Impact of Data Cleaning on Machine Learning Model Accuracy with Labeled Sections

Bhavesh Sheshnath Prasad
Master of Computer Applications
Tilak Maharashtra Vidyapeeth University, India

Abstract — Data cleaning is widely acknowledged as a critical step in preparing datasets for machine learning (ML). This review examines how data cleaning influences ML model accuracy by synthesizing recent literature. We survey systematic studies and empirical experiments addressing cleaning tasks (e.g., handling missing values, label errors, duplicates) and their effects on classification, regression, and clustering models. Key papers include the CleanML benchmark study, a broad systematic review of data cleaning for ML, an empirical analysis of data quality dimensions, and the COMET system for prioritizing cleaning efforts. Overall, we find that targeted cleaning generally improves accuracy, but gains vary by error type, data context, and resource constraints. For example, imputing missing values or correcting label errors often enhances performance, whereas removing duplicates or fixing minor inconsistencies may have little or no effect. We highlight limitations such as high cleaning costs and unpredictable benefits in real-world settings, and discuss strategies like automated tools and iterative methods (e.g., COMET, ActiveClean) to focus effort on the most impactful data issues. Our synthesis points to a "data-centric" ML paradigm: effective cleaning must be guided by downstream tasks. We conclude with practical insights (e.g., prioritize feature/label accuracy) and future directions, including tighter ML–cleaning integration and automated, cost-aware cleaning processes.

Keywords: Data Cleaning; Data Quality; Machine Learning; Model Accuracy; Data-Centric AI; Data Preprocessing; Error Correction

I. Introduction

Machine learning systems rely fundamentally on the data used for training and evaluation. As multiple authors observe, "the performance of an ML model is highly dependent on the quality of the data it has been trained on". In practice, real-world datasets often contain errors such as missing values, mislabels, duplicates, or inconsistent formats, which can degrade model accuracy. Ensuring data quality is therefore crucial; data cleaning—which involves detecting and correcting such errors—is viewed as "essential for reliable and accurate ML predictions". Historically, data cleaning was treated as a separate preprocessing step, but the paradigm is shifting toward data-centric AI, where cleaning and modeling are interwoven. In this "cleaning for ML" view, data quality and model performance are treated as symbiotic components of the pipeline. Consequently, there is growing interest in understanding how specific cleaning actions affect ML outcomes, and in developing methods to automate or optimize the cleaning process. This review synthesizes recent studies on the impact of data cleaning on ML accuracy, outlining their methods, findings, and implications.

II. Review of Literature

1) CleanML Benchmark (Li et al., 2021): Li and colleagues introduced CleanML, a benchmark to systematically evaluate data cleaning effects on classification accuracy. This study applied multiple cleaning algorithms to 14 real-world datasets exhibiting five types of errors (missing values, duplicate records, mislabels, outliers, and inconsistencies), and measured the downstream impact on seven common classifiers. The CleanML analysis yielded several nuanced observations. For missing data, they found that imputing values typically improves or matches the performance achieved by simply deleting rows, especially when appropriate imputation methods are chosen. In contrast, cleaning outliers often produced only insignificant accuracy gains. Notably, correcting label errors usually had a positive effect: fixing mislabeled instances tended to increase accuracy, particularly when using ensemble models like boosting. Conversely, removing duplicate records more often failed to improve performance and, in some cases, slightly hurt it. The authors stress that the impact of cleaning varies widely across datasets and error types. These results imply that data scientists should not assume all cleaning yields benefits; rather, each cleaning action must be evaluated in context. CleanML provides an extensible framework and has influenced how researchers design empirical cleaning studies.
2) Systematic Literature Review (Côté et al., 2024): A broader perspective is offered by a systematic literature review covering papers from 2016–2022. Côté et al. classify data cleaning activities relevant to ML into categories such as feature value cleaning, label cleaning, entity matching, outlier detection, imputation, and holistic cleaning. They report synthesizing 101 studies and highlight a dual trend: not only are many methods developed to clean data for ML, but ML techniques are also used for cleaning tasks (e.g. using learning to detect errors). Their analysis emphasizes that a variety of tools and frameworks exist, yet many are still at an exploratory stage. The authors list 24 recommendations for future work, pointing to challenges like scalability, automation, and integration of cleaning with ML pipelines. Importantly, they conclude that the literature contains "many promising data cleaning techniques that can be further extended". This survey underscores the community's recognition that improved data quality is key to ML success and encourages development of more robust, automated cleaning approaches.
3) Effects of Data Quality (Budach et al., 2024): In a large-scale empirical study, Budach and colleagues examined how six dimensions of data quality affect a wide range of ML tasks. They systematically introduced controlled

errors (e.g. missingness, noise, class imbalance) into several datasets and measured the performance of 15 algorithms across classification, regression, and clustering scenarios. A key output is a summary table that rates each quality dimension's impact on each task. For classification, the most detrimental issues were feature accuracy, target accuracy (label correctness), and completeness (missing values). For example, they found that if more than ~20% of labels are incorrect, classifier performance can degrade sharply, whereas up to ~20% mislabeling had only moderate effect. Conversely, duplicates, minor class imbalance, and minor consistency issues generally had low impact. Their findings suggest practical guidelines: data scientists might prioritize handling missing values and verifying labels, while de-emphasizing perfect deduplication or balancing when resources are limited. For regression tasks, they similarly observe that incomplete or noisy features heavily degrade performance, whereas continuous target imbalance and uniqueness issues have less effect. Overall, this study provides a quantitative basis for anticipating which data issues will hurt model accuracy the most.

4) COMET System (Mohammed et al., 2025): Mohammed and co-authors propose COMET (Cleaning Optimization and Model Enhancement Toolkit), a system that recommends which data fields to clean in order to maximize ML accuracy under budget constraints. COMET works by iteratively "polluting" features (injecting known errors) and observing the drop in prediction accuracy. By estimating how much each feature's cleaning would improve accuracy relative to its cost, COMET assigns a cleaning priority score. In empirical evaluation over diverse datasets and error scenarios, COMET achieved substantially higher accuracy than baseline approaches: up to 52 percentage points improvement in some cases, and on average around 5 points better than feature-importance or random cleaning strategies. These results illustrate that ML-aware, incremental cleaning strategies can yield significant gains. COMET embodies the shift to automated, task-driven cleaning: rather than cleaning data indiscriminately, it focuses effort where it yields the largest benefit for the model.

III. Limitations and Challenges

A. Subjectivity in Cleaning Decisions

Some cleaning actions, such as removing outliers or imputing missing values, involve subjective decisions that may vary by domain expert or data scientist. This subjectivity can introduce bias and reduce the reproducibility of ML experiments.

B. Risk of Over-cleaning

Excessive cleaning or aggressive filtering may remove valuable data or natural variability, potentially introducing bias or reducing the model's generalization ability. Finding the right balance is often non-trivial.

C. Lack of Standardization and Best Practices

There is no universally accepted framework or set of guidelines for data cleaning across industries. Teams may use ad-hoc methods or undocumented processes, leading to inconsistent quality and difficult-to-reproduce results.

D. Integration Difficulties in ML Pipelines

Data cleaning is often treated as a pre-processing step, disconnected from the model training phase. This siloed approach limits feedback loops where model performance could guide more effective cleaning efforts.

E. Dependence on Domain Knowledge

Effective cleaning often requires contextual understanding of the data. Without domain expertise, automated methods may fail to recognize critical issues or apply inappropriate transformations.

F. Strategies to Overcome Challenges

To address the numerous challenges associated with data cleaning, researchers and practitioners have developed a multi-faceted set of strategies focusing on automation, prioritization, integration, and human-in-the-loop systems:

1) Automation Through Toolkits and Frameworks:

The rise of automated systems like COMET, ActiveClean, and HoloClean exemplifies a move toward reducing human intervention by programmatically identifying and correcting data errors. These systems often leverage existing ML models to estimate the importance of specific features or records in relation to prediction performance, thereby prioritizing which portions of the data to clean. For instance, ActiveClean uses model gradients to identify which training records, if cleaned, would yield the largest improvement in accuracy.

2) Prioritization Based on Impact and Cost:

Given that data cleaning can be resource-intensive, many recent solutions focus on cleaning for maximum return on investment. By evaluating the marginal utility of cleaning a feature or instance, systems like COMET can produce prioritized cleaning schedules. These schedules ensure that practitioners use limited resources (e.g., time, annotation budgets) on areas with the highest potential to improve model accuracy. This type of cost-aware decision-making transforms data cleaning from a generic preprocessing step into a strategic, task-aligned activity.

3) Integration into ML Workflows (Data-Centric AI):

Instead of treating data cleaning as a siloed activity, new workflows embed cleaning into ML pipeline stages such as feature selection, model training, and validation. For example, a cleaning action might be followed by re-training, and model error analysis can point to remaining data issues. This tight feedback loop forms the foundation of data-centric AI, in which improving training data quality often yields better results than modifying the algorithm.

4) Human-in-the-Loop and Iterative Cleaning:

While automation can handle large volumes, human judgment is often needed for complex or domain-specific errors (e.g., resolving ambiguous labels or interpreting inconsistent formats). Human-in-the-loop strategies allow practitioners to incrementally validate cleaning suggestions, thereby improving model robustness while minimizing manual labor. Tools like LabelStudio and Snorkel support

iterative, semi-automated approaches that combine weak supervision with human labeling in cycles.

IV. Key Findings

Through a synthesis of benchmark studies, empirical evaluations, and literature reviews, several critical insights emerge about the role and effectiveness of data cleaning in improving ML accuracy:

A. Selective Cleaning is More Effective Than Blanket Cleaning:

Contrary to common assumptions, not all cleaning activities result in performance improvement. Studies show that targeted correction of high-impact errors—especially label errors and missing values—tends to offer the greatest returns. Cleaning low-impact issues like exact duplicates or minor format inconsistencies often wastes resources without improving accuracy.

B. Different Models Have Different Sensitivities to Data Errors:

For example, ensemble methods such as random forests and gradient boosting often show greater robustness to noisy data compared to linear models. Conversely, neural networks may overfit mislabeled data if not cleaned properly. This suggests that the need for and impact of cleaning depend on the model choice and that cleaning strategies may need to be tailored to specific algorithms.

C. Data Cleaning Has Diminishing Returns:

There exists a threshold effect: for some datasets, cleaning beyond a certain point yields marginal or even negative gains. For instance, fixing 10% of mislabeled examples may boost accuracy significantly, but additional effort beyond 50% may not justify the cost. This diminishing return on cleaning investment highlights the need for prioritization and early stopping strategies.

D. Data Quality Has an Uneven Impact Across ML Tasks:

Budach et al. (2024) showed that classification tasks are more sensitive to label noise, while regression tasks suffer more from missing or noisy features. Clustering tasks, on the other hand, were particularly affected by outliers. These findings emphasize that cleaning must be task-specific: what is important in one ML setting may be less so in another.

E. Cleaning Improves Generalization, Not Just Accuracy:

Several studies indicate that data cleaning enhances not only accuracy on test data but also reduces overfitting, leading to better generalization on unseen data. For example, models trained on clean datasets tend to require less regularization and converge faster, which translates to better performance in deployment settings.

V. Conclusion

Data cleaning has a profound impact on machine learning accuracy, but its effects are nuanced. The surveyed literature consistently shows that improving data quality – when done strategically – enhances model performance, yet indiscriminate cleaning is not always warranted. The emerging "data-centric AI" viewpoint calls for tighter integration of cleaning with modeling. Moving forward, researchers and practitioners should emphasize automated, ML-guided cleaning workflows and evidence-based priorities (e.g., focus on label and feature accuracy). Future work, as noted by Côté et al., should develop better end-to-end solutions and address gaps such as continuous data streams, fairness implications of cleaning, and adaptive cleaning methods. As data volumes grow, investing in intelligent cleaning processes will be crucial for reliable, accurate ML. This review underscores that, when balanced against cost, carefully targeted data cleaning remains a powerful lever for improving ML outcomes and should be pursued alongside algorithmic advances.

References

[1] Li, P., Rao, X., Blase, J., Zhang, Y., Chu, X., & Zhang, C. (2021). CleanML: A Study for Evaluating the Impact of Data Cleaning on ML Classification Tasks. In Proc. IEEE Int'l Conf. Data Engineering (ICDE).

[2] Côté, P.-O., Nikanjam, A., Ahmed, N., Humeniuk, D., & Khomh, F. (2024). Data cleaning and machine learning: A systematic literature review. Automated Software Engineering, 31, article 54. DOI:10.1007/s10515-024-00453-w.

[3] Budach, L., Feuerpfeil, M., Mohammed, S., Ihde, N., Nathansen, A., Noack, N., Patzlaff, H., Naumann, F., & Harmouch, H. (2025). The effects of data quality on machine learning performance on tabular data. Information Systems, 118, 103836.

[4] Mohammed, S., Naumann, F., & Harmouch, H. (2025). Step-by-Step Data Cleaning Recommendations to Improve ML Prediction Accuracy. In Proc. 28th Int'l Conf. Extending Database Technology (EDBT), 540–553.

[5] Goldberger, J., Rigollet, P., Lee, J. D., & Papalexakis, E. E. (2016). ActiveClean: Interactive Data Cleaning for Predictive Modeling. Proc. VLDB Endow., 9(12), 948–959. DOI:10.14778/2994509.2994514.

