International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 09 | Sep 2026 www.irjet.net p-ISSN: 2395-0072

When Does Data Quality Matter More Than Model-Side Improvement? A Controlled Comparative Study of Data Quality and Machine Learning Performance

Dr. Pranjali Patil[1], Ritesh Pathak[2], Satya Rai[3], Rohit Kulkarni[4]

[1]*Assistant Professor, D. Y. Patil Deemed to be University, School of Humanities & Sciences, Belapur, Maharashtra, India.*
[2]*Student, Department of MCA, D. Y. Patil Deemed to be University, School of Humanities & Sciences, Belapur, Maharashtra, India.*
[3]*Student, Department of MCA, D. Y. Patil Deemed to be University, School of Humanities & Sciences, Belapur, Maharashtra, India.*
[4]*Student, Department of MCA, D. Y. Patil Deemed to be University, School of Humanities & Sciences, Belapur, Maharashtra, India.*

Abstract-*Machine-learning performance can be improved in two main ways: by fixing problems in the training data or by using a stronger model. This study compares these two choices and examines when data repair gives a larger improvement. Experiments were carried out on Bank Marketing, South German Credit, and Adult Income. Label noise and feature-value corruption were added to the training data at 5%, 10%, 20%, and 30% severity. For each degraded condition, 50% of the added defects were restored and Logistic Regression was trained again. This improvement was compared with Random Forest and XGBoost trained on the same degraded data. Two stratified train/test splits and three corruption seeds produced 576 primary model-condition results. ROC-AUC was used as the main evaluation metric. Intervention Advantage was defined as the improvement from data repair minus the improvement from model replacement. Data degradation reduced performance, while partial repair improved performance in every dataset-defect combination. However, one method was not always better. South German Credit strongly favored repair under label noise, Bank Marketing moved toward repair at high degradation when compared with Random Forest, and Adult Income generally favored model-side improvement, especially XGBoost. Therefore, the better choice depends on the dataset, type of defect, severity, and model being considered.*

Key Words: Data-centric AI, Data Quality, Label Noise, Feature Corruption, Model Selection, XGBoost, Intervention Advantage

1. INTRODUCTION

Machine-learning work often starts by focusing on the model: choosing an algorithm, tuning it, and increasing its complexity when performance is poor. However, poor performance can also come from problems in the training data. Data-centric AI focuses on improving the dataset throughout the machine-learning process [1]. Studies of real AI projects also show that unresolved data problems can continue through later stages of a system and affect its results [2]. Therefore, data quality should not be treated only as a preprocessing step; it can directly affect model performance.

This creates an important practical question. If the training data have quality problems and there is limited time or effort available, should we repair the data or keep the data as it is and use a stronger model? Previous research shows that data cleaning can improve prediction, but the amount of improvement depends on the error type, dataset, cleaning method, and model [3]. Other studies have directly compared data-centric and model-centric approaches and found cases where improving the data performs better [4]. At the same time, experiments on tabular data show that different algorithms react differently to different data-quality problems [5].

This study compares these two choices under the same controlled conditions. Label errors and feature-value errors are deliberately added to the training data at four severity levels. For the data-side approach, 50% of the known added defects are repaired while Logistic Regression (LR) is kept as the model. For the model-side approach, the degraded data are kept unchanged and LR is replaced with Random Forest (RF) or XGBoost. Both approaches are compared from the same degraded LR starting point.

The main contributions of this study are: (i) a direct comparison of data repair and model improvement from the same degraded baseline; (ii) Intervention Advantage (IA), which measures the difference between the improvement from the two approaches; (iii) testing across two defect types and four severity levels on three tabular datasets; and (iv) identifying situations where the better choice changes between data repair and model improvement.

2. RELATED WORK AND RESEARCH GAP

CleanML systematically evaluated 14 real datasets, five error types, seven ML models, and multiple cleaning methods while controlling experimental randomness and false discoveries [3]. Earlier systems such as ActiveClean [6] and BoostClean [7] connected cleaning decisions to downstream predictive performance. COMET frames cleaning as a resource-allocation problem by recommending which feature to clean next under a limited cleaning budget [8]. Together, these studies establish that cleaning value is downstream-task dependent.

Label quality is especially important because incorrect targets directly alter the supervised learning signal. Northcutt, Athalye, and Mueller [9] identified widespread label errors in benchmark test sets and showed that such errors can destabilize model rankings. This supports treating label corruption separately from feature corruption.

Bhatt et al. [4] explicitly compared data-centric and model-centric approaches using a common ResNet-18 framework on MNIST, Fashion-MNIST, and CIFAR-10. Tschalzev et al. [10] showed that dataset-specific preprocessing and feature engineering can alter model rankings in tabular evaluation. A 2025 study examined six data-quality dimensions and 19 ML algorithms and concluded that quality effects vary with dimension, algorithm, and pollution location [5]. A recent systematic review similarly reports heterogeneous comparative outcomes [11].

The research gap is therefore more specific than simply asking whether clean data help. Existing studies do not fully show, under the same degraded training condition, whether repairing a fixed part of known defects gives more improvement than keeping those defects and switching to a stronger model. They also do not fully show how this comparison changes with the type and severity of the data problem.

3. RESEARCH QUESTIONS AND HYPOTHESES

RQ1: Under what combinations of data-quality defect type, degradation severity, and dataset does repairing degraded training data provide greater predictive improvement than switching to a stronger model? RQ2: How does the relative advantage change as degradation severity increases? RQ3: Does the comparison differ between label errors and feature-value errors? RQ4: Are intervention effects consistent across datasets?

H1 predicts that increasing degradation reduces predictive performance relative to clean training conditions. H2 predicts that repairing 50% of injected defects improves the degraded LR baseline. H3 predicts a severity interaction, with repair becoming more competitive at sufficiently high degradation in at least some conditions. H4 predicts different intervention behavior for label and feature corruption. H5 predicts dataset-dependent rather than universal intervention rankings.

4. METHODOLOGY

4.1 Datasets

Three UCI tabular binary-classification datasets were used. Bank Marketing contains 45,211 observations and concerns term-deposit subscription [12]; the call-duration variable was excluded because it is only known after the call. South German Credit contains 1,000 observations with 20 predictor variables and corrects coding problems in the older Statlog representation [13]. Adult Income contains 48,842 observations with 14 predictors and naturally occurring missing values [14].

Table -1: DATASETS USED IN THE CONFIRMATORY STUDY

Dataset	Instances	Predictors	Characteristic
Bank Marketing	45,211	15 used	Large, imbalanced
South German Credit	1,000	20	Small credit-risk
Adult Income	48,842	14	Large mixed-type

4.2 Preprocessing and Models

Each dataset was split into 80% training and 20% test data using stratification with split seeds 2026 and 31415. Numerical missing values were median-imputed and standardized; categorical missing values were mode-imputed and one-hot encoded with unseen-category handling. Preprocessing was fitted only on the relevant training condition and applied to its unchanged test set. The baseline learner was class-weighted LR; model-side alternatives were RF and XGBoost. Configurations were fixed across degradation levels.

Table -2: FIXED MODEL CONFIGURATIONS

Model	Configuration
LR	liblinear; balanced classes; max_iter=300
RF	15 trees; depth=12; leaf=3; max_samples=0.75
XGBoost	25 trees; depth=4; learning_rate=0.10; subsample=0.80

4.3 Controlled Degradation and Repair

Two primary defect types were injected only into the training data at 5%, 10%, 20%, and 30% severity. For label noise, selected binary labels were flipped. For feature corruption, selected observed numerical cells were perturbed by Gaussian noise with standard deviation equal to 1.5 times the feature training standard deviation. This corruption was added directly to the original numerical feature values before standardization. Selected categorical cells were replaced by another valid category from the same feature.

Because the experiment itself creates the defects, their locations and original values are known. Therefore, the data-side approach uses oracle-controlled repair: 50% of the added defects are randomly selected and restored to their original values. This measures the benefit of partial repair rather than the ability of an algorithm to detect errors. Three corruption seeds (101, 202, and 303) were used for each split, severity level, and defect type. The test data were never deliberately corrupted.

4.4 Experimental Conditions and Intervention Advantage

For every dataset, defect type, severity, split, and corruption seed, four conditions were evaluated: degraded data + LR; 50% repaired data + LR; degraded data + RF; and degraded data + XGBoost. The full primary design generated 576 model-condition results (3 datasets × 2 defect types × 4 severity levels × 2 splits × 3 corruption seeds × 4 model conditions).

Let P_B denote ROC-AUC for LR on degraded data, P_D the performance after 50% repair using LR, and P_M the performance of a model-side alternative M on the same degraded data. Data gain is Delta_D = P_D - P_B and model gain is Delta_M = P_M - P_B. Intervention Advantage is IA_M = Delta_D - Delta_M = P_D - P_M. IA > 0 favors repair; IA < 0 favors model replacement.

4.5 Metrics and Statistical Analysis

ROC-AUC was the primary metric, with F1-score and balanced accuracy as secondary metrics. Paired Wilcoxon signed-rank tests were used because interventions were evaluated on matched split/corruption conditions. Bootstrap 95% confidence intervals were computed for severity-specific IA estimates. Holm adjustment was used for families of multiple comparisons. Interpretation emphasizes effect direction, magnitude, and uncertainty.

5. RESULTS

5.1 Clean Baseline and Effect of Degradation

Table -3: CLEAN LOGISTIC REGRESSION BASELINE

Dataset	ROC-AUC	F1	Balanced Acc.
Adult Income	0.9062	0.6797	0.8222
Bank Marketing	0.7721	0.3835	0.7062
South German Credit	0.7917	0.6063	0.7196

The clean LR baseline ranged from approximately 0.7721 ROC-AUC on Bank Marketing to 0.9062 on Adult Income. H1 was supported across all six dataset-defect combinations. Mean ROC-AUC degradation loss ranged from 0.0052 for Bank feature corruption to 0.0500 for South German label noise, with one-sided paired Wilcoxon tests indicating degradation loss in every combination ($p < 0.001$).

5.2 Effect of Partial Data Repair

Table -4: DEGRADATION LOSS AND 50% REPAIR GAIN

Dataset	Defect	Loss	Repair gain
Adult Income	Feature corruption	0.0107	0.0044
Adult Income	Label noise	0.0096	0.0058
Bank Marketing	Feature corruption	0.0052	0.0042
Bank Marketing	Label noise	0.0108	0.0066
South German Credit	Feature corruption	0.0104	0.0090
South German Credit	Label noise	0.0500	0.0348

H2 was supported. Repairing 50% of injected defects produced positive mean ROC-AUC recovery in every dataset-defect combination, with one-sided paired Wilcoxon p-values below 0.001. The largest average repair gain occurred for South German Credit under label noise (+0.0348 ROC-AUC).

5.3 Data Repair versus Model-Side Improvement

Table -5: MEAN INTERVENTION ADVANTAGE (ROC-AUC)

Dataset	Defect	IA vs RF	IA vs XGB	Overall
Adult	Feature	-0.0010	-0.0035	Models
Adult	Label	-0.0025	-0.0083	Models
Bank	Feature	-0.0085	-0.0072	Models
Bank	Label	+0.0077	-0.0120	Mixed
South German	Feature	+0.0299	-0.0059	Mixed
South German	Label	+0.0457	+0.0231	Repair

The main result is that there was no single best approach for every condition. South German Credit showed the clearest advantage for data repair. Under label noise, mean IA was +0.0457 against RF and +0.0231 against XGBoost. The comparison against RF remained significant after Holm correction (adjusted p = 0.00020). The comparison against XGBoost was also positive, but it was not significant after correction (adjusted p = 0.137). Under feature corruption, repair performed much better than RF (+0.0299; adjusted p < 0.001), while the average comparison with XGBoost was slightly negative (-0.0059).

Adult Income generally benefited more from model-side improvement, especially XGBoost. Mean IA against XGBoost was -0.0083 under label noise and -0.0035 under feature corruption, and both comparisons remained significant after Holm adjustment. Bank Marketing showed a mixed result. Model-side improvement was better on average for feature corruption. Under label noise, repair performed better than RF on average (+0.0077), but XGBoost still had an overall advantage (-0.0120).

5.4 Severity and Intervention Switching

The severity of degradation also changed which intervention was more useful, supporting H3. At 30% label noise, IA against RF reached +0.0403 for Bank Marketing and +0.0953 for South German Credit. South German Credit also favored repair over XGBoost at 30% label noise (+0.0848; bootstrap 95% CI approximately 0.0658 to 0.1037). Adult Income was almost equal against RF (+0.0012) but still favored XGBoost (-0.0094).

Feature corruption showed a weaker change. At 30% severity, Bank Marketing slightly favored repair against RF (+0.0017) and XGBoost (+0.0027). South German Credit strongly favored repair against RF (+0.0570) but was nearly equal against XGBoost (+0.0035). This means there is no single degradation level at which repair always becomes the better option.

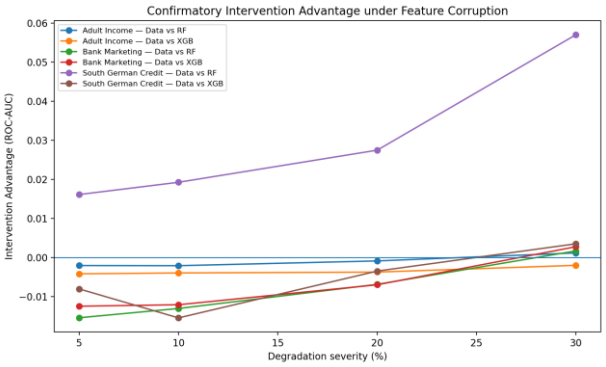

Fig -1: Confirmatory Intervention Advantage under label noise.

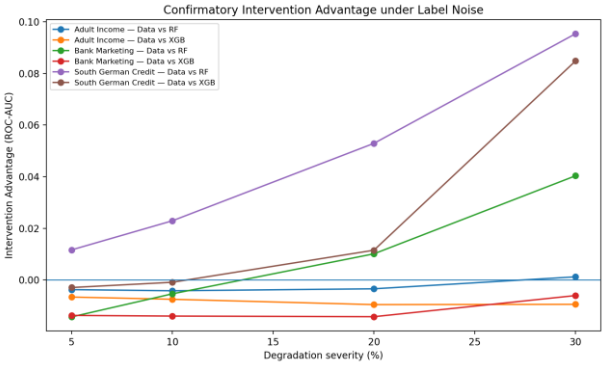

Fig -2: Confirmatory Intervention Advantage under feature corruption.

5.5 Hypothesis Summary

Table -6: HYPOTHESIS OUTCOMES

Hypothesis	Outcome	Summary
H1	Supported	Degradation reduced ROC-AUC in all six combinations.
H2	Supported	50% repair produced positive mean recovery in all six.
H3	Supported	Repair became more competitive with severity in several conditions.
H4	Supported	Label and feature corruption differed.
H5	Supported	Rankings differed materially by dataset.

6. DISCUSSION

The results agree with earlier research showing that the effect of data cleaning depends on the dataset, error type, and model [3]. This experiment adds a direct comparison between two practical choices: repairing the data and changing the model. Both are measured from the same degraded LR baseline. IA therefore helps answer a simple question: does repairing part of the known data problems give more improvement than switching to a particular stronger model?

South German Credit showed the strongest benefit from repair. One possible reason is that it is a relatively small dataset, so incorrect labels may have a larger effect on the useful training information. At 30% label noise, repairing half of the added errors gave much more improvement than changing the learning algorithm. Adult Income has many more observations, and XGBoost remained strong enough to perform better than the standard repair approach across the tested label-noise levels. This is a possible explanation of the result, but it cannot be treated as a general cause based on only three datasets.

Bank Marketing shows why the model used for comparison is also important. As label noise became more severe, repair performed increasingly better than RF, but XGBoost remained stronger on average. Therefore, it is not enough to simply say 'clean the data instead of changing the model.' The answer also depends on which alternative model is available.

In practical use, these results suggest that the condition of the data should be checked before focusing only on making the model more complex. When degradation is mild and a strong model is available, changing the model may be enough. When known data problems become more severe, especially label errors, repairing part of the data can become more useful and may give a larger improvement than model replacement.

7. LIMITATIONS AND THREATS TO VALIDITY

This study has some limitations. Oracle repair assumes that the locations of the corrupted values and their original values are already known. Therefore, it measures the benefit of fixing errors, not the cost or difficulty of finding those errors in real applications. Only a 50% repair level was tested, and the synthetic corruption used here cannot represent every real-world data problem. The study also used only three tabular binary-classification datasets and three model families. Two outer splits and three corruption seeds were used, so very small IA differences should be interpreted carefully. The model settings were fixed and were not heavily optimized. More extensive model tuning could change the size of the model-side improvements. Finally, the study compares prediction improvement, not the actual financial, human, or computational cost of each approach.

8. CONCLUSION

This study examined when repairing degraded training data gives more predictive improvement than switching from Logistic Regression to Random Forest or XGBoost. Across 576 primary model-condition results, increasing degradation reduced performance, while repairing 50% of the added defects consistently improved ROC-AUC. However, data repair was not always better. South German Credit strongly favored repair under label noise. Bank Marketing showed changes with severity when compared with Random Forest, while XGBoost remained competitive. Adult Income generally benefited more from model-side improvement. Feature corruption also showed a different pattern from label corruption.

The study does not support a simple rule that we should always clean the data or always change the model. The better choice changed with the dataset, type and severity of corruption, and the alternative model being considered. Intervention Advantage makes this comparison clear by measuring both choices from the same degraded baseline. Future studies should include realistic costs for finding and cleaning errors, different repair levels, more datasets and models, and data problems taken from real application areas.

ACKNOWLEDGEMENT

The authors sincerely thank Dr. Pranjali Patil, Assistant Professor, D. Y. Patil Deemed to be University, School of Humanities & Sciences, for her guidance, academic support, and valuable suggestions throughout this research work.

REFERENCES

[1] D. Zha, Z. P. Bhat, K.-H. Lai, F. Yang, Z. Jiang, S. Zhong, and X. Hu, "Data-centric Artificial Intelligence: A Survey," arXiv:2303.10158, 2023.

[2] N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. K. Paritosh, and L. Aroyo, "Everyone wants to do the model work, not the data work: Data Cascades in High-Stakes AI," in Proc. CHI Conference on Human Factors in Computing Systems, 2021, pp. 1-15, doi: 10.1145/3411764.3445518.

[3] P. Li, X. Rao, J. Blase, Y. Zhang, X. Chu, and C. Zhang, "CleanML: A Study for Evaluating the Impact of Data Cleaning on ML Classification Tasks," in Proc. 2021 IEEE 37th International Conference on Data Engineering (ICDE), pp. 13-24, 2021, doi: 10.1109/ICDE51399.2021.00009.

[4] N. Bhatt, N. Bhatt, P. Prajapati, V. Sorathiya, S. Alshathri, and W. El-Shafai, "A Data-Centric Approach to improve performance of deep learning models," Scientific Reports, vol. 14, Art. no. 22329, 2024, doi: 10.1038/s41598-024-73643-x.

[5] S. Mohammed, L. Budach, M. Feuerpfeil, N. Ihde, A. Nathansen, N. Noack, H. Patzlaff, F. Naumann, and H. Harmouch, "The effects of data quality on machine learning performance on tabular data," Information Systems, vol. 132, Art. no. 102549, 2025, doi: 10.1016/j.is.2025.102549.

[6] S. Krishnan, J. Wang, E. Wu, M. J. Franklin, and K. Goldberg, "ActiveClean: Interactive Data Cleaning For Statistical Modeling," Proceedings of the VLDB Endowment, vol. 9, no. 12, pp. 948-959, 2016, doi: 10.14778/2994509.2994514.

[7] S. Krishnan, M. J. Franklin, K. Goldberg, and E. Wu, "BoostClean: Automated Error Detection and Repair for Machine Learning," CoRR, abs/1711.01299, 2017.

[8] S. Mohammed, F. Naumann, and H. Harmouch, "Step-by-Step Data Cleaning Recommendations to Improve ML Prediction Accuracy," in Proc. 28th International Conference on Extending Database Technology (EDBT), pp. 542-554, 2025, doi: 10.48786/edbt.2025.43.

[9] C. G. Northcutt, A. Athalye, and J. Mueller, "Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks," NeurIPS Datasets and Benchmarks, 2021.

[10] A. Tschalzev, S. Marton, S. Lüdtke, C. Bartelt, and H. Stuckenschmidt, "A Data-Centric Perspective on Evaluating Machine Learning Models for Tabular Data," in Advances in Neural Information Processing Systems 37, Datasets and Benchmarks Track, 2024, doi: 10.52202/079017-3039.

[11] A. Oludele, N.-O. Aisosa, A. E. Emmanuel, S. B. Nneamaka, A. Ifeoluwa, P. Iyogun, A. Y. Adisa, et al., "Data-Centric Versus Algorithm-Centric Machine Learning Approaches: A Systematic Review of Comparative Effectiveness and Evaluation Frameworks," Asian Journal of Research in Computer Science, vol. 19, no. 6, pp. 88-105, 2026, doi: 10.9734/ajrcos/2026/v19i6871.

[12] S. Moro, P. Rita, and P. Cortez, "Bank Marketing" [Dataset], UCI Machine Learning Repository, 2014, doi: 10.24432/C5K306.

[13] "South German Credit" [Dataset], UCI Machine Learning Repository, 2019, doi: 10.24432/C5X89F.

[14] B. Becker and R. Kohavi, "Adult" [Dataset], UCI Machine Learning Repository, 1996, doi: 10.24432/C5XW20.