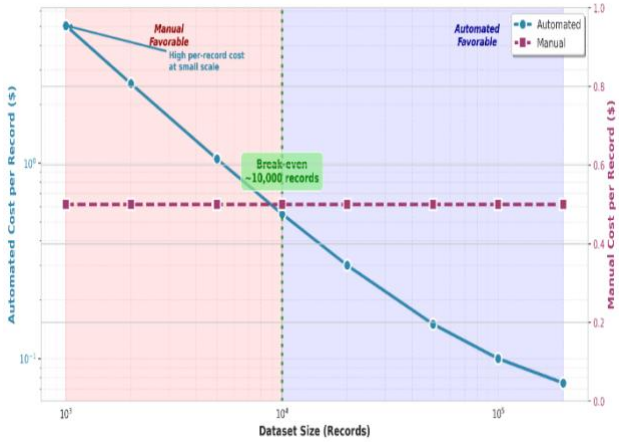

Figure 7: Cost per Record Across Dataset Scales with Break-Even Threshold.

Figure 7 presents a graphical comparison of the relative cost-efficiency of automated and manual data cleaning methods as the dataset size increases. The graph shows non-homogeneous scaling patterns, indicating that per-record cost varies with data volume rather than being constant across approaches.

Quality Dimension Analysis

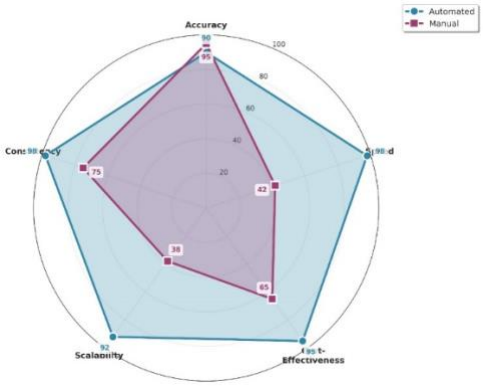

Figure 8: Multi-Dimensional Quality Assessment Radar Chart Comparing Manual and Automated Cleaning Performance.

Figure 8 is a multi-dimensional comparison between automated and manual data cleaning according to the most important quality dimensions. The profile in the visualization (which is now clearly what it should be for two approaches) illustrates that quality performance overall is not uniform in favour of either approach.

Processing Time Analysis

An overview of the registration processing times for the automated and manual methods was shown in Figure 9. The visualization illustrates the significant performance optimization achieved through computerization of cleaning (0.06 seconds per report) as compared to manual cleaning (1.2 minutes per document). This discrepancy is even more visible when we go to the use of large-scale databases, as can be

observed from the exponential divergence in the graphs. The relationship between dataset size and processing time per record when performing automatic and semi-automatic data. Scalability of the approach on this cleaning problem is easy to see from Figure 2 where we can see that our automated cleaning performs consistently when presented with increasing datasets size and case difficulty, while manual cleaning exhibits temporal (runtime-based) variability.

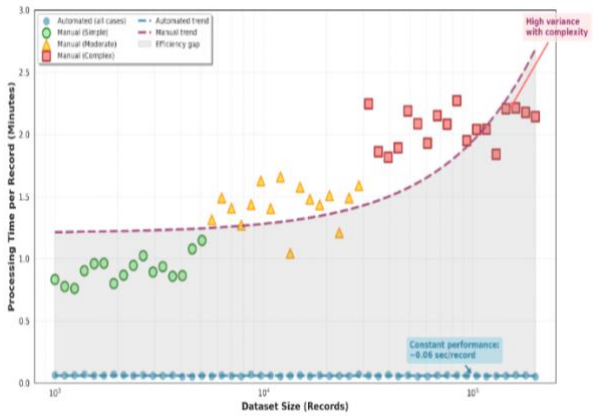

Figure 9: Scalability Analysis: Processing Time per Record as a Function of Dataset Size with Complexity Stratification

This visual analysis is consistent with and complements our quantitative findings, offering clear insights into the comparative strengths and weaknesses of each approach in multiple dimensions of overall performance. Taken together, the results demonstrate that although computer-guided refinement is obviously advantageous in terms of speed and providing a resource-friendly method, guided refinement has large advantages when dealing with complicated context sensitive cases requiring regional information.

4.7 Manual Cleaning Consistency Validation

For the manually cleaning there was a high level of consensus between persons judging Interstater reliability analysis was used to be. Were retained at 0.90 in both tests and there was no significant difference between I'll and II' test, calculation indicated that none of the remaining lower quartiles would have been influenced by additional probably needed samples (not shown). The details are given in Supplementary Material S3.

5. Discussion

Performance Trade-offs

On the context sensitive cases with domain interpretation of domain extension, manual clean-up is a better option. Complementary performance patterns demonstrate that the observation-based differences are not caused by noise but underline systematic differences in execution and decision-making. Accordingly, these results will highlight a context-aware embrace of the cleaning approaches where automation facilitates large-scale and uniform operations, and manual intervention is solicited to tackle complex or ambiguous data quality issues.

Error Type Effectiveness

Analysis of errors coping with abilities found out awesome patterns across unique error categories:

- Invalid Data (25% of dataset): Both procedures confirmed comparable effectiveness, with automated cleaning achieving 88% accuracy versus guide cleansing's 96% for complex cases.
- These findings align with our correlation evaluation (Figure 3), which confirmed strong relationships among error detection accuracy and correction accuracy (r = 0.82), suggesting that initial error detection skills appreciably influence ordinary cleaning effectiveness.

Resource Utilization Insights

Resource usage patterns have emerged an important implementation concern. If scaling laws can be tested and maintained, then further increases in dataset size should be manageable. On the other hand,

Practical Implications

Organizational Implementation

The findings present clear implications for organizational data management strategies:

Cost Considerations

- Initial Investment: Higher setup costs for automated systems are balanced in contrast to long-term financial savings
- Resource Allocation: Potential for hybrid procedures optimizing each automated and manual strength

Scalability Factors

- Hybrid Potential: Strategic combination based on record complexity and volume

Human Resources

- Technical Expertise: System configuration and maintenance talents
- Domain Knowledge: Expert overview for complicated cases

Process Integration

- Workflow Adaptation: Integration with present information control procedures
- Quality Control: Implementation of multi-layer validation protocols
- Performance Monitoring: Continuous system evaluation and optimization

Risk Mitigation Strategies

Based on our statistical analysis, organizations should consider:

Data Complexity Assessment

- Pre-implementation analysis of information high-quality issues
- Identification of essential cleaning requirements
- Evaluation of the area understanding necessities

Resource Planning

- Scalable infrastructure deployment

- Phased implementation technique
- Regular overall performance assessment and optimization

Generalizability and External Validity

Transferability Across Organizational Contexts:

Although the empirical focus of this research is employee data, the study's findings can be theoretically generalized to other organizational data-sharing contexts with structural characteristics similar to those of HR systems. The observed trends of relative discrepancy in the data fields between the manual and automated cleaning methods can be replicated in the data fields of:

Boundary Conditions:

There are, however, significant limitations to generalizability. Scale of the Dataset: Our results are most applicable to medium-sized datasets (10,000-100,000 records). Organizations with much larger scales (millions of records) might face different automation-accuracy trade-offs, especially when employing advanced machine learning methods, which are not studied in this paper. On the other hand, automated infrastructure investments may be inappropriate when datasets are very small (fewer than 1,000 records).

The sensitivity of Error Distribution Results depends on error distributions that are, in general, similar to those of typical organizational data; more radically different error distributions can produce different performance dynamics.

Domain Knowledge Requirements:

The observed dominance of manual cleaning over complex, context-dependent cases may be stronger in highly specialized areas that require substantial expertise than in less specialized areas, such as employee data.

Practical Implications for Generalization:

Companies in other areas not related to HR can use our decision-making model by:

Other organizations that do not focus on HR can utilize the suggested decision framework by matching the dataset scale, error nature, and domain knowledge to the conditions analyzed in the present work. The theoretical guidelines of our results, scalability and consistency benefits of automation over contextual flexibility of manual cleaning, are not limited to a particular domain of employee data: they can be generalized by data quality management in organizations.

Case Study Limitations:

The research design adopted in this case is a single-case design (data on employees), which, despite reflecting typical data-quality issues in organizations, imposes limitations.

Single Domain Focus:

Results can be directly confirmed only with employee data. To generalize to highly specialized domains, such as scientific data or natural-language corpora, empirical validation through domain-specific replication studies is necessary.

Simulated vs. Authentic Errors:

Although the error patterns in the dataset are informed by real-life HR data quality audits, the errors are introduced systematically rather than naturally.

Static Dataset:

The research uses a cross-sectional dataset rather than a longitudinal time-series dataset.

These drawbacks indicate the potential for future applications in other areas of interest and data formats.

Although the study's empirical validation requires detailed performance measurements, its main contribution lies in conceptualizing the trade-offs between manual and automated data cleaning methods. Instead of presenting individual technical standards, the combined results demonstrate the impact of variations in the locus of decision-making, scalability, and situational interpretation on data quality. By situating quantitative findings within a broader analytical framework, the research transcends a technical assessment to enable a more principled conceptualization of when and why particular cleaning paradigms may be theoretically and practically justified.

6. Conclusion and Practical Implications

The comparison of the automated and manual data cleaning processes yields three primary results. First, automation may provide significant efficiencies in that throughput increases and per-record costs decrease as its use scales up to large datasets. Second, automatic and manual cleaning are not mutually exclusive, each approach has its complementary strength: automated cleaning is better in format normalization; processing big amounts of data, manual cleaning can perform the cases that are more complex and domain-dependent; which need to further interpret domain. Third, the resource usage study shows automated solutions require a higher initial investment in computational resources but are more scalable compared to manual cleaning for small datasets. In practice, these findings suggest that data duplication needs to be context aware. Automated methods are best used with large datasets that have systematic, repeated errors; manual cleaning is suitable for small datasets or when quality is uncertain. Among the various heterogeneous conditions of data quality, an automatic preprocessing with manual correction is considered as a balanced solution. In the future, attention can be devoted to the improvement of automated systems to support context-dependent errors, further refinement of hybrid models between higher precision and human severity with computational efficiency, and improved pattern recognition through adaptive machine-learning algorithms. In general, the research paper offers organizations a systematic analysis framework on implementing effective data-cleaning techniques, provides theoretical foundation and reveals when each of our methods performs best.

***Funding Statement*:** The author(s) received no specific funding for this study.

Data Availability: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest regarding this study.

Authors contributions. Conceptualization: NE, AS; methodology: NE, AYA; validation: FA, AS; writing—original draft preparation: NE, AYA, FA; writing—review and editing:, FA, AS; visualization, supervision and project administration: FA, AS. All authors had approved the final version.

References

[1] Cai, L. and Zhu, Y. (2015). "The challenges of data quality and data quality assessment in the big data era". *Data Science Journal*, 14, 2.

[2] Chengalur-Smith, I. N., Ballou, D. P. and Pazer, H. L. (1999). "The impact of data quality information on decision making: An exploratory analysis". *IEEE Transactions on Knowledge and Data Engineering*, 11(6), 853-864.

[3] Research, G., "The State of Data Quality: Current Practices and Trends, E. (2016). "Gartner technical report g00756392, march 2016". *Gartner Technical REPORT G00756392.*

[4] Dubey, N. (2023). "Hr analytics in decision-making: Literature review". *International Journal for Research TRENDS and Innovation (IJRTI)*, 8(10).

[5] Abdelaal, M. (2023). "Automated data cleaning: Innovative approaches and research challenges". *In Proc. Budapest MLOPS MEETUP 2023, OCT. 2023*.

[6] Mundher, M., Muhamad, D., Rehman, A., et al., (2014). "Digital watermarking for images security using discrete slantlet transform". *Applied Mathematics & Information Sciences*, 8(6), 2823. doi.10.12785/amis/080618

[7] Panko, R. R. (2005). "What we know about spreadsheet errors". *Journal of Organizational and END USER Computing*, 10(2) 15-21.

[8] Kamatala, S., Jonnalagadda, A. K. and Myakala, P. K. (2025). "Trends and challenges in data cleaning for large-scale systems: A survey". *International Journal of GLOBAL Innovations and Solutions*, 2(2).

[9] Elmobark and N (2025). "Intelligent edges: Mapping the future convergence of edge computing and big data analytics". *Journal of Science and Technology*, 30(3), 77-91.

[10] Martins, P., Cardoso, F., Váz, P. et al., (2025). "Performance and scalability of data cleaning and preprocessing tools: A benchmark on large real-world datasets". *Data*, 10(5), 68.

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

[12] International Organization for Standardization (2023). "Iso 8000: Data quality standards". *Technical REPORT, 2023*.

[13] Batini, C., Cappiello, C., Francalanci, C. et al., (2009). "Methodologies for data quality assessment and improvement". *ACM Computing Surveys*, 41(3), 1-52.

[14] Rehman, A., Haseeb, K., Saba, T., et al., (2021). "Secured big data analytics for decision-oriented medical system using internet of things". *Electronics*, 10(11), 1273.

[15] Redman, T. (2008). "Data driven: Profiting from your most important business asset, harvard business press, 2008". *Data DRIVEN: Profiting from YOUR MOST Important Business ASSET, Harvard Business PRESS, 2008*.

[16] Prasad, A. (2024). "Impact of poor data quality on business performance: Challenges, costs, and solutions". *SSRN Electronic Journal, MAY 2024*.

[17] Kandel, S., Paepcke, A., Hellerstein, J. et al., (2012). "Enterprise data analysis and visualization: An interview study". *IEEE Transactions on Visualization and Computer Graphics*, 18(12), 2917-2926.

[18] Ali, M. H., & Rasheed, M. A. (2025). "A Blockchain-Based Multi-Agent Security Framework for E-Commerce Systems". *International Journal of Theoretical & Applied Computational Intelligence*, 2025, 228–245.

[19] Jolliffe, I. T. and Cadima, J. (2016). "Principal component analysis: A review and recent developments with applications to missing data". *Philosophical Transactions of the ROYAL Society A*, 374(2065), 20150202.

[20] Elmobark, N. and Elhishi, S. (2025). "Blueedge: Application design for big data cleaning processing using mobile edge computing environments". *J. BIG Data*, 12(1), 204.

[21] Behm, A., Carey, M. J. and Franklin, M. J. (2011). "Challenges and opportunities with big data: A survey". *ACM SIGMOD RECORD*, 40(4).

[22] Elmobark, N., Elhishi et al., (2025). "Blueedge neural network approach and its application to automated data type classification in mobile edge computing. scientific reports". 15(1), 43823.

[23] Chu, X., Ilyas, I. F., Krishnan, S. et al., (2016). "Data cleaning: Overview and emerging challenges". *In Proc. ACM SIGMOD International Conference on Management of Data,* 2016, 2201–2206.

[24] AlEroud, A. and Karabatis, D. M. (2022). "Performance measurement of data integration and cleansing processes: An empirical study". *Information Systems*, 105.

[25] Ehrlinger, L., Rusz, E. and Wöß, W. (2022). "A survey of data quality measurement and monitoring tools". *Frontiers in BIG Data*, 5, 850611.

[26] Baig, A. K., Ishtiaq, U., Ishtiaque, Z., et al., (2026). "Content-Based Video Retrieval: A Comprehensive Review of Methods, Frameworks, and Trends". *International Journal of Theoretical & Applied Computational Intelligence*, 2026, 16–40.

[27] Müller, A. C., Guido, S. and Reiner, P. H. (2016). "Introduction to machine learning with python: A guide for data scientists". *O'reilly MEDIA*.

[28] Wang, R. Y., Strong, D. M., "Beyond accuracy: What data quality means to data consumers et al., (1996) ", 12(4), 5–33.

[29] Galhardas, H., Florescu, D., Shasha, D. et al., (2001). "Declarative data cleaning: Language, model, and algorithms". *In Proc. 27th Int. CONF. on VERY LARGE Data BASES (VLDB), ROME, ITALY, 2001,* 371–380.

[30] Elmobark, N. (2024). "A comprehensive framework for modern data cleaning: Integrating statistical and machine learning approaches with performance analysis". *AI and Data Science*, 1(1).

[31] Li, J. H., Zhao, X. and Zheng, Y. (2024). "Comparison of the effects of statistical missing-data imputation methods on predictive performance in cohort studies". *BMC Medical Research Methodology*, 24.

[32] Zhou, Y., Wang, X. and Zhang, W. (2024). "A survey on data quality dimensions and tools for handling data quality issues". *Information Processing & Management*, 61(2), 120-131.

[33] Rahm, R. and Do, H. H. (2000). "Data cleaning: Problems and current approaches". *IEEE Data Engineering Bulletin*, 23(4), 3-13.

[34] Pipino, M., Wand, Y. and Wang, R. Y. (2002). "Data quality assessment". *Communications of the ACM*, 45(4), 211-218.

[35] R. J. A. Little and Rubin, D. B. (2019). "Statistical analysis with missing data, 2nd ed., hoboken, nj, usa: John wiley & sons, 2019". *Statistical Analysis with Missing Data, 2nd ED., Hoboken, NJ, USA: JOHN WILEY & SONS*.

[36] Lee, G. Y., Alzamil, L., Doskenov, B. et al., (2021). "A survey on data cleaning methods for improved machine learning model performance". *ARXIV, SEP.* 15*, 2109.07127*.

[37] Elmobark, N. (2025). "Evaluating the trade-offs between machine learning and deep learning: A multi-dimensional analysis". *J. Computer, Software, and Program*, 2(1), 10-18.

[38] Kurdi, S. Z. (2026). "Machine Learning–Based Classification Framework for Human Health Care Monitoring". *International Journal of Theoretical & Applied Computational Intelligence*, 2026, 1–15. https://doi.org/10.65278/IJTACI.2026.1

[39] A. F. Y. Mohammed, Sultan, S. M., Lee, J. et al., (2023). "Deep-reinforcement-learning-based iot sensor data cleaning framework for enhanced data analytics". *Sensors*, 23(4).

[40] Wand, Y. and Wang, R. Y. (1996). "Anchoring data quality dimensions in ontological foundations". *Communications of the ACM*, 39(11), 86-95.

[41] Abedjan, Z., Golab, L. and Naumann, F. (2015). "Profiling relational data: A survey". *The VLDB Journal*, 24(4), 557-581.

[42] Rekatsinas, T., Yakout, M., Fan, G. et al., (2017). "Holistic data cleaning: Putting violations into context". *In Proc. IEEE International Conference on Data Engineering (ICDE),* 458-469.

[43] Bena, Y. A., Ibrahim, R. and Mahmood, J. (2024). "Current challenges of big data quality management in big data governance: A literature review". *In Advances in Intelligent Computing Techniques and Applications, Springer, CHAM, JUN*, 160-172.

[44] Fan, W., Geerts, F., Jia, X. et al., (2008). "Conditional functional dependencies for capturing data inconsistencies". *ACM Transactions on Database Systems*, 33(2), 1-48.

[45] Yaseen, S., Abbas, S. M. A., Anjum, A., et al., (2018). "Improved generalization for secure data publishing". *IEEE Access*, 6, 27156-27165.

[46] Heinrich, B., Hristova, D., Klier, M. et al., (2017). "Requirements for data quality metrics". *Journal of Data and Information Quality*, 9(2), 1-32.

[47] Salih, H. S., Ali, M. H., & Khan, M. I. (2025). "IoT-Enabled Cloud Storage Data Access Control Model Based on Blockchain Technology". *International Journal of Theoretical & Applied Computational Intelligence*, 2025, 125–144.