International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 03 | March – 2026 DOI:10.55041/ISJEM05905
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Machine Learning Techniques for Cleaning Raw Data

Mrs. Jaspreet Kaur
M.Tech Scholar, Computer Science and Engineering RCET, Bhilai, Chhattisgarh, India

Dr. Neelabh Sao
Associate Professor, Department of Information Technology RCET, Bhilai, Chhattisgarh, India

A**BSTRACT**

Machine learning (ML) systems typically rely on high-quality, well-structured datasets for effective performance. However, data collected from real-world environments is often incomplete, noisy, inconsistent, and heterogeneous. Such imperfections negatively impact model accuracy, generalization ability, and reliability, particularly in dynamic and large-scale applications. This study reviews existing research on data cleaning techniques and their role in improving machine learning outcomes. It examines how different types of data imperfections arise and how they influence various stages of the ML pipeline. The review also evaluates current approaches for handling issues such as missing values, outliers, and data inconsistencies, along with their limitations in real-world scenarios. Furthermore, it highlights the need for more adaptive and scalable solutions that integrate data cleaning within the learning process. The study concludes that data quality should be treated as a critical factor throughout the ML lifecycle rather than as a standalone pre-processing step.

Keywords: Data Cleaning, Machine Learning, Data Pre-processing, Outlier Detection, Imputation

1. INTRODUCTION

Machine learning has become a key technology in modern applications, including healthcare, finance, transportation, and intelligent systems. The effectiveness of these systems largely depends on the quality of data used during training and deployment. In many research settings, machine learning models are developed using clean and well-prepared benchmark datasets. However, such conditions rarely exist in practical environments.

In real-world scenarios, data is often collected from multiple sources such as sensors, user interactions, enterprise systems, and online platforms. This data frequently contains issues such as missing values, noise, inconsistencies, outliers, and bias. Additionally, data distributions may change over time, leading to challenges such as concept drift. These imperfections can distort data patterns and reduce the performance and reliability of machine learning models.

The gap between ideal datasets used in research and imperfect real-world data presents a major challenge for deploying robust ML systems. As a result, data cleaning and pre-processing have become essential steps in the machine learning pipeline. However, traditional cleaning methods are often insufficient for handling large-scale and dynamic data. This has led to increasing interest in developing advanced techniques that combine data cleaning with machine learning to improve overall system performance.

2. LITERATURE REVIEW

Vidyalakshmi (2017) explains that the rapid increase in data due to web technologies and automated systems has led to very large datasets, but these are often low in quality for analysis. She highlights that data mining, or knowledge discovery, requires clean and reliable data to extract useful patterns.
Therefore, data cleansing is essential to improve data quality and make large datasets suitable for accurate analysis and meaningful insights.

Ridzuan and Zainon (2019) highlight the growing importance of data quality in the era of big data and data-driven decision-making. They emphasize that inaccurate, incomplete, or inconsistent data can lead to unreliable analysis and poor business decisions. The study

explains that data cleansing plays a key role in improving data quality by detecting and correcting errors, as well as handling missing values.

The authors also point out that many organizations suffer financial losses due to poor data quality, especially in systems like customer relationship management and supply chain management. Therefore, they conclude that effective data cleansing is essential to ensure accurate analysis and support better decision-making

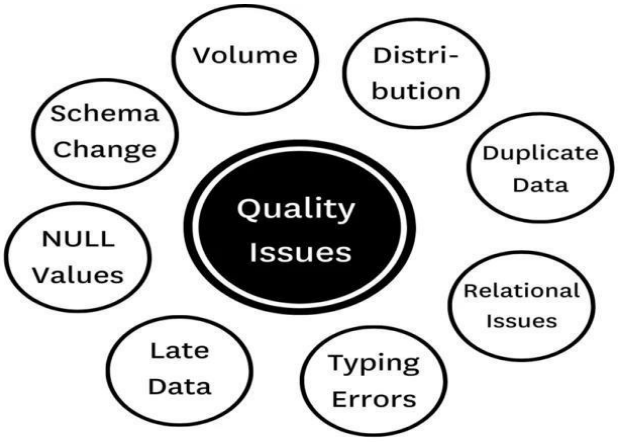


Fig. 2.1: Common Data Quality Issues in Large Datasets (Source: [1], [2], [7])

Li et al. (2020) examined how different data cleaning techniques influence the performance of machine learning classification models. Their study highlights that ML models are sensitive to various forms of noise present in both input features and labels. Instead of relying solely on data cleaning, the research community has also focused on developing noise-tolerant algorithms, such as robust decision trees, regularization methods, and ensemble techniques like bagging. The findings suggest that improving model robustness and applying selective cleaning strategies are both important for handling noisy datasets.

Côté et al. (2022) conducted a comprehensive systematic literature review to explore the relationship between data cleaning and machine learning. The study categorizes research into two main directions: using machine learning to improve data cleaning processes and applying data cleaning techniques to enhance ML performance. By analyzing a large collection of research papers, the authors provide an overview of current methods, tools, and challenges in this domain. The study also emphasizes the need for structured and unbiased review methodologies to ensure reliable insights and identifies several future research directions.

Alotaibi et al. (2023) investigated data cleaning challenges in big data environments, particularly focusing on streaming data. The study explains that big data differs from traditional datasets due to its scale, speed, and diversity of formats. These characteristics make data cleaning more complex and require specialized techniques capable of handling continuous data flow and heterogeneous structures. The authors conclude that conventional data cleaning approaches are insufficient for big data scenarios and must be extended to support real-time and scalable processing.

Côté et al. (2024) further explored the growing integration of machine learning into real-world systems across various domains. The study highlights that ML models are increasingly embedded within software systems that require high reliability and quality assurance. As these systems become more complex, ensuring data quality becomes essential for maintaining accurate and trustworthy outputs. The authors stress that data cleaning should be treated as an integral component of the machine learning lifecycle rather than a separate preprocessing step.

ISSN: 2583-6129
DOI:10.55041/ISJEM05905

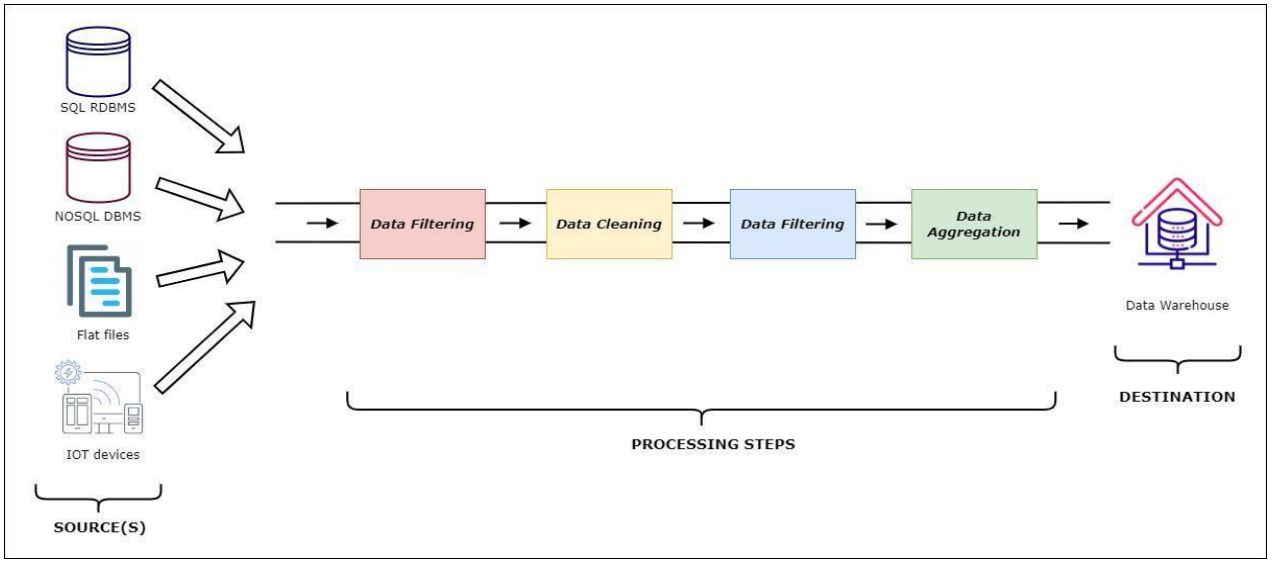


Fig. 2.2: Impact of Data Cleaning on Machine Learning Performance (Source: [3], [4], [6])

Adhithya Srija et al. (2024) discuss the challenges caused by the rapid growth and complexity of modern datasets. They highlight that data often contains issues such as missing values, outliers, and inconsistencies, which negatively affect the accuracy of analysis and machine learning models.

The study emphasizes that data cleaning is a crucial step to improve data quality and ensure reliable results in downstream tasks. Proper handling of these imperfections helps organizations generate more accurate insights and make better decisions

Panwar (2024) highlights the importance of maintaining accurate and reliable data in the big data era. He explains that traditional data cleansing methods, which rely on manual efforts and rule-based techniques, are time-consuming and less effective in handling large and complex datasets.

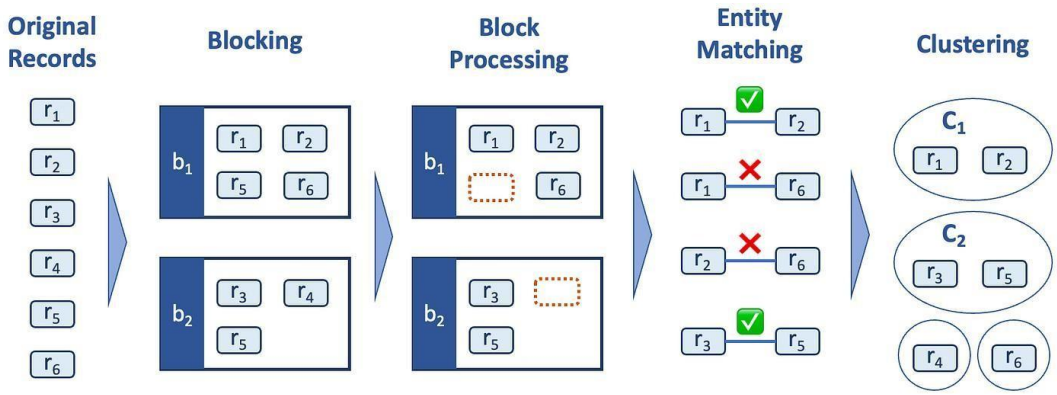
The study emphasizes that AI-based data cleansing offers a more efficient solution by using machine learning and natural language processing to automatically detect and correct errors. This approach improves both the speed and accuracy of data cleaning, making it more suitable for modern data environments.

Martins et al. (2025) evaluated the efficiency and scalability of various data cleaning and preprocessing tools using large real-world datasets. Their findings indicate that data cleaning is one of the most time-intensive stages in data analysis, often consuming a significant portion of effort. While many tools are designed to automate cleaning tasks, their performance tends to decline when applied to large-scale datasets. The study highlights the importance of developing scalable solutions that can efficiently handle increasing data volumes without compromising accuracy.

Entity Resolution (ER) is the process of identifying records that refer to the same real-world entity, which is essential for improving data quality and integration. It typically involves comparing multiple attributes of data records to detect duplicates, but this process becomes computationally expensive for large datasets.

ISSN: 2583-6129
DOI:10.55041/ISJEM05905

Fig. 2.3: Entity Resolution and Deduplication Process (Source: [10])



To address this issue, techniques like blocking are used to reduce the number of comparisons by grouping similar records. However, ER still remains time-consuming at scale. The Dedoop framework improves efficiency by using Hadoop and MapReduce, enabling parallel processing of large datasets, which significantly reduces execution time and enhances scalability.

3. METHODOLOGY

This study adopts a **systematic literature review approach** to analyze machine learning techniques for cleaning raw data.

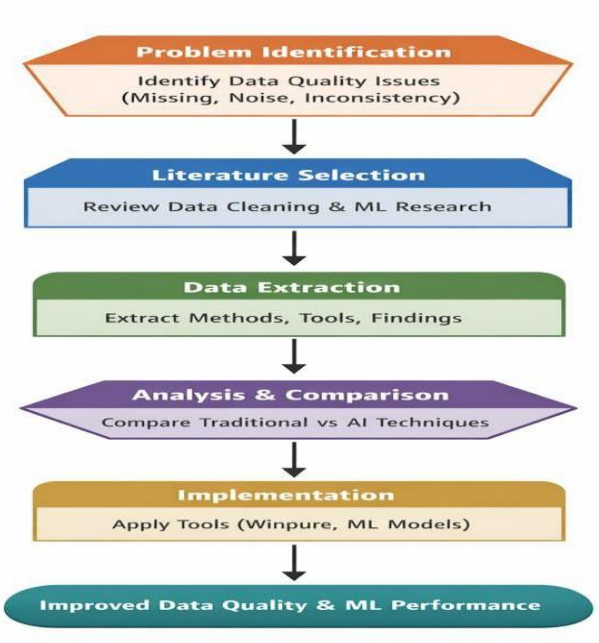


Fig. 3.1: Workflow of Data Cleaning and Analysis in ML Systems

Literature Collection

Relevant research papers were collected from academic sources such as Google Scholar,

IEEE Xplore, and other journals. The focus was on studies related to data cleaning, preprocessing, big data, and machine learning.

3.1. Selection of Studies

Papers were selected based on:

- Relevance to data cleaning in machine learning
- Coverage of real-world data issues (missing values, noise, inconsistency)
- Recent publications (2017–2025) to ensure updated analysis

3.2. Categorization of Approaches

The selected studies were grouped into:

- Traditional data cleaning methods
- Big data cleaning techniques
- AI/ML-based data cleaning approaches

3.3. Comparative Analysis

Different techniques were analyzed and compared based on:

- Effectiveness in improving data quality
- Impact on machine learning performance
- Scalability and handling of large datasets

3.4. Synthesis of Findings

The insights from all studies were combined to identify:

- Common challenges in data cleaning
- Emerging trends (AI-based cleaning, automation)
- Best practices for improving data quality in ML systems

4. LIMITATIONS AND CHALLENGES

A. Data Quality Issues in Real-World Data

Real-world datasets contain multiple imperfections such as missing values, noise, outliers, and inconsistent labels. As highlighted in the Clean ML Study, machine learning models are highly sensitive to such noise, especially label and input noise, which can significantly degrade model performance. Existing approaches often fail to handle complex, non-uniform data errors effectively.

B. Limited Effectiveness of Traditional Data Cleaning Methods

Conventional data cleaning techniques such as imputation, outlier removal, and normalization are not universally effective. According to Data Cleaning and Machine Learning: A Systematic Literature Review, the impact of cleaning methods varies depending on dataset characteristics, making it difficult to generalize a single approach across domains.

C. Lack of Standardized Evaluation Frameworks

A major challenge identified in the same study by Data Cleaning and Machine Learning: A Systematic Literature Review is the absence of standardized benchmarks and evaluation metrics. This makes it difficult to compare different data cleaning techniques and assess their effectiveness objectively.

D. Scalability Issues with Large Datasets

Handling large-scale data remains a critical challenge. The study Performance and Scalability of Data Cleaning Tools highlights that many data cleaning tools fail to scale efficiently when dealing with massive datasets, leading to increased processing time and reduced performance.

E. Complexity of Big Data Characteristics

Big data introduces additional challenges such as volume, velocity, and variety. As discussed in Cleaning Big Data Streams, traditional data cleaning methods are not suitable for streaming and heterogeneous data environments, requiring more advanced and adaptive techniques

Dependency on Clean Data Assumptions in ML Models

Most machine learning models are designed under the assumption of clean and well-structured data. However, as emphasized in your review topic *Machine Learning Techniques for Cleaning Raw Data*, this assumption does not hold in real-world scenarios, leading to reduced reliability and poor generalization.

F. High Time and Resource Consumption

Data cleaning is highly resource-intensive. According to Performance and Scalability of Data Cleaning Tools, up to 80% of data science effort is spent on preprocessing tasks, making it inefficient and costly.

G. Lack of Adaptive and Automated Cleaning Systems

Current approaches lack automation and adaptability. As indicated in Data Cleaning and Machine Learning Review, there is a need for intelligent systems that can dynamically adjust cleaning strategies based on data characteristics.

H. Challenges in Handling Data Drift and Dynamic Environments

Real-world data is dynamic and evolves over time (concept drift). Existing cleaning methods are mostly static and fail to adapt to such changes, leading to performance degradation in deployed ML systems.

I. Trade-off between Cleaning and Information Loss

Excessive cleaning (e.g., removing outliers or duplicates) can lead to loss of useful information. Studies like Clean ML Study show that some cleaning operations may even reduce model accuracy.

5. KEY FINDINGS

A. Data cleansing is essential for handling big data issues such as inconsistency, missing values, and duplication, which directly affect data quality.
B. Traditional data cleansing methods are useful but are less effective for large-scale and complex big data environments compared to modern approaches.
C. AI-based data cleansing techniques significantly improve efficiency, accuracy, and scalability by automating error detection and correction.
D. Clean data has a strong positive impact on downstream tasks like machine learning and analytics, leading to better model performance and decision-making.
E. The use of diverse datasets (public, synthetic, and real-world) shows that data quality problems vary across domains, requiring adaptable cleansing techniques.
F. Tools like Winpure and similar automated solutions help in identifying duplicate and inconsistent data, making the cleansing process faster and more reliable.
G. Proper data cleaning reduces uncertainty in analysis and enhances the extraction of meaningful insights from large datasets.
H. Overall, effective data cleansing—especially when combined with AI—plays a critical role in improving data reliability, analytical accuracy, and business outcomes.

6. CONCLUSION

This review examined the role of data cleaning in machine learning and highlighted its critical importance in ensuring model performance, reliability, and robustness. The analysis of existing studies indicates that real-world data is inherently imperfect, containing issues such as missing values, noise, inconsistencies, and bias, which significantly affect machine learning outcomes. While various data cleaning techniques have been developed, their effectiveness is highly context-dependent and varies across datasets and applications.

The review also revealed that traditional data cleaning methods are often insufficient for handling large-scale and dynamic data environments. Challenges such as scalability, lack of standardized evaluation frameworks, and high computational costs continue to limit the efficiency of existing approaches. Furthermore, the assumption of clean data in many machine learning models creates a gap between research and real-world deployment.

Overall, data cleaning should not be considered merely a preprocessing step but a fundamental component of the entire machine learning lifecycle. Future research should focus on developing adaptive, automated, and scalable data cleaning techniques that can handle real-world data complexities and improve the reliability of intelligent systems.

REFERENCES

1. Vidyalakshmi, V. (2017). *Data Mining and Data Cleansing Techniques*.
2. Ridzuan, F., & Zainon, W. M. N. W. (2019). *A Review on Data Cleansing Methods for Big Data*.
3. Li, et al. (2020). *CleanML: A Study for Evaluating the Impact of Data Cleaning on Machine Learning Classification Tasks*.
4. P.-O. Côté, et al. (2022). *Data Cleaning and Machine Learning: A Systematic Literature Review*.
5. O. Alotaibi, E. Pardede, & S. Tomy (2023). *Cleaning Big Data Streams: A Systematic Literature Review*.
6. P.-O. Côté, et al. (2024). *Data Cleaning and Machine Learning: Recent Advances and Challenges*.
7. Srija, A. A., Kedareswari, K., Sahithya, P., Aswini, N., & Vuyyuru, V. A. (2024). *Unveiling the Significance of Data Cleaning: A Review of its Impact on Downstream Tasks in Machine Learning and Analytics*.
8. Panwar, V. (2024). *AI-Powered Data Cleansing for Big Data: Enhancing Accuracy and Efficiency*.
9. P. Martins, et al. (2025). *Performance and Scalability of Data Cleaning and Preprocessing Tools: A Benchmark on Large Real-World Datasets*..
Kolb, L., Thor, A., & Rahm, E. (2012). *Dedoop: Efficient Deduplication with Hadoop*.