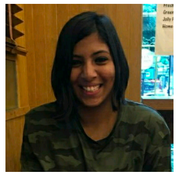
Automated Data Cleaning Can Hurt Fairness in ML-based Decision Making

Shubha Guha[*], Falaah Arif Khan[**], Julia Stoyanovich[**], Sebastian Schelter[*]
[*]*University of Amsterdam* [**]*New York University*

IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING

UNIVERSITY OF AMSTERDAM

NYU

Machine Learning in the Real World

- ML used in critical decision-making processes
- If left unchecked, can often **reproduce or even amplify pre-existing bias** in the data, leading to **unlawful discrimination**
- Ongoing efforts to mitigate with research on fairness and **responsible data management**

DOI:10.1145/3488717

Perspectives on the role and responsibility of the data-management research community in designing, developing, using, and overseeing automated decision systems.

BY JULIA STOYANOVICH, SERGE ABITEBOUL, BILL HOWE, H.V. JAGADISH, AND SEBASTIAN SCHELTER

Responsible Data Management

INCORPORATING ETHICS AND legal compliance into data-driven algorithmic systems has been attracting significant attention from the computing research community, most notably under the umbrella of fair[8] and interpretable[16] machine learning. While important, much of this work has been limited in scope to the "last mile" of data analysis and has disregarded both the *system's design, development, and use life cycle* (What are we automating and why? Is the system working as intended? Are there any unforeseen consequences post-deployment?) and the *data life cycle* (Where did the data come from? How long is it valid and appropriate?). In this article, we argue two points. First, the decisions we make during data collection and preparation profoundly impact the robustness, fairness, and interpretability of the systems we build. Second, our responsibility for the operation of these systems does not stop when they are deployed.

Example: Automated hiring systems. To make our discussion concrete, consider the use of predictive analytics in hiring. Automated hiring systems are seeing ever broader use and are as varied as the hiring practices themselves, ranging from resume screeners that claim to identify promising applicants[a] to video and voice analysis tools that facilitate the interview process[b] and game-based assessments that promise to surface personality traits indicative of future success.[c] Bogen and Rieke[2] describe the hiring process from the employer's point of view as a series of decisions that forms a funnel, with stages corresponding to

a https://www.crystalknows.com
b https://www.hirevue.com
c https://www.pymetrics.ai

64 COMMUNICATIONS OF THE ACM | JUNE 2022 | VOL. 65 | NO. 6

"MIT Researcher Exposing Bias in Facial Recognition Tech"
https://www.insurancejournal.com/news/national/2019/04/08/523153.htm

Stoyanovich et al.: "Responsible data management," Communications of the ACM, 2022

Data Quality & Fairness

- Production ML typically **requires automated cleaning techniques**
- Relationship between data quality & fairness unclear
- **Research gap**: research on **joint cleaning and learning focuses on prediction accuracy only**, while **research on fairness ignores low-quality data** or focuses on coverage only

Schelter et al.: “Fairprep: Promoting data to a first-class citizen in studies on fairness-enhancing interventions”, EDBT, 2019

Chen et al.: “Why is my classifier discriminatory?”, NeurIPS, 2018

Automated Data Cleaning & Fair Decision-Making

- **Goal**: to obtain **insights on the impact of data quality and automated data cleaning on fair decision-making**

- **Challenge**: **no clean ground truth data available** for datasets commonly used in fairness research, difficult to manually obtain such ground truth

- Research questions address two common stages of automated data cleaning:

  - Error detection stage (RQ1): **Does the incidence of data errors track demographic group membership in ML fairness datasets?**

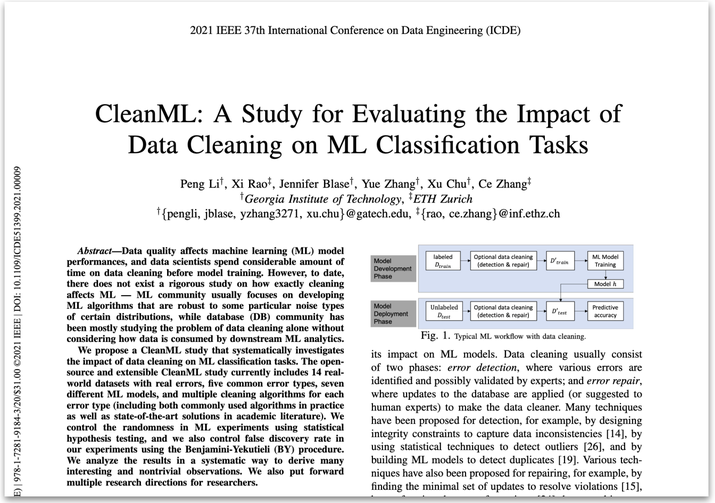
  - Data repair stage (RQ2): **Do common automated data cleaning techniques impact the fairness of ML models trained on the cleaned datasets?**

Datasets & Error Detection Strategies

- **Five benchmark datasets** commonly used in fairness research
- Datasets **partitioned into privileged group and disadvantaged group** based on sensitive demographic attributes
- **Common error detection strategies** from previous work on joint cleaning and learning
  - Missing values (NULL, NaN)
  - Outliers (stddev, IQR, isolation forest)
  - Label errors (cleanlab)

name	source	number of tuples	number of attributes	sensitive attribute(s)
adult	census	48,844	12	sex, race
folk	census	378,817	10	sex, race
credit	finance	150,000	8	age
german	finance	1,000	18	age
heart	healthcare	70,000	11	sex

TABLE I
BENCHMARK DATASETS USED IN ML FAIRNESS RESEARCH.



2021 IEEE 37th International Conference on Data Engineering (ICDE)

CleanML: A Study for Evaluating the Impact of Data Cleaning on ML Classification Tasks

Peng Li†, Xi Rao‡, Jennifer Blase†, Yue Zhang†, Xu Chu†, Ce Zhang‡
†*Georgia Institute of Technology*, ‡*ETH Zurich*
†{pengli, jblase, yzhang3271, xu.chu}@gatech.edu, ‡{rao, ce.zhang}@inf.ethz.ch

DOI: 10.1109/ICDE51399.2021.00009
©2021 IEEE | 978-1-7281-9184-3/20/$31.00

***Abstract*—Data quality affects machine learning (ML) model performances, and data scientists spend considerable amount of time on data cleaning before model training. However, to date, there does not exist a rigorous study on how exactly cleaning affects ML — ML community usually focuses on developing ML algorithms that are robust to some particular noise types of certain distributions, while database (DB) community has been mostly studying the problem of data cleaning alone without considering how data is consumed by downstream ML analytics.**

We propose a CleanML study that systematically investigates the impact of data cleaning on ML classification tasks. The open-source and extensible CleanML study currently includes 14 real-world datasets with real errors, five common error types, seven different ML models, and multiple cleaning algorithms for each error type (including both commonly used algorithms in practice as well as state-of-the-art solutions in academic literature). We control the randomness in ML experiments using statistical hypothesis testing, and we also control false discovery rate in our experiments using the Benjamini-Yekutieli (BY) procedure. We analyze the results in a systematic way to derive many interesting and nontrivial observations. We also put forward multiple research directions for researchers.

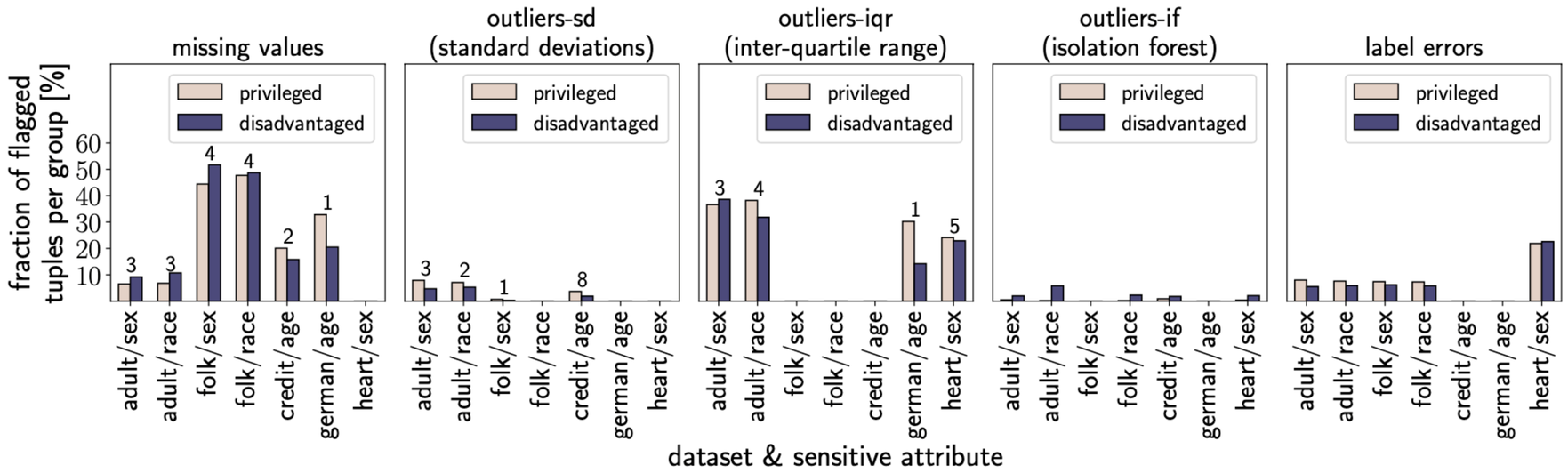
Fig. 1. Typical ML workflow with data cleaning.

its impact on ML models. Data cleaning usually consist of two phases: *error detection*, where various errors are identified and possibly validated by experts; and *error repair*, where updates to the database are applied (or suggested to human experts) to make the data cleaner. Many techniques have been proposed for detection, for example, by designing integrity constraints to capture data inconsistencies [14], by using statistical techniques to detect outliers [26], and by building ML models to detect duplicates [19]. Various techniques have also been proposed for repairing, for example, by finding the minimal set of updates to resolve violations [15],

Li et al.: “CleanML: A Study for Evaluating the Impact of Data Cleaning on ML Classification Tasks,” ICDE, 2019

RQ1: Incidence of Demographically Disparate Data Errors

- Compared **fractions of tuples flagged by common error detection strategies** for privileged and disadvantaged groups
- Found **higher fraction of tuples with missing values for disadvantaged groups** (in 14 out of 17 attributes)
- **No clear evidence for disparity** in other error types



RQ2: Impact of Automated Data Cleaning on Fairness

- Experimental study adapted from CleanML benchmark
- Measured **impact on accuracy and fairness** of several hundred cleaning configurations over “dirty” baselines, trained and evaluated **26,400 models in total**
- Generated cleaning configurations from:
  - **5 datasets** with corresponding **sensitive attributes, 3 ML models, 5 error detection strategies** and corresponding **repair methods** (mean/mode/dummy imputation, flipping labels)
- Trained **100 models per configuration** (20 train/test splits, 5 random seeds for hyperparameter search)
- Evaluated on accuracy and **2 fairness metrics** (predictive parity and equal opportunity)

Experimentation Framework

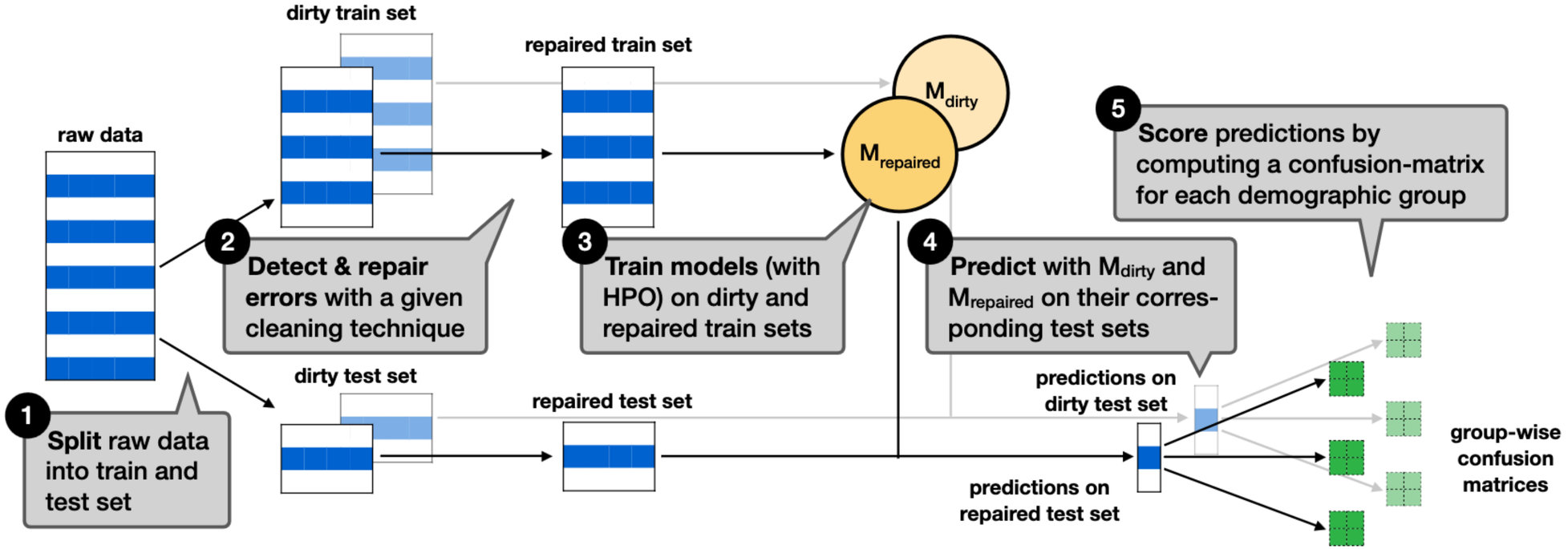


Fig. 3. Overview of our experimentation framework. For each experimental configuration (dataset/model/error/repair), we ❶ split the dataset into train/test sets; ❷ save the original raw data as a dirty version and apply the repair strategy to the raw data to generate a repaired version; ❸ train a classifier on the dirty train data and another classifier on the repaired train data; ❹ generate predictions on the dirty test set using the classifier trained on dirty data and predictions on the repaired test set using the classifier trained on the repaired train data; and ❺ score each model on accuracy and fairness and compare the scores computed from repaired data with the scores computed from dirty data to assess the impact of auto-cleaning for this configuration.

Findings on the Impact of Auto-Cleaning

- **Most of the time**: non-negative impact on accuracy and insignificant impact on fairness
- Worrying finding: **in cases where auto-cleaning impacts fairness, this impact is more likely to be negative than positive**
- **Example - auto-cleaning label errors**: strong positive impact on accuracy across all configurations, fairness impact highly dependent on chosen fairness metric
- **More details and findings in the paper** (including experimental results for intersectional group definitions)

TABLE III
IMPACT OF AUTO-CLEANING MISSING VALUES FOR SINGLE-ATTRIBUTE GROUPS, WITH EQUAL OPPORTUNITY AS FAIRNESS METRIC.

		accuracy			
		worse	**insignificant**	**better**	
fair.	**worse**	1.9% (2)	15.7% (17)	19.4% (21)	37.0% (40)
	insign.	9.3% (10)	25.9% (28)	13.0% (14)	48.1% (52)
	better	1.9% (2)	1.9% (2)	11.1% (12)	14.8% (16)
		13.0% (14)	43.5% (47)	43.5% (47)	

TABLE IV
IMPACT OF AUTO-CLEANING MISSING VALUES FOR SINGLE-ATTRIBUTE GROUPS, WITH DEMOGRAPHIC PARITY AS FAIRNESS METRIC.

		accuracy			
		worse	**insignificant**	**better**	
fair.	**worse**	3.7% (4)	13.0% (14)	19.4% (21)	36.1% (39)
	insign.	9.3% (10)	12.0% (13)	18.5% (20)	39.8% (43)
	better	0.0% (0)	18.5% (20)	5.6% (6)	24.1% (26)
		13.0% (14)	43.5% (47)	43.5% (47)	

TABLE XIV
IMPACT OF AUTO-CLEANING LABEL ERRORS FOR SINGLE-ATTRIBUTE GROUPS, WITH PREDICTIVE PARITY AS FAIRNESS METRIC.

		accuracy			
		worse	**insignificant**	**better**	
fair.	**worse**	14.3% (3)	14.3% (3)	19.0% (4)	47.6% (10)
	insign.	9.5% (2)	0.0% (0)	9.5% (2)	19.0% (4)
	better	0.0% (0)	0.0% (0)	33.3% (7)	33.3% (7)
		23.8% (5)	14.3% (3)	61.9% (13)	

TABLE XV
IMPACT OF AUTO-CLEANING LABEL ERRORS FOR SINGLE-ATTRIBUTE GROUPS, WITH EQUAL OPPORTUNITY AS FAIRNESS METRIC.

		accuracy			
		worse	**insignificant**	**better**	
fair.	**worse**	0.0% (0)	4.8% (1)	0.0% (0)	4.8% (1)
	insign.	0.0% (0)	0.0% (0)	14.3% (3)	14.3% (3)
	better	23.8% (5)	9.5% (2)	47.6% (10)	81.0% (17)
		23.8% (5)	14.3% (3)	61.9% (13)	

Call To Action: Fairness-Aware Data Cleaning

- Need to **think holistically** about **disparities in data quality**, **disparities in the effectiveness of data cleaning methods**, and **impacts** of such disparities **on ML model performance for different demographic groups**

- Need to **support data scientists with principled methods for selecting appropriate cleaning procedures** (many configurations do not negatively impact the fairness of model predictions)

- **Open questions and research directions**

  - Obtain datasets with clean ground truth
  - Evaluate more advanced data cleaning techniques
  - Evaluate data from non-US sources

Thanks!

- Code and results available at:

  https://github.com/amsterdata/demodq




University of Amsterdam

NYU