Science World Journal Vol. 21(No 1) 2026 https://dx.doi.org/10.4314/swj.v21i1.39
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

Review Article

FEEDBACK-DRIVEN AUTOMATION IN DATA PREPARATION: A SYSTEMATIC LITERATURE REVIEW

*[1,2]Sa'adatu Abdulkadir, [1]Philip O. Odion, [1]Darius T. Chinyio, [1]Isah R. Saidu, [2]Muhammad A. Ahmad

[1]Department of Computer Science, Nigerian Defence Academy, Kaduna – Nigeria
[2]Department of Secure Computing, Kaduna State University, Kaduna – Nigeria

*Corresponding Author Email Address: saa.abdul@kasu.edu.ng

ABSTRACT
Data preparation remains a major bottleneck in analytics and machine learning workflows, often consuming up to 80% of the total project effort for data scientists. Although recent systems automate portions of the preparation pipeline, most lack mechanisms to capture or reuse user feedback, thereby limiting their ability to adapt to evolving data characteristics and domain-specific requirements. This study aims to systematically review the state of research on feedback-driven automation in data preparation, with emphasis on how feedback mechanisms are designed, integrated, and evaluated in contemporary systems. A systematic literature review was conducted following PRISMA 2020 guidelines, targeting peer-reviewed publications from 2010 to 2025 indexed in IEEE Xplore, ACM Digital Library, SpringerLink, Semantic Scholar, and Google Scholar. Twenty-nine primary studies met the predefined eligibility criteria and were included in the analysis. The findings indicate that feedback is predominantly utilised at schema-alignment, error-detection, and transformation stages, where explicit user input or implicit behavioural signals guide corrective processes. Feedback has been shown to improve data quality and reduce the need for repeated manual intervention; however, existing solutions are fragmented and do not support the persistence or reuse of feedback across future datasets or sessions. The review identifies a significant gap in unified architectures that treat feedback as a sustained knowledge asset. Addressing this gap presents opportunities for developing adaptive, feedback-driven data-preparation platforms that enhance transparency, reliability, and long-term automation benefits.

Keywords: Data Preparation, Feedback-driven Automation, Human-in-the-Loop, Data Cleaning, Systematic Literature Review

INTRODUCTION

The widespread adoption of data-driven systems across scientific, commercial, and industrial research has intensified demand for scalable, reliable data-preparation pipelines (Lai et al., 2021). Data preparation is a process that involves data profiling, cleaning, transformation, schema alignment, and quality validation, all of which ensure that raw, heterogeneous data can be consumed by analytical models or machine-learning workflows (Azeroual, 2020).
Several researchers have observed that data preparation consumes a disproportionate amount of effort in practical projects, frequently requiring more time and attention than the modelling or deployment stages themselves (Fernandes et al., 2023). As a result, there is growing interest in methods that improve the efficiency and accuracy of data preparation tasks, making the optimisation of this phase a recurring concern in data engineering research.
Traditional approaches to automating data preparation rely on predefined scripts, rule-based transformations, and static extraction–transformation–loading (ETL) pipelines (Paton, 2019). While suitable for stable environments, these methods offer limited adaptability to evolving data semantics, heterogeneous formats, and domain-specific constraints (Konstantinou & Paton, 2020; Somasundaram, 2022). These limitations are amplified in emerging application contexts where data schemas are frequently changing, and interpretation requires contextual or expert judgement. As a result, organisations continue to expend considerable manual effort to resolve ambiguities, correct erroneous values, and refine transformation logic.
Recent research has explored human-in-the-loop (HITL) and feedback-based approaches to address this challenge (Yin et al., 2024). In these systems, user feedback, whether explicitly provided through corrections or implicitly captured from interaction traces, serves as a supervisory signal to refine and guide preparation tasks (Konstantinou & Paton, 2020). Feedback thereby augments automated processes by enabling systems to infer intent, resolve transformation uncertainties, and adapt to previously unseen patterns (Somasundaram, 2022). This integration shifts data preparation from a deterministic activity to an iterative process in which automation can evolve rather than remain statically configured (Liu et al., 2024).
Despite promising developments, research in feedback-driven data preparation remains fragmented across disparate subdomains, including data management, interactive analytics, and machine-learning–assisted repair (Konstantinou & Paton, 2020). Existing studies differ considerably in how they conceptualise feedback, which pipeline phases employ it, and whether captured feedback persists beyond a single interaction. The absence of a consolidated synthesis limits theoretical development and complicates the design of systems capable of reusing or generalising learned corrections across datasets and sessions (Page et al., 2021).
To address these gaps, this paper conducts a systematic literature review (SLR) of feedback-driven automation in data preparation, covering publications from 2010 to 2025. The review is structured around three research questions:

RQ1: What feedback mechanisms have been employed to automate components of the data-preparation pipeline?
RQ2: At which stages of the pipeline does feedback provide measurable benefit?
RQ3: What design limitations constrain current systems, and what requirements emerge for unified, feedback-driven architectures?
The contributions of this paper are threefold. First, it provides one of the systematic syntheses of studies that explicitly integrate feedback into automated data-preparation processes. Second, it proposes a classification framework that maps feedback modalities to pipeline stages and architectural strategies. Third, it identifies

unresolved challenges and outlines a research agenda for scalable platforms capable of capturing, persisting, and reusing feedback as a computational asset. The insights derived from this review offer a foundation for advancing the next generation of adaptive data-preparation systems.

Data Preparation

Data preparation refers to the set of processes required to convert raw, incomplete, or inconsistent data into a structured representation suitable for analytical use (Azeroual, 2020). Common activities include profiling, cleaning, transformation, schema alignment, and validation, each contributing to the correctness, consistency, and interpretability of downstream analytics (Fernandes et al., 2023). As modern systems integrate heterogeneous and dynamically evolving data sources, preparation tasks increasingly require semantic interpretation rather than mechanical restructuring (Konstantinou & Paton, 2020). Within lifecycle models such as the Cross-Industry Standard Process for Data Mining (CRISP-DM), data preparation is consistently identified as the most time-intensive stage, frequently consuming a disproportionate share of project resources (Fernandes et al., 2023). Although CRISP-DM provides a structured workflow for analytics development, it implicitly assumes stable data semantics. It lacks mechanisms for iterative refinement in response to user-generated insights or contextual changes (Guduru, 2025). As preparation environments become more iterative, static assumptions lead to operational inefficiencies and increase reliance on manual intervention (Wirth & Hipp, 2000).

Limitations of Traditional Automation

Early automation approaches to data preparation relied on deterministic extraction–transformation–loading (ETL) pipelines and handcrafted scripts (Cormier et al., 2025). These mechanisms are effective when data characteristics are predictable, and transformation rules are explicitly defined. However, they offer limited adaptability when confronted with irregular schemas, evolving data semantics, or domain-specific anomalies (Paton, 2019; Konstantinou & Paton, 2019).
A key structural limitation of deterministic automation is its inability to infer contextual meaning or adjust transformation logic in response to unforeseen data patterns. Systems that lack adaptive capabilities must be repeatedly re-engineered, resulting in cumulative operational overhead and increased risk of error propagation. As organisations deploy data-driven processes in regulated domains, such as healthcare, finance, and scientific instrumentation, the absence of contextual responsiveness introduces risks concerning trustworthiness, reproducibility, and compliance (Somasundaram, 2022; Fernandes et al., 2023). These limitations collectively indicate that static automation is insufficient for emerging data environments.

Human-in-the-Loop (HITL) Paradigm

The Human-in-the-Loop (HITL) paradigm provides a framework for incorporating human expertise into computational processes. Unlike automated workflows that execute predefined rules, HITL architectures enable users to validate intermediate results, correct inconsistencies, and contribute domain knowledge essential for resolving ambiguous or semantically rich transformations (Konstantinou & Paton, 2020; Somasundaram, 2022).
Feedback constitutes the principal operational construct in HITL systems. It may be explicit (where users provide direct corrections, confirmations, or annotations), or implicit (where systems infer intent from interaction traces, behavioural patterns, or decision histories).
These feedback signals extend beyond simple error reporting. They function as supervisory information that guides automation processes, modifies execution paths, and calibrates system behaviour. As a result, HITL enables automation mechanisms to incorporate semantic reasoning, thereby reducing reliance on rigid configuration.

Feedback-Driven Automation

Feedback-driven automation represents an evolutionary step beyond HITL by enabling systems not only to accept feedback but also to integrate it into adaptive decision processes. In this paradigm, feedback becomes a computational artefact that informs subsequent operations, enabling systems to revise transformation rules, prioritise error resolution, and propagate corrections across multiple datasets or execution scenarios (Konstantinou & Paton, 2020).
Such systems combine algorithmic inference, provenance modelling, and incremental learning to achieve the assimilation of correction patterns, the contextual propagation of validated transformations to future datasets, the reduction of repetitive interventions, and the progressive refinement of pipeline behaviour (Asai & Yamauchi, 2023).
This operational shift transforms data preparation from a linear execution sequence into an iterative learning loop. Whereas traditional pipelines execute static logic, feedback-driven systems adjust transformation strategies based on accumulated evidence, aligning with contemporary needs for resilient, domain-sensitive automation (Vanalkar & Pazare, 2025).

Prior Surveys and the Need for a Hybrid Theoretical Lens

Existing surveys on data preparation and data-quality management have primarily focused on cleaning strategies, tooling, or machine-learning-assisted repair (Lai et al., 2021; Côté et al., 2023). While informative, these studies adopt a narrow scope and do not address the integration of feedback mechanisms across the entire preparation pipeline (Konstantinou & Paton, 2020). Critically, they do not examine whether systems persist or generalise learned corrections, nor do they analyse architectural dependencies that enable the reuse of feedback knowledge (Liu & Yu, 2022).
To address these limitations, this study adopts a hybrid theoretical lens that combines CRISP-DM (to model the structural progression of the preparation pipeline) and HITL (to characterise the mechanisms by which feedback informs and modifies automation). This dual perspective enables a systematic comparison of existing systems based on their capacity to capture, interpret, and operationalise feedback. It also reveals structural assumptions embedded within current research, particularly the lack of architectures that unify data preparation, adaptive learning, and feedback reuse (Konstantinou & Paton, 2020).
Feedback-driven automation reframes data preparation as an iterative computational process rather than a static engineering task. However, the literature lacks a comprehensive synthesis of how feedback is harnessed, represented, and operationalised across the stages of preparation. This gap provides the rationale for the systematic review conducted in this study.

Science World Journal Vol. 21(No 1) 2026 https://dx.doi.org/10.4314/swj.v21i1.39
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

MATERIALS AND METHODS

The review methodology was structured to address the research questions presented in the Introduction section. This section outlines the procedures for identifying, selecting, and analysing studies on feedback-driven automation in data preparation. The review follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) guidelines to ensure transparency, replicability, and methodological consistency. The methodological design was established prior to executing the search queries and remained unchanged throughout the review.

Review Protocol

A predefined review protocol was adopted to structure the investigation. The protocol specified the research questions, search strategy, eligibility criteria, quality assessment measures, and synthesis procedures. The procedures adopted in this study follow the reporting structure and transparency requirements set out in PRISMA 2020 reporting principles. All procedural decisions, search logs, inclusion rationales, and screening outcomes were documented to enable independent verification.

PICOC Framework

To articulate the scope and boundaries of this systematic review, the Population–Intervention–Comparison–Outcome–Context (PICOC) framework was adapted. It ensures consistent alignment between the research objectives and the evidence base, as shown in Table 1.

Table 1: PICOC Framework for the Systematic Review

Element	Specification for this Review
Population	Data-preparation systems, workflows, or pipelines
Intervention	Feedback-driven mechanisms (explicit, implicit, hybrid, or adaptive)
Comparison	Traditional deterministic or non-feedback automation approaches
Outcome	Improvements in automation efficiency, data quality, rule refinement, or usability
Context	Computational, engineering, and machine-learning environments

Eligibility Criteria

Eligibility criteria were defined to ensure methodological rigour and to restrict the review to studies with demonstrable feedback-driven contributions to data-preparation processes.

Inclusion Criteria

Studies were included in this review if they were published between January 2010 and December 2025, written in English, and appeared as peer-reviewed journal articles, conference papers, or book chapters. Eligible studies were required to present a system, framework, tool, or methodology that automated or semi-automated one or more data-preparation tasks such as cleaning, transformation, schema mapping, or integration. Additionally, they had to incorporate a feedback mechanism, either explicit, implicit, or adaptive, ranging from user corrections and interaction logs to iterative repair processes or learning-based feedback models. Finally, only studies that provided sufficient technical detail, including system architecture, algorithmic specifications, evaluation metrics, or user studies, to allow classification within our proposed taxonomy were considered.

In addition to formally peer-reviewed publications, a limited number of high-impact pre-publication studies hosted on arXiv were included. Although arXiv does not conduct formal peer review, it is widely used in computer science as a pre-publication dissemination platform, particularly for emerging research on data systems, human-in-the-loop automation, and large language model–assisted workflows. Such studies were included only when they demonstrated strong methodological clarity, relevance to feedback-driven data preparation, and evidence of scholarly uptake, such as citations or subsequent peer-reviewed publication. These studies were treated as peer-review-adjacent and were not weighted equivalently to formally peer-reviewed articles during synthesis.

Exclusion Criteria

Studies were excluded if they were published outside the 2010–2025 timeframe, not written in English, or presented as reports, posters, editorials, white papers, vendor brochures, blogs, or any other non-archival content. Studies that focused exclusively on data cleaning for model performance without addressing broader data-preparation workflows, or those that examined feedback solely within machine-learning model training rather than data preparation, were also omitted. Additionally, studies without publicly accessible full texts or those that did not provide architectural, empirical, or algorithmic contributions were excluded. These criteria ensured that only studies with demonstrable, feedback-driven contributions to data preparation pipelines were retained for review.

Information Sources

Five major scientific electronic databases, IEEE Xplore, ACM Digital Library, SpringerLink, Semantic Scholar, and Google Scholar, were queried for relevant studies. Although Web of Science was initially considered, it was excluded due to institutional access limitations.

Additionally, arXiv was queried selectively as a supplementary source to capture emerging and pre-publication research not yet published in archival venues. Preprint studies retrieved from the arXiv were screened using the same eligibility and quality assessment criteria as those applied to peer-reviewed literature, with additional scrutiny of methodological transparency, evaluation rigour, and relevance to feedback-driven automation in data preparation.

Search Strategy

A combination of Boolean expressions and controlled keywords was used to retrieve relevant literature. The primary query pattern was:

("data preparation" OR "data wrangling" OR "data cleaning" OR "schema matching" OR "data transformation")
AND
("feedback" OR "user feedback" OR "human-in-the-loop" OR "interactive" OR "adaptive" OR "implicit feedback")
AND
("automation" OR "automated" OR "semi-automated" OR

Science World Journal Vol. 21(No 1) 2026 https://dx.doi.org/10.4314/swj.v21i1.39
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

"intelligent system" OR "machine learning")

Queries were executed between March and September 2025. The temporal boundary (2010–2025) reflects the emergence of machine-learning-assisted preparation pipelines and HITL frameworks. Earlier works predominantly employed static ETL processes without interactive or adaptive feedback capabilities; therefore, inclusion of pre-2010 research was unnecessary for this synthesis.

Study Selection Procedure

A multi-stage screening procedure was adopted to ensure a rigorous and unbiased selection of studies, which is consistent with PRISMA 2020 recommendations and the inclusion/exclusion criteria in the Information Sources Section:

1. Identification: A total of 23,866 records were retrieved from IEEE Xplore, ACM Digital Library, SpringerLink, Semantic Scholar, and Google Scholar using the search strings described in the Search Strategy Section. The high initial record count reflects the inclusion of broad Google Scholar and Semantic Scholar queries, which return approximate hit counts rather than exact bibliographic records. Only records with accessible metadata were subsequently retained. Records retrieved from arXiv were treated as pre-publication sources and were tracked separately during screening to ensure transparent differentiation from formally peer-reviewed publications.
2. **Initial Screening (Inclusion/Exclusion Criteria):** After applying publication-year limits (2010–2025), English-language filters, peer-review status, and document-type constraints (journal articles, conference papers, book chapters), 6,614 studies remained for further screening.
3. **Duplicate Removal**: Cross-database duplicates (identical title and DOI) were removed, yielding 210 **unique studies**.
4. **Title Screening**: The titles of 210 records were screened for relevance to automated or semi-automated data-preparation pipelines with feedback or human-in-the-loop mechanisms. Clearly irrelevant works, such as those in unrelated domains or without a data preparation component, were excluded. 120 were excluded, leaving 90 studies for abstract-level assessment.
5. **Abstract and Conclusion Screening:** After evaluating the abstracts and conclusions of the remaining 90 papers against the PICOC framework and established eligibility criteria, studies that did not incorporate automation, lacked any form of feedback mechanism, or focused solely on downstream model training without addressing data preparation were excluded. Following this screening process, 60 studies met the requirements, while 30 were removed at this stage.
6. Full-Text Eligibility Assessment: Full texts of the 60 remaining studies were assessed, and 20 were excluded for not describing a concrete system, mentioning data preparation only superficially, or using feedback for general organisational decisions rather than data-preparation tasks. Excluded works included conceptual AI strategy papers, blockchain supply chain studies, hyperautomation reports, and threat modelling frameworks without data preparation components.
7. Quality Assessment: The remaining 40 studies underwent quality scoring using the rubric presented in the Data Extraction and Coding Section, which comprised six criteria with a maximum obtainable score of 6. Any study that scored below 3 out of 6 (50%) was deemed methodologically weak and therefore removed from the primary evidence base. Of the 40 studies assessed, 11 were excluded for failing to meet the minimum quality threshold, leaving the remaining 29 to form the core dataset for subsequent synthesis and analysis.
8. Final Inclusion: After the quality assessment stage, a total of 29 studies were retained as primary sources for data extraction and synthesis. These selected works met all inclusion criteria by simultaneously addressing one or more data preparation tasks, incorporating explicit, implicit, or adaptive feedback mechanisms, and providing sufficient technical, architectural, or empirical detail to meaningfully support the development of our taxonomy and the investigation of the study's research questions.

The remaining high-quality but peripheral papers, such as general HITL surveys or feedback frameworks without concrete data-preparation implementations, were used only as **contextual or theoretical references** and are not counted as primary SLR studies.

Table 2 presents the search and screening summary for studies identified from each database, and Figure 1 illustrates the PRISMA flow diagram of the study selection process.

Table 2: Search and Screening Summary of Identified Studies

Stage	IEEE Xplore	ACM DL	SpringerLink	Semantic Scholar	Google Scholar	Total
Initial Search	246	15,415	1,455	1,240	5,510	23,866
First Screening (year, language, publication type)	133	6,095	28	111	247	6,614
Second Screening (relevance to data preparation + feedback)	83	99	25	38	15	260
Duplicates Removed	22	12	4	8	4	50
Unique Records	61	87	21	30	11	210
Full-Text Reviewed	61	79	21	30	11	202
Excluded from full text	43	74	18	28	10	173
Quality Assessment	18	5	3	2	1	29
Final Included Studies	18	5	3	2	1	29

Science World Journal Vol. 21(No 1) 2026
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University
https://dx.doi.org/10.4314/swj.v21i1.39

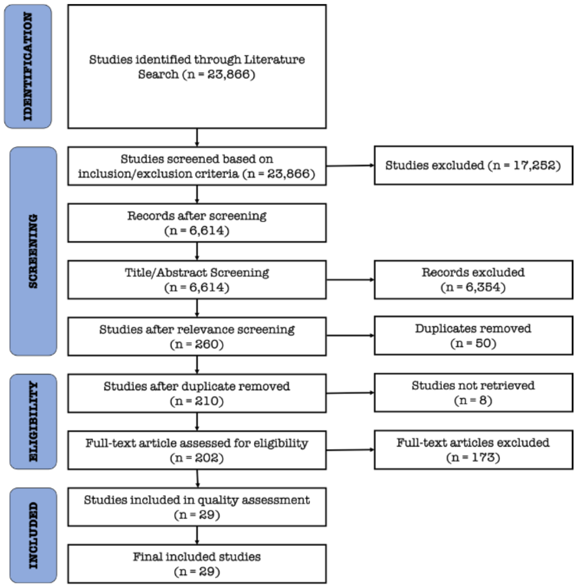

Figure 1. PRISMA Flow Diagram of the Study Selection Process

Quality Assessment

The methodological rigour of each study in feedback-driven data-preparation research was evaluated using a six-point rubric. The criteria and scoring scheme employed in the assessment are presented in Table 3.

Table 3. Quality Assessment Rubric for Included Studies

Criterion	Scoring
Clear definition of data-preparation tasks	0 / 0.5 / 1
Presence of a feedback mechanism	0 / 0.5 / 1
Description of automation strategy	0 / 0.5 / 1
Evidence of evaluation (accuracy, efficiency, usability)	0 / 0.5 / 1
Reproducibility (architecture, algorithms, examples)	0 / 0.5 / 1
Feedback reuse or learning capability	0 / 0.5 / 1

Studies scoring below 3 out of 6 were excluded to avoid conceptual or anecdotal evidence from influencing the synthesis. A complete breakdown of the scoring outcomes is provided in Appendix A (Table A1). However, two borderline studies scoring slightly below the threshold were excluded from quantitative synthesis but retained due to their conceptual completeness to feedback-driven architectures and their frequent citation in the literature. To mitigate potential bias introduced by pre-publication studies, all arXiv papers were subjected to the same six-point quality assessment rubric as peer-reviewed studies, with stricter interpretation applied to criteria for evaluating evidence and reproducibility. Studies whose primary contribution relied solely on speculative claims or lacked empirical validation were excluded. Findings derived primarily from pre-publication studies were interpreted with caution and triangulated with peer-reviewed evidence wherever possible, ensuring transparency and reproducibility of the selection process.

Data Extraction and Coding

The fields used for data extraction were structured to standardise interpretation across studies, and these are summarised in Table 4. This coding allows grouping and classification, which later forms the taxonomy and comparative analysis. The complete extracted dataset corresponding to these fields is provided in Appendix B (Table B1).

Table 4. Data Extraction and Coding Schema for Included Studies

Field	Description
Authors & Year	Bibliographic information
Venue & Publication Type	Journal, conference, workshop, etc.
Feedback Type	Explicit feedback (user corrections), Implicit feedback (usage logs, error reports), Mixed
Data-Preparation Stage(s) Addressed	Cleaning, transformation, schema matching, integration, validation, etc.
Automation Technique	Rule-based, ML-based, LLM-assisted, heuristic, hybrid
Feedback Integration Architecture	How and where feedback is injected — interactive GUI, pipeline hook, iterative loop, feedback repository, etc.
Evaluation Method	Qualitative (user study), quantitative (error reduction, data quality metrics, time savings), performance benchmarks
Datasets Used	Synthetic, real-world (open or private), domain context
Strengths & Limitations (as reported by authors)	What authors claim, and what issues remain
Additional Comments	Notable design decisions, special features, reuse capabilities, and open-source availability

Synthesis Method

Due to heterogeneity in evaluation metrics, architectures, and reporting formats, a narrative synthesis was adopted rather than meta-analysis. Comparative tabulation and taxonomic mapping were used to aggregate patterns and identify trends in architectural integration.

Threats to Validity

Potential validity concerns identified in this study include publication bias arising from limited disclosure of industrial deployments; language bias, as only English-language sources were considered; and classification subjectivity, which was mitigated through predefined inclusion criteria and repeated screening. Additionally, metric inconsistency posed a challenge and was managed through qualitative aggregation rather than numerical generalisation. Nonetheless, adherence to PRISMA reporting practices reduces interpretive bias and enhances the reproducibility of the review process. An additional threat to validity arises from the inclusion of a limited number of pre-publication studies sourced from arXiv. While common practice in computer science, such studies may undergo substantive revision during formal peer review. To address this risk, arXiv studies were treated as indicative of emerging research directions rather than definitive evidence, and their findings were not weighted equally with those

Science World Journal Vol. 21(No 1) 2026 https://dx.doi.org/10.4314/swj.v21i1.39
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

of peer-reviewed publications.

RESULTS

This section presents the outcomes of the synthesis conducted on the 29 primary studies retained after the screening and quality assessment procedures described in the Methodology Section. The findings are organised along three analytical dimensions corresponding to the research questions: (i) the feedback mechanisms adopted, (ii) the pipeline stages where feedback is operationalised, and (iii) the architectural strategies that support automation within these systems.

Overview of Included Studies

The 29 studies meeting the inclusion criteria were published between 2019 and 2025, reflecting a recent and accelerating research trajectory in feedback-driven automation for data preparation. Although the search window spanned 2010–2025, all studies meeting the inclusion criteria were published from 2019 onward, reflecting the recent emergence of feedback-driven automation paradigms. Three thematic clusters were identified. The distribution of the 29 included studies across these thematic clusters is presented in Table 5, which summarises the dominant research orientations underpinning feedback-driven automation in data preparation.

Table 5. Thematic Distribution of Included Studies

Cluster	Description	% of Studies
Data-preparation systems	Systems supporting interactive wrangling, schema correction, transformation, and repair	55%
Feedback analytics frameworks	Models for preference extraction, behavioural inference, and feedback introspection	31%
Pipeline automation foundations	Approaches enabling rule evolution, provenance-aware reuse, and runtime adaptation	14%

This distribution indicates a dominant emphasis on operational data-preparation tasks, with supporting research increasingly directed toward analytics mechanisms that enable the interpretation and reuse of feedback.

As illustrated in **Figure 2**, research activity in feedback-driven data preparation has risen markedly since 2021, coinciding with the growing adoption of HITL paradigms, large language models (LLMs), and provenance-centric architectures. This upward trend reflects a broader shift toward adaptive, user-informed automation in contemporary data-preparation workflows.

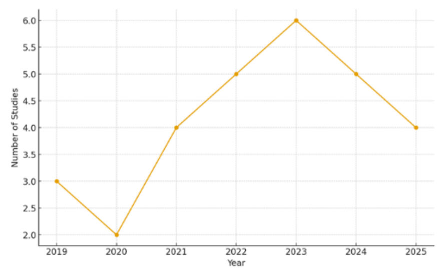

Figure 2. Publication Trend of Feedback-Driven Data Preparation Studies (2019-2025)

Taxonomy of Feedback Mechanisms

The analysed studies adopt four classes of feedback modalities, distinguished by capture strategy, operational scope, and degree of learning capability. Table 6 summarises their deployment characteristics.

Table 6. Hybrid Taxonomy of Feedback Modalities, Integration Stages, and Operational Strategies

Feedback Type	Primary Pipeline Stage(s)	Operational Strategy / Mechanisms	Frequency Across Studies
Explicit Feedback	Schema Matching, Data Cleaning, Quality Validation	User confirmations, rule approvals, correction dialogs, annotation-based repair, and interactive transformation refinement	Very High - Dominant in classical HITL systems; present in most early and mid-period studies (2011–2020)
Implicit Feedback	Transformation, Profiling, Error Detection	Behavioural logs, correction patterns, change-propagation heuristics, pattern reuse, user-action mining	Moderate - Increasing adoption after 2018; leveraged by ML-enabled platforms
Hybrid Feedback	Multi-stage pipelines combining cleaning, schema mapping, and enrichment	Reinforcement loops, provenance tracking, contextual learning, LLM-assisted program synthesis, preference-generalization engines	High - Rapid growth post-2021 due to LLMs and adaptive orchestration frameworks
Autonomous / Learned Feedback*(emergent)*	End-to-end automated pipelines and reuse across datasets	Self-tuning orchestration, cross-dataset rule generalization, persistent feedback repositories, continuous model adaptation	Low - Experimental; limited to a small subset of recent studies (2023–2025)

Science World Journal Vol. 21(No 1) 2026
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University
https://dx.doi.org/10.4314/swj.v21i1.39

Explicit feedback remains predominant, particularly in early studies, whereas hybrid and implicit approaches appear increasingly in recent systems due to advances in behavioural analytics and model-driven inference.
These findings provide a consolidated synthesis of the feedback mechanisms employed across contemporary data preparation systems, thereby addressing RQ1.

Pipeline Stages of Feedback Integration
Feedback mechanisms were identified across five recurring stages of the data-preparation pipeline. The distribution of feedback utilisation across these stages is summarised in **Table 7**, which shows a marked concentration of interventions in schema matching and data-cleaning tasks. This pattern reflects the high semantic and contextual ambiguity associated with these phases, where automated techniques still rely heavily on corrective signals.

Table 7. Distribution of Feedback Integration across Data Preparation Stages

Pipeline Stage	Feedback Role	% of Studies
Data Profiling	Identify anomalies and missing values	10%
Schema Matching / Alignment	Validate mapping correctness	32%
Data Cleaning / Repair	Correct values, deduplicate inconsistencies	40%
Transformation and Enrichment	Refine transformation rules and propagate corrections	30%
Quality Validation	Confirm correctness and enable reuse	18%

To complement this tabular summary, Figure 3 visualises the proportion of studies operating at each stage, further reinforcing the dominance of feedback-driven interventions within cleaning and schema-alignment processes, and indicating comparatively limited utilisation during profiling and validation.
Taken together, the distribution of feedback utilisation across the data-preparation pipeline demonstrates where measurable benefits occur, thereby satisfying RQ2.

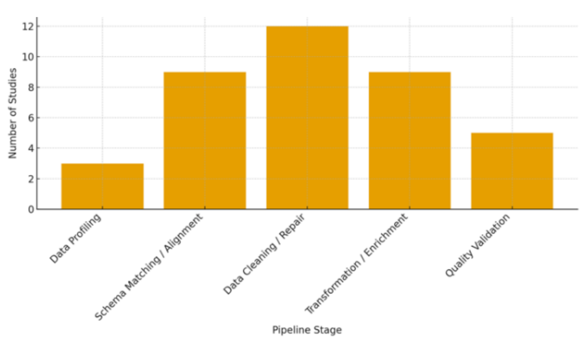

Figure 3. Distribution of Feedback across Data Preparation Pipeline Stages

Architectural Strategies for Feedback Operationalisation
To compare how existing systems embed and utilise feedback, four recurring architectural classes, differing in their integration scope and automation capabilities, were identified, as summarised in

Table 8. Architectural Classes and Strategies for Operationalising Feedback

Architecture Class	Core Strategy	Typical Techniques	Characteristics
Rule-based systems	Apply validated transformation rules	Deterministic rules, heuristics	brittle without feedback reuse
Interactive workflows	Humans intervene iteratively	GUIs, dashboards	rely heavily on explicit user effort
ML-driven adaptive systems	Learn from feedback patterns	clustering, classifiers	Feedback influences future decisions
Feedback repositories	Persist, organize, reuse captured feedback	provenance logs, knowledge stores	foundation for unified platforms

A notable pattern is the absence of systems that implement repository-based reuse across the full pipeline, thereby confirming the architectural discontinuity that motivates this study.

Summary of Primary Studies
Table 9 summarises the 29 primary studies, categorised by feedback type, pipeline stage, and architectural approach. The table provides a structured reference for comparative analysis.

Table 9. Summary of Primary Studies on Feedback-Driven Automation

Author(s) & Year	Feedback Type	Pipeline Stage(s)	Architectural Approach	Core Contribution
Heer et al. (2019)	Explicit	Cleaning, Transformation	Interactive workflows	Visual wrangling with guided correction
Bogatu et al. (2019a)	Explicit	Transformation	Rule-based	Edit-operation-based transformation synthesis
Bogatu et al. (2019b)	Explicit	Transformation	Rule-based	Automatic example generation for wrangling
Paton (2019)	Conceptual	All	Conceptual	Automation feasibility framework
Konstantinou et al. (2019)	Hybrid	Cleaning, Matching	Feedback repository	VADA architecture for user-driven pipelines
Konstantinou et al. (2020)	Hybrid	Cleaning, Matching, Repair	ML-driven	Statistical prioritisation of user feedback
Rezig et al. (2019)	Explicit	Cleaning	Interactive	Human-centric cleaning design
Azeroual (2020)	Explicit	Cleaning	Interactive	Database cleaning

https://dx.doi.org/10.4314/swj.v21i1.39
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

Author(s) & Year	Feedback Type	Pipeline Stage(s)	Architectural Approach	Core Contribution
				explanation
Somasundaram (2022)	Explicit	Cleaning	Rule-based	Automated wrangling guidelines
Rajenthiram (2025)	Hybrid	Cleaning, Validation	Adaptive workflows	User-driven pipeline optimisation
Bilal et al. (2022)	Hybrid	Cleaning	ML-driven	Auto-PreP automated preprocessing
Fan et al. (2024)	Hybrid	Transformation	ML-driven	Multi-agent AutoPrep pipeline
Goyle et al. (2024)	Implicit	Cleaning	ML-driven	DataAssist automation
Liu et al. (2024)	Implicit	Cleaning	ML-driven	AutoDW for full wrangling
Liu et al. (2025)	Implicit	Cleaning, Transformation	ML-driven	AutoDW-TS for time-series
Krajnc et al. (2022)	Explicit	Cleaning	ML-driven	Optimised preprocessing for tumour pipelines
Yasin &Khorsheed (2025)	Explicit	Cleaning	ML-driven	Automated learning-based cleaning
Narayan et al. (2022)	Implicit	Cleaning, Integration	ML-driven	Foundation models for wrangling
Li &Döhmen (2024)	Implicit	Transformation	ML-driven	Efficient LLM code generation
Huang et al. (2024)	Implicit	Transformation	ML-driven	Contextual code wrangling (CoCoMine)
Akella et al. (2025)	Implicit	Cleaning, Transformation	ML-driven	CodeGenWrangler task automation
Chang et al. (2025)	Hybrid	Pipeline construction	RL-LLM	Adaptive pipeline building
Chen et al. (2024)	Hybrid	Schema & Transformation	ML-driven	Data-Juicer processing
Somayajula (2025)	Explicit	Integration	Rule-based	Enterprise ETL optimisation
Seedat et al. (2022)	Implicit	Profiling, Validation	ML-driven	Data congruence detection
Wu et al. (2022)	Conceptual	All	Survey	HITL synthesis
Shome et al. (2024)	Implicit	Validation	Behaviour mining	Notebook feedback taxonomy
Jaimovitch-López et al. (2022)	Implicit	Transformation	LLM-assisted	LLMs for wrangling tasks
Xu et al. (2023)	Explicit	Validation	ML-driven	Obsolete comment detection impact

Cross-Study Comparative Summary

Comparative synthesis indicates that **62%** of the studies report measurable improvements in data quality, and **48%** demonstrate reductions in manual intervention or preparation time. However, only **17%** implement persistent or reusable feedback, and **none** provide a unified architecture spanning the entire preparation pipeline.

This fragmentation suggests that although feedback has demonstrable utility for local optimisation, current systems do not support generalisation across datasets, transformation cycles, or operational environments.

Emerging Directions

A longitudinal analysis of the selected studies indicates a progressive shift in the design and operational assumptions of data-preparation systems. Early systems predominantly relied on deterministic scripts and explicit user corrections; however, recent systems increasingly incorporate adaptive computational components that interpret user signals, infer transformation intent, and adjust preparation logic at runtime. Three trends are particularly evident:

1. **Transition from Static Pipelines to Adaptive Learning Loops:** Recent systems embed machine-learning components that utilise feedback not merely for error correction, but as a driver for iterative refinement of preparation rules and transformation decisions. This marks a departure from linear ETL workflows toward cyclical preparation processes that evolve and adapt to changing data semantics.
2. **Integration of Large Language Models and Semantic Inference:** Several studies leverage LLMs to infer transformation patterns, generate wrangling code, and interpret user behaviour as implicit feedback. This development expands the feedback modality beyond direct annotations to include behavioural and contextual signals, enabling systems to generalise user intent without explicit intervention. Beyond their operational use, large language models introduce a fundamental shift in how feedback is captured and utilised within data-preparation pipelines. Unlike traditional human-in-the-loop systems that rely primarily on explicit user corrections, LLM-based approaches enable the interpretation of implicit feedback through interaction context, natural language prompts, and behavioural patterns. These models support automated code generation for data transformation, infer transformation intent from user queries, and abstract feedback into reusable transformation logic

Science World Journal Vol. 21(No 1) 2026 https://dx.doi.org/10.4314/swj.v21i1.39
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

without requiring direct annotation. This capability reduces the dependency on continuous user intervention and allows systems to generalise feedback across tasks and datasets. However, the integration of LLMs also introduces notable limitations, including the risk of hallucinated transformations, lack of deterministic reproducibility, and challenges in validating generated outputs, particularly in high-stakes or regulated environments.

3. **Emergence of Provenance-Aware and Compliance-Sensitive Architectures:** Increasing regulatory emphasis on transparency and reproducibility has motivated systems that record not only transformation outcomes, but the decision rationale behind them. Feedback is thereby treated as a traceable artefact that supports auditability, accountability, and system introspection.

Despite these advances, several unresolved issues persist in current research efforts, particularly regarding how feedback is represented, propagated, and operationalised across different components of the preparation pipeline.

DISCUSSION AND IMPLICATIONS

This section interprets the synthesised results and examines the implications of feedback-driven automation for data-preparation research and practice. The discussion is structured around the conceptual, methodological, and architectural themes emerging from the reviewed studies.

Interpretation of Findings

The empirical evidence confirms that feedback mechanisms play a decisive role in improving the reliability, correctness, and adaptability of automated data-preparation workflows. Most existing systems apply feedback to correct values, refine transformation logic, and validate schema mappings; however, these interventions remain primarily localised to specific pipeline stages. This selective integration limits the pipeline's overall automation potential and results in repeated corrective effort, since insights generated during cleaning or alignment are not preserved for subsequent tasks.

The predominance of explicit feedback techniques indicates that automation continues to rely on direct human intervention to resolve ambiguity. While implicit and hybrid mechanisms demonstrate growing capability, particularly in studies employing machine-learning models, their adoption remains limited by architectural complexity and a lack of standardised representations for storing and reusing feedback. Consequently, current systems improve the efficiency of task execution but fall short of lifecycle-level automation.

Comparison with Prior Reviews

Existing systematic reviews on data preparation and data quality have primarily focused on specific sub-tasks, such as data cleaning techniques or the application of machine learning for preprocessing. For instance, Lai et al. (2021) emphasise data-driven optimisation and bottleneck detection in industrial systems, while Côté et al. (2023) provide a comprehensive synthesis of data-cleaning methods to improve machine learning performance. Although these studies offer valuable insights, their scope is largely confined to isolated stages of the data-preparation process. It does not explicitly consider feedback as a unifying mechanism across the pipeline.

In contrast, this study adopts a broader and integrative perspective by systematically examining how feedback mechanisms are designed, operationalised, and distributed across multiple stages of the data-preparation lifecycle. It further advances the literature by proposing a hybrid taxonomy that links feedback modalities to pipeline stages and architectural strategies, and by highlighting the absence of persistent feedback repositories that enable reuse across datasets and sessions. Consequently, this review extends prior work by positioning feedback not merely as an auxiliary feature but as a central architectural component that enables adaptive, scalable, and context-aware data-preparation systems.

Theoretical Implications

The findings of this review reveal three substantive theoretical implications for the evolving discourse on data preparation. First, the evidence supports reconceptualising data preparation from a linear, deterministic Extract–Transform–Load (ETL) activity to an adaptive, iterative process. Rather than operating as fixed execution pipelines, contemporary systems demonstrate that user interaction, contextual cues, and evolving data semantics introduce variability that demands dynamic recalibration. Feedback mechanisms, therefore, function not as peripheral refinements but as integral components that guide continuous adjustment and reconfiguration.

Second, the review advances an epistemic interpretation of feedback. While conventional perspectives regard feedback as merely operational metadata or corrective annotation, emerging systems show that it frequently embodies domain knowledge, transformation rationale, and user intent. In this sense, feedback serves as a knowledge-bearing artefact that shapes algorithmic decisions, influences future transformations, and enhances the traceability and reproducibility of data workflows. This elevation of feedback from procedural metadata to an epistemic construct challenges existing taxonomies and invites deeper theoretical engagement with its representational capacities.

Third, the synthesis points to the emergence of hybrid human–machine architectures that reposition feedback as a supervisory signal rather than a terminal corrective action. These architectures delegate routine, well-defined operations to automated inference engines while reserving human judgment for context-sensitive, ambiguous, or value-laden decisions. Consequently, feedback becomes a mechanism that steers automation trajectories, enabling systems to evolve with usage patterns and organisational contexts rather than remaining static after initial deployment. This reframing highlights feedback as a structural driver of adaptation within data preparation ecosystems, rather than merely an instrument of error correction.

Practical Implications

The synthesis of reviewed studies highlights five interrelated implications for the design and deployment of feedback-driven data-preparation platforms. First, feedback must be treated as a persistent asset rather than a transient interaction point. Architectures that retain, index, and generalise feedback beyond its initial application enable cumulative automation, minimise repetitive input, and progressively reduce user effort. Second, user-interaction mechanisms should prioritise capturing intent over discrete corrective actions. By inferring transformation semantics from behavioural cues, systems can scale user-derived intelligence without requiring continuous explicit confirmations. Third, establishing structured feedback repositories is critical. In their absence, feedback remains localised to individual sessions or

 https://dx.doi.org/10.4314/swj.v21i1.39

datasets, limiting its reuse and preventing pipeline-level optimisation; persistent repositories, in contrast, facilitate knowledge transfer across transformations, workflows, and organisational contexts. Fourth, increasing automation levels require transparency. Feedback-informed decisions, particularly in regulated domains such as healthcare and finance, require explainability, provenance tracking, and auditability to demonstrate how specific user inputs shaped data transformations. Finally, regulatory expectations surrounding data governance and traceability recast feedback as a documented decision point rather than an auxiliary mechanism. Consequently, provenance-aware system architectures are not merely beneficial but essential for compliant and operational deployment in contemporary data ecosystems.

Research Gaps Identified

A synthesis of the reviewed studies reveals four persistent deficiencies that constrain the evolution of feedback-driven data preparation systems. These gaps are summarised in Table 10.

Table 10. Gaps in Feedback-Driven Data Preparation Research

Gap	Description
Lack of Unified Architectures	Existing systems address isolated stages of the data preparation workflow; none provide an end-to-end architecture that persistently captures, reuses, and propagates feedback across sessions or datasets.
Localised Feedback Generalisation	User corrections are typically bound to a single task or dataset, resulting in limited cross-context inference and minimal transferability of learned rules or patterns.
Absence of Evaluation Benchmarks	Heterogeneity in datasets, metrics, and reporting practices prevents rigorous cross-system comparison, making it difficult to assess performance improvements attributed to feedback mechanisms.
Limited Regulatory Consideration	Feedback artefacts are not conceptualised as governed data objects, despite increasing regulatory obligations surrounding data provenance, auditability, and user consent.

Collectively, these gaps highlight a structural limitation in current research. Although feedback has been shown to enhance the correctness and efficiency of data preparation processes, it is rarely operationalised as a persistent and reusable knowledge resource. As a result, no existing approach provides a unified, feedback-driven platform that can assimilate, store, and reapply user feedback across sessions, users, and datasets. Addressing these limitations, therefore, represents a necessary direction for future research in adaptive and scalable data-preparation systems. These findings directly address RQ3 by identifying the architectural constraints that limit current feedback-driven data preparation systems and by clarifying the requirements necessary for developing unified, scalable, and feedback-aware platforms.

Implications for Future Systems

The convergence of feedback analytics, provenance-aware computing, and automated orchestration indicates an emergence of the next generation of data preparation platforms with capabilities that extend beyond conventional human-in-the-loop paradigms. These next-generation systems are envisioned to incorporate (i) persistent feedback repositories that enable cross-dataset rule generalisation; (ii) semantic inference mechanisms capable of interpreting and operationalising user actions; (iii) adaptive transformation engines that evolve in response to accumulated feedback; (iv) provenance-aware components that maintain end-to-end traceability and support regulatory compliance; and (v) domain-independent architectural frameworks that facilitate applicability across heterogeneous contexts rather than isolated, task-specific deployments.

The absence of such integrated platforms in current literature highlights a significant research opportunity and provides the foundational motivation for the present study. Fully realising feedback as a computational resource requires architectural principles that treat user input not as a transient annotation but as a reusable knowledge artefact.

Conclusion

This systematic literature review examined the role of feedback mechanisms in automating data-preparation pipelines. It analysed how contemporary systems operationalise user interaction to improve data quality, transformation accuracy, and workflow efficiency. The review synthesised evidence from 29 primary studies published between 2019 and 2025 and identified feedback as a pivotal element in transitioning from deterministic, rule-based pipelines to adaptive, context-aware data-preparation architectures.

The findings demonstrate that while feedback significantly enhances individual tasks, most notably cleaning, schema alignment, and transformation, the implementations remain fragmented and task-specific. Feedback is predominantly consumed at runtime, without being stored or generalised for future use, thereby constraining advances in automation and preventing cross-dataset knowledge transfer. Moreover, emerging approaches incorporating machine learning, behavioural analytics, and large language models indicate a trend toward implicit and hybrid feedback mechanisms. However, these systems lack unified repositories or governance structures to capture feedback as a durable knowledge asset.

The review further identified four unresolved limitations in current research. These are (i) the absence of end-to-end architectures capable of integrating feedback across all preparation stages, (ii) the lack of mechanisms for generalising feedback beyond local corrections, (iii) insufficient benchmark datasets for comparable evaluation, and (iv) minimal attention to provenance, auditability, and compliance concerns. These gaps underscore the need for architectural frameworks that treat feedback not as an auxiliary annotation but as a core computational paradigm.

Future Research Directions

The findings of this review highlight several unresolved challenges and, consequently, three promising avenues for advancing feedback-driven automation in data preparation:

a. Unified, Feedback-Oriented Architectures: Future platforms should employ persistent, context-aware feedback repositories that capture, index, and reuse corrective actions across datasets, domains, and operational environments. Such architectures would transform feedback from isolated, task-specific interventions into reusable knowledge for transformation, thereby enhancing consistency and reducing repetitive manual effort.

b. Semantic and Behavioural Feedback Generalisation: There is a need for models that abstract patterns from explicit user interactions and implicit behavioural cues, enabling systems to autonomously propagate learned corrections, infer transformation intents, and disambiguate mappings without recurrent user supervision. Achieving this generalisation is essential for scalable automation and a core prerequisite for reducing user dependency in complex workflows.

c. Compliance-Aware and Explainable Data-Preparation Pipelines: As regulatory, ethical, and organisational constraints intensify, data-preparation systems must incorporate mechanisms for transparent lineage tracking, justification logging, and verifiable transformation paths. Embedding feedback within provenance models will ensure that automated decisions remain interpretable, auditable, and aligned with domain and legal requirements.

In summary, the evidence reviewed in this study establishes feedback as a central enabler of adaptive automation in data-preparation workflows. While existing systems demonstrate meaningful progress, the absence of unified, feedback-aware architectures limits the scalability and reusability of current solutions. Addressing this gap provides a clear direction for subsequent research and represents an essential step toward realising fully autonomous, context-sensitive data-preparation platforms.

ACKNOWLEDGEMENT
This research is supported by the Tertiary Education Trust Fund (TETFund), Nigeria, under the Academic Staff Training and Development (AST&D) intervention.

REFERENCES

Ahmadvand, A. (2020). User intent inference for web search and conversational agents. In *Proceedings of the 13th International Conference on Web Search and Data Mining* (pp. 911–912). ACM. https://doi.org/10.1145/3336191.3372187

Akella, A., Manatkar, A., Narayanam, K., & Mehta, S. (2025). CodeGenWrangler: Data Wrangling task automation using Code-Generating Models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Vol. 3: Industry Track)* (pp. 949–960). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-industry.70

Asai, T., & Yamauchi, K. (2023). Swap-Deep Neural Network: incremental inference and learning for embedded systems. In *Proceedings of the 13th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2024)* (pp. 418–427). SCITEPRESS. https://doi.org/10.5220/0012465500003654

Azeroual, O. (2020). Data wrangling in database systems: Purging of dirty data. *Data, 5*(2), Article 50. https://doi.org/10.3390/data5020050

Bilal, M., Ali, G., Iqbal, M. W., Anwar, M., Malik, M. S. A., & Kadir, R. A. (2022). Auto-PreP: Efficient and automated data preprocessing pipeline. *IEEE Access, 10*, 107764–107784. https://doi.org/10.1109/access.2022.3198662

Bogatu, A., Fernandes, A. A. A., Paton, N. W., & Konstantinou, N. (2019a). SynthEdit: Format transformations by example using edit operations. *Extending Database Technology*, 714–717. https://doi.org/10.5441/002/edbt.2019.94

Bogatu, A., Paton, N. W., Fernandes, A. A., & Koehler, M. (2019b). Towards automatic data format transformations: Data wrangling at scale. *The Computer Journal, 62*(7), 1044–1060. https://doi.org/10.1093/comjnl/bxy118

Chang, J., Liu, C., Huang, J., Mao, R., & Qin, J. (2025). LLaPipe: LLM-Guided Reinforcement Learning for Automated Data Preparation Pipeline Construction. *arXiv*, 1-20. https://doi.org/10.48550/arxiv.2507.13712

Chen, D., Huang, Y., Ma, Z., Chen, H., Pan, X., Ge, C., Gao, D., Xie, Y., Liu, Z., Gao, J., Li, Y., Ding, B., & Zhou, J. (2024). Data-Juicer: A One-Stop data processing system for large language models. In *Companion of the 2024 International Conference on Management of Data (SIGMOD)* (pp. 120–134). ACM. https://doi.org/10.1145/3626246.3653385

Cormier, K., Gagnier, K., Padron-Uy, J., Sareen, D., Parihar, A., Khmelevsky, Y., Hains, G., & Wong, A. (2025). Data Extraction, Transformation, and Loading (ETL) process automation and data warehouse implementation for algorithmic trading machine learning modelling. In *Proceedings of the IEEE International Systems Conference (SysCon)* (pp. 1–8). IEEE. https://doi.org/10.1109/syscon64521.2025.11014812

Côté, P., Nikanjam, A., Ahmed, N., Humeniuk, D., & Khomh, F. (2023). Data cleaning and machine learning: A systematic literature review. *arXiv*, 1-82. https://doi.org/10.48550/arXiv.2310.01765

Fan, M., Fan, J., Tang, N., Cao, L., & Du, X. (2024). AutoPrep: Natural Language Question-Aware Data Preparation with a Multi-Agent Framework. *arXiv*. https://doi.org/10.48550/arxiv.2412.10422

Fernandes, A. A. A., Koehler, M., Konstantinou, N., Pankin, P., Paton, N. W., & Sakellariou, R. (2023). Data preparation: A technological perspective and review. *SN Computer Science, 4*(4), Article 425. https://doi.org/10.1007/s42979-023-01828-8

Goyle, K., Xie, Q., & Goyle, V. (2024). DataAssist: A machine learning approach to data cleaning and preparation. In *Lecture Notes in Networks and Systems* (pp. 476–486). https://doi.org/10.1007/978-3-031-66431-1_33

Guduru, S. (2025). Cybersecurity workforce upskilling: CRISP-DM automation with Jupyter notebooks and immersive lab Terraform modules. *International Journal of Science and Research (IJSR), 14*(4), 907–911. https://doi.org/10.21275/sr25408010102

Heer, J., Hellerstein, J. M., & Kandel, S. (2019). Data wrangling. In S. Sakr & A. Zomaya (Eds.), *Encyclopedia of big data technologies* (pp. 1–8). Springer. https://doi.org/10.1007/978-3-319-63962-8_9-1

Huang, J., Guo, D., Wang, C., Gu, J., Lu, S., Inala, J. P., Yan, C., Gao, J., Duan, N., & Lyu, M. R. (2024). Contextualized data-wrangling code generation in computational notebooks. In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering (ASE)* (pp. 1282–1294). https://doi.org/10.1145/3691620.3695503

Jaimovitch-López, G., Ferri, C., Hernández-Orallo, J., Martínez-Plumed, F., & Ramírez-Quintana, M. J. (2022). Can language models automate data wrangling? *Machine Learning, 112*(6), 2053–2082. https://doi.org/10.1007/s10994-022-06259-9

Konstantinou, N., Abel, E., Bellomarini, L., Bogatu, A., Civili, C., Irfanie, E., Koehler, M., Mazilu, L., Sallinger, E., Fernandes, A. A. A., Gottlob, G., Keane, J. A., & Paton, N. W. (2019). VADA: An architecture for end-user informed data preparation. *Journal of Big Data, 6*(74), 1–32. https://doi.org/10.1186/s40537-019-0237-9

Konstantinou, N., & Paton, N. W. (2020). Feedback-driven improvement of data preparation pipelines. *Information Systems, 92*, Article 101480. https://doi.org/10.1016/j.is.2019.101480

Krajnc, D., Spielvogel, C. P., Grahovac, M., Ecsedi, B., Rasul, S., Poetsch, N., Traub-Weidinger, T., Haug, A. R., Ritter, Z., Alizadeh, H., Hacker, M., Beyer, T., & Papp, L. (2022). Automated data preparation for in vivo tumor characterization with machine learning. *Frontiers in Oncology, 12*, Article 1017911. https://doi.org/10.3389/fonc.2022.1017911

Lai, X., Shui, H., Ding, D., & Ni, J. (2021). Data-driven dynamic bottleneck detection in complex manufacturing systems. *Journal of Manufacturing Systems, 60*, 662–675. https://doi.org/10.1016/j.jmsy.2021.07.016

Lee, G. Y., Alzamil, L., Doskenov, B., & Termehchy, A. (2021). A survey on data cleaning methods for improved machine learning model performance. arXiv. https://doi.org/10.48550/arXiv.2109.07127

Li, X., & Döhmen, T. (2024). Towards efficient data wrangling with LLMs using code generation. In *Proceedings of the Eighth Workshop on Data Management for End-to-End Machine Learning (DEEM)* (pp. 62–66). ACM. https://doi.org/10.1145/3650203.3663334

Liu, L., Hasegawa, S., Sampat, S. K., Xenochristou, M., Chen, W., Kato, T., Kakibuchi, T., & Asai, T. (2024). AutoDW: Automatic data wrangling leveraging large language models. In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering (ASE)* (pp. 2041–2052). ACM. https://doi.org/10.1145/3691620.3695267

Liu, L., Hasegawa, S., Sampat, S. K., Bahrami, M., Chen, W., Toyota, K., Kato, T., Akazaki, T., Ura, A., & Asai, T. (2025). AutoDW-TS: Automated Data Wrangling for Time-Series Data. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM)* (pp. 5888–5897). ACM. https://doi.org/10.1145/3746252.3761524

Liu, M., & Yu, D. (2022). Towards intelligent e-learning systems. *Education and Information Technologies*, *28*(7), 7845–7876. https://doi.org/10.1007/s10639-022-11479-6

Narayan, A., Chami, I., Orr, L., & Ré, C. (2022). Can foundation models wrangle your data? *Proceedings of the VLDB Endowment, 16*(4), 738–746. https://doi.org/10.14778/3574245.357425

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, Article n71. https://doi.org/10.1136/bmj.n71

Paton, N. (2019). Automating data preparation: Can we? Should we? Must we? Research Explorer, the University of Manchester. https://research.manchester.ac.uk/en/publications/automating-data-preparation-can-we-should-we-must-we/

Rajenthiram, K. (2025). Optimizing Data Analytics Workflows Through User-Driven Experimentation: Progress and Updates. In *Proceedings of the 2025 IEEE/ACM International Conference on AI Engineering – Software Engineering for AI (CAIN)* (pp. 236–240). IEEE. https://doi.org/10.1109/cain66642.2025.00035

Rezig, E. K., Ouzzani, M., Elmagarmid, A. K., Aref, W. G., & Stonebraker, M. (2019). Towards an end-to-end human-centric data cleaning framework. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics (HILDA)* (pp. 1–7). ACM. https://doi.org/10.1145/3328519.3329133

Seedat, N., Crabbé, J., & Van Der Schaar, M. (2022). Data-SUITE: Data-centric identification of in-distribution incongruous examples. In *Proceedings of the 39th International Conference on Machine Learning* (ICML) (Vol. 162, pp. 19467–19496). https://proceedings.mlr.press/v162/seedat22a.html

Shome, A., Cruz, L., Spinellis, D., & Arie, V. D. (2024). Understanding feedback mechanisms in machine learning Jupyter notebooks. *arXiv*, 1-37. https://doi.org/10.48550/arxiv.2408.00153

Somasundaram, P. (2022). Streamlining data wrangling processes through automation and tooling. *International Journal of Science and Research (IJSR), 11*(2), 1323–1325. https://doi.org/10.21275/sr24418101747

Somayajula, S. K. (2025). Advanced ETL optimization: A framework for next-generation data integration. *International Journal of Computer Engineering & Technology, 16*(1), 381–406. https://doi.org/10.34218/ijcet_16_01_035

Vanalkar, A. V., & Pazare, V. (2025). Comparative study: Regular furnace vs smart furnace based on IoT integration. *International Journal of Scientific Research in Engineering and Management (IJSREM), 9*(4), 1–4. https://doi.org/10.55041/IJSREM45373

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining* (pp. 29–39).

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., & He, L. (2023). A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems, 135*, 364–381. https://doi.org/10.1016/j.future.2022.05.014

Xu, S., Yao, Y., Xu, F., Gu, T., Xu, J., & Ma, X. (2023). Data quality matters: A case study of obsolete comment detection. In *Proceedings of the 2023 IEEE/ACM 45th International Conference on Software Engineering*

(ICSE) (pp. 781–793). IEEE. https://doi.org/10.1109/icse48619.2023.00074

Yasin, H. M., & Khorsheed, A. K. (2025). Automated Data Cleaning in Large Databases Using Machine Learning Methods. *Asian Journal of Research in Computer Science, 18*(5), 364–386. https://doi.org/10.9734/ajrcos/2025/v18i5661

Yin, Y., Wang, B., Zuo, H., & Childs, P. (2024). Effects of different Human-In-The-Loop approaches on Human-AI Co-Design: A comparison between Human-Learning HITL approach and Machine-Learning HITL Approach. In *Proceedings of the ASME 2024 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference* (Vol. 2A, Article V02AT02A050). ASME. https://doi.org/10.1115/detc2024-143295

APPENDIX A (QUALITY ASSESSMENT SCORES)

The quality assessment scores presented in Table A1 were assigned using the six-point rubric described in Table 3. Each study was independently evaluated against the criteria of task clarity, presence of feedback, automation strategy, evidence of evaluation, reproducibility, and feedback-reuse capability. Scores were assigned on a scale of 0 (not addressed), 0.5 (partially addressed), or 1 (fully addressed), with a maximum possible score of 6. Only studies scoring ≥3 were included in the final synthesis.

Table A1: Quality Assessment Scores of Included Studies

Study	Task Clarity	Feedback Presence	Automation	Evaluation	Reproducibility	Feedback Reuse	Total (/6)
Heer et al. (2019)	1	1	0.5	0.5	0.5	0	3.5
Bogatu et al. (2019a)	1	1	1	0.5	1	0	4.5
Bogatu et al. (2019b)	1	1	1	0.5	1	0	4.5
Paton (2019)	0.5	0.5	0.5	0	0.5	0	2.0
Konstantinou et al. (2019)	1	1	1	1	1	1	6.0
Konstantinou et al. (2020)	1	1	1	1	1	1	6.0
Rezig et al. (2019)	1	1	0.5	0.5	0.5	0	3.5
Azeroual (2020)	1	1	0.5	0.5	0.5	0	3.5
Somasundaram (2022)	1	1	0.5	0.5	0.5	0	3.5
Rajenthiram (2025)	1	1	1	1	0.5	0.5	5.0
Bilal et al. (2022)	1	1	1	1	1	0.5	5.5
Fan et al. (2024)	1	1	1	1	1	0.5	5.5
Goyle et al. (2024)	1	1	1	1	0.5	0.5	5.0
Liu et al. (2024)	1	1	1	1	1	0.5	5.5
Liu et al. (2025)	1	1	1	1	1	0.5	5.5
Krajnc et al. (2022)	1	1	1	1	1	0.5	5.5
Yasin & Khorsheed (2025)	1	1	1	1	0.5	0.5	5.0
Narayan et al. (2022)	1	1	1	1	1	0.5	5.5
Li & Döhmen (2024)	1	1	1	1	1	0.5	5.5
Huang et al. (2024)	1	1	1	1	1	0.5	5.5
Akella et al. (2025)	1	1	1	1	1	0.5	5.5
Chang et al. (2025)	1	1	1	1	1	1	6.0
Chen et al. (2024)	1	1	1	1	1	0.5	5.5
Somayajula (2025)	1	1	1	0.5	0.5	0	4.0
Seedat et al. (2022)	1	0.5	1	1	1	0	4.5
Wu et al. (2022)	0.5	0.5	0.5	0.5	0.5	0	2.5
Shome et al. (2024)	1	1	0.5	0.5	0.5	0	3.5
Jaimovitch-López et al. (2022)	1	1	1	1	1	0.5	5.5
Xu et al. (2023)	1	1	1	1	1	0.5	5.5

APPENDIX B (EXTRACTED STUDY DATASET)

This appendix provides the structured dataset extracted from the 29 primary studies included in the systematic literature review. The dataset follows the coding schema in Table 4 and includes key attributes such as feedback type, data preparation stage, automation technique, architectural approach, dataset characteristics, and evaluation methods. This structured representation supports transparency, reproducibility, and traceability of the synthesis process.

Table B1: Extracted Dataset of Included Studies

https://dx.doi.org/10.4314/swj.v21i1.39
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

Authors & Year	Feedback Type	Stage(s)	Technique	Architecture	Dataset Used	Evaluation Method
Heer et al. (2019).	Explicit	Cleaning, Transformation	Heuristic	Interactive	Real-world	User study
Bogatu et al. (2019a).	Explicit	Transformation	Rule-based	Rule-based	Synthetic + real	Performance metrics
Bogatu et al. (2019b).	Explicit	Transformation	Rule-based	Rule-based	Synthetic	Performance metrics
Paton (2019)	Conceptual	All	Conceptual	Conceptual	N/A	Conceptual analysis
Konstantinou et al. (2019).	Hybrid	Cleaning, Matching	Hybrid	Feedback repository	Real-world	Quantitative + system evaluation
Konstantinou et al. (2020).	Hybrid	Cleaning, Matching	ML-based	ML-driven	Real-world	Quantitative
Rezig et al. (2019)	Explicit	Cleaning	Heuristic	Interactive	Real-world	User evaluation
Azeroual (2020)	Explicit	Cleaning	Rule-based	Interactive	Real-world	Qualitative
Somasundaram (2022)	Explicit	Cleaning	Rule-based	Rule-based	Real-world	Case-based
Rajenthiram (2025)	Hybrid	Cleaning, Validation	Adaptive	Adaptive workflows	Real-world	Experimental
Bilal et al. (2022).	Hybrid	Cleaning	ML-based	ML-driven	Real-world	Quantitative
Fan et al. (2024)	Hybrid	Transformation	LLM-assisted	ML-driven	Real-world	Benchmark + qualitative
Goyle et al. (2024).	Implicit	Cleaning	ML-based	ML-driven	Real-world	Quantitative
Liu et al. (2024).	Implicit	Cleaning	LLM-based	ML-driven	Real-world	Benchmark
Liu et al. (2025).	Implicit	Cleaning, Transformation	LLM-based	ML-driven	Real-world	Benchmark
Krajnc et al. (2022).	Explicit	Cleaning	ML-based	ML-driven	Medical datasets	Quantitative
Yasin & Khorsheed (2025).	Explicit	Cleaning	ML-based	ML-driven	Real-world	Experimental
Narayan et al. (2022).	Implicit	Cleaning, Integration	Foundation models	ML-driven	Large-scale datasets	Benchmark
Li & Döhmen (2024).	Implicit	Transformation	LLM-based	ML-driven	Real-world	Benchmark
Huang et al. (2024).	Implicit	Transformation	LLM-based	ML-driven	Notebook datasets	Benchmark
Akella et al. (2025).	Implicit	Cleaning, Transformation	LLM-based	ML-driven	Real-world	Benchmark
Chang et al. (2025)	Hybrid	Pipeline construction	RL + LLM	Adaptive	Real-world	Experimental
Chen et al. (2024).	Hybrid	Transformation	LLM-based	ML-driven	Large-scale datasets	Benchmark
Somayajula (2025)	Explicit	Integration	Rule-based	Rule-based	Enterprise data	Case study
Seedat et al. (2022).	Implicit	Profiling, Validation	ML-based	ML-driven	Real-world	Quantitative
Wu et al. (2022).	Conceptual	All	Survey	Conceptual	N/A	Literature synthesis
Shome et al. (2024).	Implicit	Validation	Behavioural mining	ML-driven	Notebook logs	Qualitative
Jaimovitch-López et al. (2022).	Implicit	Transformation	LLM-based	ML-driven	Real-world	Benchmark
Xu et al. (2023)	Explicit	Validation	ML-based	ML-driven	Real-world	Quantitative