Automated Data Error Cleaning Impact on Federated Learning Utility and Fairness

James Sudlow
INSA Lyon – CNRS – LIRIS
Lyon, France
james.sudlow@insa-lyon.fr

Baudouin Naline
INSA Lyon – CNRS – LIRIS
Lyon, France
baudouin.naline@insa-lyon.fr

Sara Bouchenak
INSA Lyon – CNRS – LIRIS
Lyon, France
sara.bouchenak@insa-lyon.fr

***Abstract*— Data cleaning is a critical process for reliable and effective Federated Learning (FL). While current FL data cleaning methods mainly focus on their impact on model accuracy, their effects on FL model fairness and overall utility remain largely unexplored. This paper presents the first comprehensive empirical study that investigates the impact of automated data cleaning for tabular data on both FL model utility *and* fairness. Our study encompasses 2,365 different FL workloads, involving 21 data cleaning methods, applied to five real-world datasets, all implemented within our extensible TITANIA framework, resulting in FL workload traces consisting of 1,764,750 records. The key finding of our analysis is twofold: (i) data cleaning actually influences both FL model utility and fairness, with disparities in impacts on different datasets; (ii) FL data cleaning is not robust to high data error rates or high non-IID data distributions, as is usually the case in FL. This opens interesting research directions for novel robust and multi-criteria FL data cleaning, where model fairness and quality are provided by design.**

***Index Terms*—Noisy Data; Data Cleaning; Federated Learning; Utility; Fairness**

I. INTRODUCTION

Federated Learning (FL) is a fundamental approach in modern machine learning for training models from distributed data sources, while preserving data privacy [1]. However, data quality remains one of the biggest challenges faced by machine learning and FL teams when acquiring distributed and heterogeneous training data [2], [3]. Indeed, erroneous, noisy, and mislabeled data may degrade model performance or even render part of the data unusable [4], [5]. Therefore, ensuring reliable FL models requires automated data cleaning techniques that are capable of correcting erroneous samples for high-quality training data.

Furthermore, FL model fairness emerges as a major concern, as imbalances in datasets during data collection and preparation may lead to biases and disparities across demographic groups [6], [7]. These data biases propagate through FL training, producing models that amplify the fairness gaps between demographic groups, resulting in racist, sexist, and non-ethical models. Consequently, fairness is a central aspect that must be considered in FL alongside high-quality data.

Extensive works have been conducted recently, on the one hand, on FL data cleaning [8]–[12], and on the other hand, on FL model fairness [6], [7], [13], [14], these two sets of works being addressed independently from each other. To the best of our knowledge, no prior work investigates how automated data cleaning impacts not only FL model quality and utility but also model fairness.

In this paper, we precisely address this problem through the first comprehensive empirical study examining how automated tabular data cleaning affects both FL model utility and fairness. Other types of data (*e.g.*, text, audio, image, video, etc.) have their own types of errors and specific methods for cleaning them, and are thus out of the scope of this study [15]–[19].

Our key contributions are summarized as follows:

- Our extensive experimental analysis encompasses 2,365 different FL workloads, involving five real-world datasets, three FL models, and 21 data cleaning methods.
- We propose TITANIA, an open-source and extensible framework, to ease the comparative evaluation of FL data cleaning methods. TITANIA implements several widely-used model utility and fairness metrics measured on both the server and client sides. Its modular design facilitates the integration of additional data cleaning methods, datasets, models, and noise injection.
- We release a set of collected FL workload traces consisting of 1,764,750 records, which is a valuable resource for future studies on reliable FL. The TITANIA framework and the collected FL workload traces are publicly available at: *https://anonymous.4open.science/r/titania-4A2C/*

Our results show that tabular data cleaning impacts both FL model utility and fairness, with the effects strongly dependent on the dataset and the data cleaning method. We also find that many cleaning methods are not robust under high data error rates, as well as with highly non-IID data, as is usually the case in FL.

The findings of our study open interesting research paths, such as the exploration of automatic identification of data cleaning techniques that are better suited to a dataset; and the investigation of a novel multi-criteria approach of FL data cleaning that accounts for FL model utility and fairness by design, as well as robustness to data non-IID-ness and high error rates.

The rest of the paper is organized as follows. We first introduce the background in §II followed by a presentation of the TITANIA framework in §III. We then outline the results of our empirical evaluation in §IV. Finally, the key findings and open research directions are discussed in §V.

II. Background

Federated Learning. A Federated Learning (FL) system consists of N clients, C_1 to C_N, and a server. Each client C_i has a local dataset D_i. In each FL round, the server selects the clients that will participate. Each selected client C_i trains a local model on its data D_i, and shares its local model updates θ_i with the server. The server then aggregates the client updates, for instance through weighted averaging [1], to produce a global FL model θ that will be shared with the clients.

Corrupt Data. Data quality is a major concern in machine learning and FL systems. Indeed, corrupt or irrelevant records in a dataset often result in very poor quality FL models, or worse, in the inability to train the models on the data. Examples of corrupted tabular data include *missing values* in data attributes. These missing values must be carefully handled before training the FL model. In some cases, statistical *outliers* may correspond to incorrect data and should be addressed before model training. When the data structure includes a label (*i.e.*, an output class) that will be predicted by a classification model, errors can also creep into these labels, often owing to human errors in data annotation.

Bias in Federated Learning. Some datasets include *sensitive attributes* that are usually defined by the law and correspond to attributes such as gender, age, race, and religion. Here, a *demographic group* is defined by a specific value associated with a sensitive attribute; for example, the group of men and the group of women with respect to the sensitive attribute of gender, or the group of elderly people and the group of young people with respect to the sensitive attribute of age. Furthermore, we define *positive outcome* and *negative outcome* as data samples whose output class values are positive or negative. For simplicity, we consider a binary ML classifier, although this can be easily generalized to non-binary classifiers. An example of such a case is employment data with the positive outcome being the group of people with a high salary, and the negative outcome the group of people with a low salary.

III. Titania Framework

A. Overview of Titania Framework

We propose Titania (federa**T**ed learn**I**ng da**T**a cle**AN**ing and b**IA**s), an extensible framework for the evaluation of the impact of data cleaning methods on FL model utility and fairness. Titania includes several datasets and models (§III-B), data cleaning methods (§III-C), and evaluation metrics (§III-D). Consequently, it produces various FL workload traces and statistical analyses (§III-E). Figure 1 shows the Titania pipeline. ❶ For each FL client's dataset D_i, and data cleaning method $\mathcal{CM}_j$, two variants of the dataset are considered: the raw unclean dataset D_i, and the clean dataset D_i^{clean} resulting from the application of $\mathcal{CM}_j$ on D_i. ❷ The clean dataset D_i^{clean} is used to train model θ_i^{clean}, whereas θ_i model is trained over the unclean dataset D_i to serve as a baseline. Then, ❸ both θ_i^{clean} and θ_i models are used to perform predictions and ❹ to measure the model utility and fairness metrics. ❺ Finally, all results are stored in experimental traces, along with the overall experimental statistics. The Titania software prototype and collected traces are publicly available at: *https://anonymous.4open.science/r/titania-4A2C/*.

TABLE I
Dataset characteristics

Dataset	Topic	#Records	#Attr.	Sensitive attr.	Label
KDD	Finance	299,285	41	*gender*, *race*, *age*	High/low income
ARS	Healthcare	75,128	9	*gender*	Lying/non-lying position
Heart	Healthcare	70,000	11	*gender*, *age*	Heart disease or not
Adult	Finance	48,842	15	*gender*, *race*, *age*	High/low income
MEPS	Healthcare	33,400	42	*gender*, *race*	Visited few times/a lot of times

B. Datasets and Models

Titania includes several well-known tabular datasets, as listed in Table I, that contain sensitive attributes, such as gender, race, and age. Adult and KDD contain census data and are used to predict whether an individual's annual income is high or low [20], [21]. ARS is a sensor-based dataset that contains human activities of elderly participants and aims to determine whether participants are lying or not [22]. Heart is a healthcare dataset containing patient measurements used to predict the presence of heart diseases [23]. MEPS is also a medical dataset that contains in-person data access to hospitals, and its purpose is to predict whether a person has used medical facilities a lot of times [24]. Furthermore, Titania includes three classification models to predict data outcomes: Logistic Regression (LR) [25], Support Vector Machine (SVM) [26], and Multi-Layer Perceptron (MLP) [27].

In Figure 2, we further analyze the demographic groups in the Adult dataset. We observe disparities between demographic groups, in particular, regarding the proportion of positive and negative outcomes. Here, the datas contain 67% of men *vs*. 33% of women, and there are almost three times more men than women who have positive outcomes (*i.e.*, have a high salary). These disparities also appear in the other datasets, as shown in [28].

C. Data Cleaning Methods

The Titania framework includes several data cleaning methods that fall into one of the following three categories depending on the type of corrupted data covered by the method: (i) missing values, (ii) outliers, or (iii) label errors, as detailed below.

Missing Value Cleaning Methods. Missing values are flagged during the data loading stage, and their imputation is performed in different ways [29]. For instance, if a missing value is a numerical attribute, the imputed value can be the most frequently occurring value (*i.e.*, the mode) in the set of values of that attribute. The imputed missing numerical value can also be the mean of the set of values for the attribute in question. In the case of a categorical missing value, data imputation is based on the mode value of the attribute. In an FL system, the calculation of mode or mean values is performed either independently by each FL client based on its local data, or based on the global statistical information of all FL clients'

Fig. 1. Overview of TITANIA

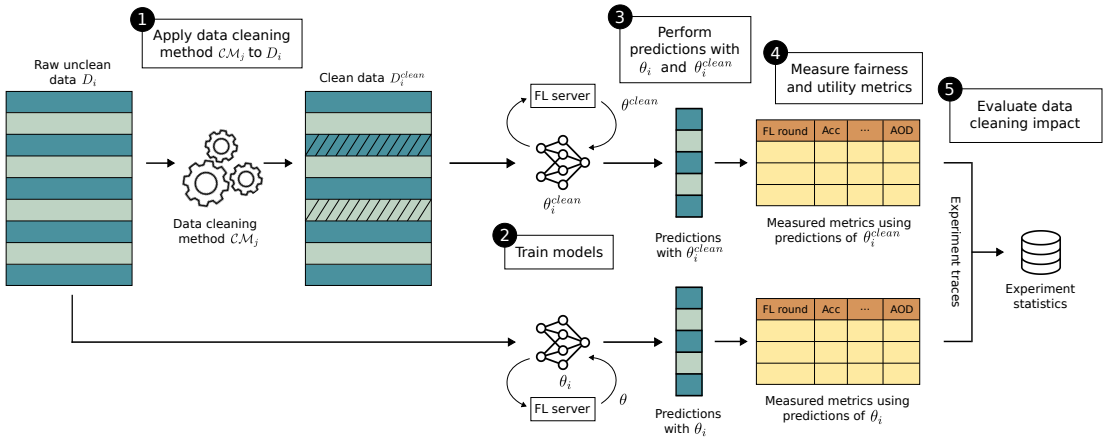
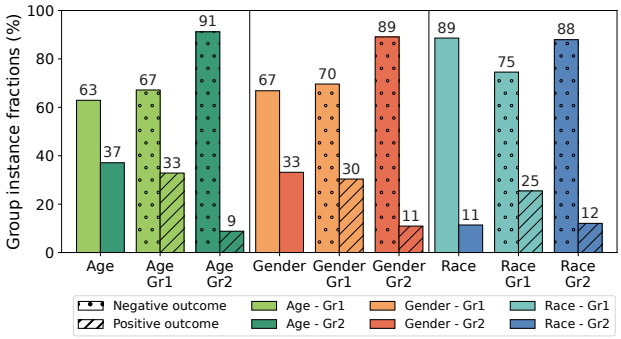


Fig. 2. Demographic disparities in Adult



data. In the latter case, the global mean or mode values are computed using a distributed protocol between the FL clients and server. In TITANIA, we implement missing value cleaning methods denoted as *mv-mn-l*, *mv-mn-g*, *mv-md-l*, and *mv-md-g*, where *mv* refers to missing value cleaning, and *mn*, *md*, *l*, and *g* represent the mean, mode, local, and global, respectively. Recent solutions also propose personalized FL approaches specifically targeting heterogeneous data imputation, such as the Cafe system [9] that we integrate into TITANIA and is referred to as *mv-cf*.

Outlier Cleaning Methods. Handling data outliers is another classical technique for cleaning the numerical data [29]. To detect outliers, it is necessary to determine the interval of values allowed for an attribute. For instance, the interquartile range (IQR) can be used to determine the interval of allowed values related to the 25^{th} and 75^{th} percentiles of the considered attributes, denoted p_{25} and p_{75}. Thus, with $iqr = p_{75} - p_{25}$, values outside the interval $[p_{25} - \gamma \cdot iqr, p_{75} + \gamma \cdot iqr]$ are considered outliers, with $\gamma \in \mathbb{R}^+$. Another method to determine the interval of allowed values for an attribute is through the standard deviation (STD) and mean value of that attribute. Values outside the interval $[mean - \mu \cdot std, mean + \mu \cdot std]$ are considered outliers, with $\mu \in \mathbb{R}^+$. Once an interval of allowed values is defined, there are several ways to handle outliers. For instance, an extreme approach is to exclude from the dataset the samples that contain outliers, but this has the risk of excluding too much data. Other techniques induce the replacement of outliers with median, mean, or mode values of the underlying attribute. Furthermore, as mentioned earlier, in an FL system, the calculation of data statistical values is performed either independently by each FL client based on its local data, or based on the global statistical information of all FL clients' data. Thus, TITANIA implements 12 outlier cleaning methods, denoted as *ol-iqr-mn-l*, *ol-iqr-mn-g*, *ol-iqr-md-l*, *ol-iqr-md-g*, *ol-iqr-dl-l*, *ol-iqr-dl-g*, *ol-std-mn-l*, *ol-std-mn-g*, *ol-std-md-l*, *ol-std-md-g*, *ol-std-dl-l*, and *ol-std-dl-g*, where *ol* refers to outlier cleaning, *iqr* and *std* correspond to IQR- and STD-based outlier detection techniques, *mn*, *md* and *dl* represent the mean, mode or deletion approaches, *l* and *g* represent the local and global, respectively.

Label Error Cleaning Methods. Data label errors are frequently introduced by human errors, particularly when the annotation process is tedious. Existing techniques to detect corrupt labels follow the same general approach, which consists of training a model on the data to learn noisy *vs.* correct labels [30]–[32]. For instance, a logistic regression model can be used to detect corrupt labels [30]. Thus, in a dataset with a binary outcome (*cf.* §III-B), correcting corrupt labels involves flipping their values. TITANIA implements this method for label cleaning, denoted as *le-cl*, as well as its variant *le-cl-dl* where samples are deleted if their labels are corrupted. In addition, recent works propose FL approaches specifically targeting heterogeneous data labels, such as the FedCorr system [8] that we integrate into TITANIA and is referred to as *le-fc*, as well as its variant *le-fc-dl* that deletes samples with label errors.

Other Approaches. In addition to data cleaning methods, there is another research path that consists in providing robust ML and FL, in order to learn models despite the existence of a fraction of corrupted data [33], [34]. These works are out of the scope of this study, that focuses on actual data cleaning methods.

D. Evaluation Metrics

The TITANIA framework considers several evaluation metrics to measure model performance to quantify different aspects of model utility and different aspects of model fairness.

Utility Metrics are used to predict the quality of the ML model. TITANIA considers four widely used utility metrics for ML classifiers, namely accuracy, precision, recall, and F1-score:

$$accuracy = (TP + TN)/(TP + TN + FP + FN)$$

$$precision = TP/(TP + FP)$$

$$recall = TP/(TP + FN)$$

$$F1 - score = 2 \cdot (\text{Precision} \cdot \text{Recall})/(\text{Precision} + \text{Recall})$$

where TP, FP, TN, and FN denote the true positives, false positives, true negatives, and false negatives, respectively.

Fairness Metrics are used to evaluate the impact of data cleaning methods on ML model fairness. TITANIA includes five commonly used fairness metrics, namely Statistical Parity Difference (SPD), Disparate Impact (DI), Discrimination Index (DcI), Equal Opportunity Difference (EOD), and Average Odds Difference (AOD). Each metric is measured for a demographic group related to a sensitive attribute S_j, as detailed below.

$$SPD_{S_j} = \left| P(\hat{Y} = 1 \mid S_j = v_1^j) - P(\hat{Y} = 1 \mid S_j = v_2^j) \right|$$

$$DI_{S_j} = P(\hat{Y} = 1 \mid S_j = v_1^j)/P(\hat{Y} = 1 \mid S_j = v_2^j)$$

$$EOD_{S_j} = \left| Recall_{S_j = v_1^j} - Recall_{S_j = v_2^j} \right|$$

$$DcI_{S_j} = \left| F1\text{-}score_{S_j = v_1^j} - F1\text{-}score_{S_j = v_2^j} \right|$$

$$AOD_{S_j} = \frac{1}{2} |(Recall_{S_j = v_1^j} - Recall_{S_j = v_2^j}) + (FPR_{v_1^j} - FPR_{v_2^j})|$$

where $FPR = FP/(FP + TN)$. Here, SPD and DI metrics aim to assess whether different demographic groups lead to different probabilities for model positive outcome $\hat{Y} = 1$, and measure either the absolute difference for SPD or the relative difference for DI between the probabilities of the demographic groups. EOD, DcI, and AOD metrics measure the difference in model quality between demographic groups, considering different quality metrics (*e.g.*, recall or F1-score).

E. FL Workload Traces and Statistical Analysis

TITANIA framework produces extensive FL workload traces containing various metrics, measured both globally and per FL client. A statistical analysis is then carried out for each utility and fairness metric to evaluate the impact of FL data cleaning on that metric. To account for randomness, each experimental scenario (cleaning method, dataset, model) is repeated several times, and at each time measurements are performed at several FL final stable rounds, which results in a total of κ measurements per metric. Then, for each metric, the set of κ values of the measurements resulting from the data cleaning method is compared with the set of κ values of the measurements resulting from the unclean baseline. To do so, we use *paired sample t-tests*, a rigorous statistical testing approach, to compare two sets of values [35], with a threshold for the p-value of 0.01. This allows us to determine whether the impact of data cleaning on the metric is positive (*i.e.*, it improves the metric), negative (*i.e.*, it deteriorates the metric), or insignificant (*i.e.*, no perceptible impact).

Based on the t-test results of all considered experimental scenarios, TITANIA produces overall statistics, that we will illustrate through the following simple example in Table II. Here, four experimental scenarios are considered, each one corresponding to a different combination of cleaning method/dataset/model. Two utility metrics and four fairness metrics are measured and reported for the global FL model (although TITANIA measures much more metrics, globally and per client). For instance, for the first experimental scenario, t-test results show that the considered cleaning method deteriorates accuracy and SPD$_{age}$, improves F1-score and SPD$_{race}$, and leaves AOD$_{age}$ and AOD$_{race}$ unaffected. Based on these results, Table III presents the overall statistics produced by TITANIA, by calculating the ratio of worse, insignificant or better cases for each metric, overall for all the experimental scenarios. These ratios can also be provided overall for all utility metrics, and overall for all fairness metrics, as reported in the last row of Table III. Furthermore, the statistics could also be produced overall for a type of fairness metric. For instance, in this example, the impact of cleaning methods on AOD (*i.e.*, AOD$_{age}$ and AOD$_{race}$) is worse in 50% of cases, insignificant in 50% of cases, and better in 0% of cases.

TABLE II
EXAMPLE OF T-TEST RESULTS

Exp. scenarios	Utility		Fairness			
			AOD		SPD	
	Accuracy	F1-score	AOD$_{age}$	AOD$_{race}$	SPD$_{age}$	SPD$_{race}$
exp1	worse	better	insig.	insig.	worse	better
exp2	worse	worse	insig.	worse	better	better
exp3	better	better	worse	insig.	worse	better
exp4	insig.	better	worse	worse	insig.	insig.

TABLE III
EXAMPLE OF OVERALL STATISTICS

Utility	worse	insig.	better
Accuracy	50%	25%	25%
F1-score	25%	0%	75%
Overall	*37.5%*	*12.5%*	*50%*

Fairness	worse	insig.	better
AOD$_{age}$	50%	50%	0%
AOD$_{race}$	50%	50%	0%
SPD$_{age}$	50%	25%	25%
SPD$_{race}$	0%	25%	75%
Overall	*37.5%*	*37.5%*	*25%*

IV. EXPERIMENTAL EVALUATION

In this section, we use TITANIA framework to conduct a comprehensive empirical study. In the following, we first describe our experimental setup, before presenting extensive results of the impact of all considered data cleaning methods, datasets and models in §IV-B-§IV-D, and analyzing the joint impact of data cleaning on both utility and fairness in §IV-E. Three use cases are added to the appendix to evaluate the robustness of FL data cleaning under various error rates in

§C, under heterogeneous FL data distributions in §B, and the benefits and limitations of the combination of FL data cleaning with bias mitigation in §D. Owing to space limitations, each of these three use cases is illustrated through one dataset. The full results of these use cases with all datasets, and their related plots can be found in [28].

A. Experimental Setup

The TITANIA framework is implemented using Python 3.13.5. Five datasets and three models presented in §III-B are considered in our experiments. As described in §III-C, TITANIA includes 12 outlier cleaning methods, 5 missing value cleaning methods (among which is the Cafe library [9]), and 4 label error cleaning methods (*i.e.*, the Cleanlab library [36], and the FedCorr library [8]). Outlier cleaning methods based on IQR have $\gamma = 1.5$, and those based on STD have $\mu = 3$ [37], [38]. We also consider a baseline consisting of raw unclean datasets for which the only modification was to remove records with missing values in KDD and Adult datasets. Handling missing values in the baseline is mandatory to allow model training. Sample deletion is frequently used when the dataset is large, or the percentage of deleted samples is small, and when the missingness is unrelated to output classes or to any other features in the dataset (which is the case for KDD and Adult) [39]. Furthermore, unless otherwise stated, all FL experiments are run with ten clients. Due to space limitation, we report on only results of globally computed FL metrics. The per client results can be found in [28]. To account for randomness, each experiment is repeated five times, and each time, all types of metrics are measured at each of the last 20 FL rounds, resulting in 100 paired measured metric for each t-test. All experiments are run on machines equipped with two Intel Xeon Gold 6130 CPUs (16 cores each) and 192 GB of RAM. In our empirical study, we collect extensive FL traces consisting of 2,365 workloads for 1,764,750 trace records. The TITANIA framework is publicly available, along with the collected FL workload traces and detailed hyper-parameters for reproducibility. All materials are available at [28].

B. Overall Impact of Data Cleaning

We first consider the impact of data cleaning on each utility metric and each fairness metric separately, considering all our experimental scenarios. Here, the objective is to analyze the impact of each considered cleaning method compared to the unclean baseline, by comparing the measurements of a given metric with and without cleaning through t-tests. This is first done separately for each experimental scenario (cleaning method/dataset/model), and each metric. Then for a given metric, we compute over all the results of all experimental scenarios the overall ratios of worse/better/insignificant results (*cf.* III-E). These results are presented in Table IV[1] Overall, the considered data cleaning methods worsen (resp. improve) fairness metrics in a non-negligible number of cases, that is up to 52% of cases with negative impacts (resp. up to 53% of cases with positive impacts). Regarding FL model utility, and as expected by data cleaning methods, the utility is improved in a non-negligible amount of cases (up to 40%). However, in up to 62% of cases, data cleaning also deteriorates utility, with F1-score being the most frequently impacted metric. This is counter-intuitive, and will be further investigated through finer-grained analysis in the following sections.

Observation 1: Although data cleaning may improve FL model utility, it may also deteriorate it in many cases.

Observation 2: Data cleaning may often have positive and negative impacts on FL model fairness.

TABLE IV
IMPACT OF DATA CLEANING ON UTILITY AND FAIRNESS METRICS [2]

Utility	worse	insig.	better
Accuracy	49%	11%	**40%**
Recall	61%	12%	27%
Precision	58%	12%	30%
F1-score	**62%**	10%	28%

Fairness	worse	insig.	better
AOD	47%	6%	47%
DcI	48%	7%	45%
EOD	47%	10%	43%
DI	**52%**	6%	42%
SPD	41%	6 %	**53%**

C. Deep Dive into Data Cleaning Methods

To better understand the previous observations, we conduct a deeper investigation of the impact of each data cleaning method on FL model utility and fairness. Table V presents a finer-grained view of the overall results of §IV-B, focusing separately on each data cleaning method. The results show that the negative impacts on model utility are generally more frequent with IQR-based outlier cleaning methods (between 65% and 77% of cases) than with STD-based outlier cleaning methods (between 42% and 67%). A similar trend is observed for the impact on FL model fairness, with more frequent negative impacts with IQR-based outlier cleaning methods (between 41% and 66% of cases) than with STD-based outlier cleaning methods (between 38% and 48%). Moreover, in up to 58% of cases, outlier cleaning methods may improve model fairness. STD-based methods assume that data approximately follows a normal distribution, whereas IQR-based methods are more suitable for highly skewed data [37], which explains their impact on the datasets considered in our study. It is worth noting that with the private and distributed nature of data in FL, the type of statistical distribution of the overall data is unknown by FL clients, which makes it difficult to choose the most appropriate technique (STD *vs.* IQR).

[1] AOD refers to the overall statistics of AOD_{gender}, AOD_{age}, and AOD_{race}, and the same applies for every category of fairness metric. Detailed numbers per fairness metric for each sensitive attribute are in [28].

[2] In addition to these results with t-test p-value of 0.01, similar analyses using p-values of 0.03 and 0.05 show consistent trends (available at [28].

TABLE V
IMPACT OF DIFFERENT DATA CLEANING METHODS ON FL MODEL FAIRNESS AND UTILITY

OUTLIER CLEANING METHODS

Outlier cleaning method	Utility			Fairness		
	worse	insig.	better	worse	insig.	better
ol-iqr-md-l	**77%**	10%	13%	57%	5%	38%
ol-iqr-md-g	**77%**	13%	10%	58%	4%	38%
ol-iqr-mn-l	68%	12%	20%	**66%**	6%	28%
ol-iqr-mn-g	65%	13%	22%	**66%**	7%	27%
ol-iqr-dl-l	76%	2%	22%	41%	1%	**58%**
ol-iqr-dl-g	76%	2%	22%	41%	2%	57%
ol-std-md-l	55%	23%	22%	47%	18%	35%
ol-std-md-g	53%	18%	**29%**	48%	11%	41%
ol-std-mn-l	42%	32%	27%	46%	16%	38%
ol-std-mn-g	45%	28%	27%	48%	14%	38%
ol-std-dl-l	65%	20%	15%	38%	6%	56%
ol-std-dl-g	67%	18%	15%	38%	5%	57%

MISSING VALUE CLEANING METHODS

Missing value cleaning method	Utility			Fairness		
	worse	insig.	better	worse	insig.	better
mv-md-l	63%	4%	33%	32%	10%	**58%**
mv-md-g	58%	0%	**42%**	45%	2%	53%
mv-mn-l	58%	0%	**42%**	38%	9%	53%
mv-mn-g	58%	0%	**42%**	45%	2%	53%
mv-cf	**79%**	4%	17%	**57%**	4%	39%

LABEL ERROR CLEANING METHODS

Label error cleaning method	Utility			Fairness		
	worse	insig.	better	worse	insig.	better
le-cl	0%	3%	**97%**	**49%**	8%	43%
le-cl-dl	0%	5%	95%	44%	7%	49%
le-fc	**72%**	0%	28%	39%	1%	**60%**
le-fc-dl	**72%**	0%	28%	39%	3%	58%

Observation 3: For cleaning data outliers in the considered datasets, it is better to favor methods based on STD than IQR, for fewer cases with negative impacts on FL model utility and fairness.

Regarding the cleaning of missing values, the different statistical methods (*i.e.*, *mv-md-l*, *mv-md-g*, *mv-mn-l*, and *mv-mn-g*) have a similar overall behavior, with a negative impact on model utility in at least half of cases and an improvement of utility in 33% to 42% of cases. These statistical methods also impact model fairness, more frequently positively (53% to 58% of cases) than negatively (32%-45%). In contrast, the *mv-cf* cleaning method, which specifically targets FL data heterogeneity, negatively impacts FL model utility and fairness more frequently than other statistical methods.

Finally, cleaning label errors, if applied locally to FL clients using a ML-based technique (*i.e.*, *le-cl*), is very effective for improving FL model utility in up to 97% of cases. The negative and positive impacts on model fairness are close, with 49% and 43% of cases, respectively. Whereas cleaning label errors using the considered FL-based technique (*i.e.*, *le-fc*) often deteriorates model utility (in 72% of cases), while often improving model fairness (in 60% of cases).

Overall, we observe that the considered data cleaning methods that are specifically designed for FL systems may deteriorate model utility more than other cleaning methods, *e.g.*, *mv-cf* and *le-fc* (and its variant *le-fc-dl*). These methods make strong assumptions or error distributions, *e.g.*, *le-fc* assumes that all unclean data is concentrated in some FL clients, whereas the other clients have clean data, and *le-fc* assumes that error distributions among clients are very heterogeneous. Such assumptions do not hold in all cases, and may explain the poor performance.

Observation 4: FL-based cleaning methods for missing values and label errors often make strong assumptions on error distributions which, if not applicable, become counterproductive in termes of FL model performance.

TABLE VI
IMPACT OF DATA CLEANING ON DIFFERENT DATASETS

Dataset	Utility			Fairness			Detected errors		
	worse	insig.	better	worse	insig.	better	outliers	miss. val.	lab. err.
Adult	**80%**	2%	18%	**55%**	5%	40%	✓	✓	✓
KDD	68%	1%	31%	38%	4%	**58%**	✓	✓	✓
Heart	56%	9%	35%	46%	5%	49%	✓	✗	✓
ARS	29%	48%	23%	43%	31%	26%	✓	✗	✓
MEPS	43%	4%	**53%**	53%	7%	40%	✓	✗	✓

D. Impact of Data Cleaning on Different Datasets

Table VI presents the overall impact of data cleaning on FL model utility and fairness, with a finer-grained view on each dataset separately. The results show that the different datasets present significant disparities in terms of both model utility and fairness. Regarding the impact of data cleaning on FL model utility, Adult and KDD are the most frequently deteriorated (with 80% and 68% of cases, respectively), while ARS mainly presents unaffected FL model utility (48% of cases). Moreover, data cleaning may improve model utility (in up to 53% of cases). Regarding the impact of data cleaning on FL model fairness, all datasets are negatively impacted in a non-negligible number of cases (*i.e.*, between 38% and 55% of cases). Furthermore, model fairness with all datasets may be improved by data cleaning (in 26% to 58% of cases). ARS is the least negatively impacted data in terms of utility. This can be explained by the fact that this dataset achieves a very high level of model utility with the unclean baseline (*i.e.*, 99%, *cf.* [28]) with, thus, a relatively limited impact of data cleaning on utility, which is not the case of the other datasets. Table VI also provides the types of errors detected by data cleaning, which allows us to derive the following observation.

Observation 5: The more a dataset includes several error types (*i.e.*, outliers, missing values, and label errors), the more their correction may induce more frequent FL model utility deterioration.

E. Joint Impact on Utility and Fairness

We now analyze cross-tabulated results considering the joint impact of cleaning methods on FL model utility fairness. First, for each experimental scenario (cleaning method, dataset, model), and for each metric, we use the t-test result produced by TITANIA (*i.e.*, worse, insignificant or better result). We then consider all possible pairs of metrics, one utility metric and one fairness metric, and determine their joint effect, *e.g.*, better for the former metric and worse for the latter, or worse for both metrics, etc. Finally, based on the results of all experimental scenarios, and all possible pairs of fairness-utility metrics, the overall ratios are computed, and presented in Table VII. We observe that in almost third of cases, both utility and fairness are deteriorated. This is explained by the fact that some fairness metrics depend on utility (*e.g.*, EOD, DCI in III-D), thus, an impact on the former would induce and impact on the latter.

TABLE VII
JOINT IMPACT OF DATA CLEANING ON FAIRNESS AND UTILITY

		Utility		
		worse	insig.	better
Fairness	worse	31%	2%	14%
	insig.	3%	2%	2%
	better	30%	2%	14%

V. LESSONS LEARNED & OPEN RESEARCH DIRECTIONS

We presented the first comprehensive empirical study that analyzes the impact of automated tabular data cleaning on FL model utility and fairness. Our extensible and open-source TITANIA framework includes 21 FL data cleaning methods applied to five widely used datasets, as well as a set of collected traces from 2,365 FL workloads, for 1,764,750 records. We believe that TITANIA's results will help researchers and practitioners explore novel FL data cleaning solutions. In summary, our empirical study identifies the following key takeaways:

Robustness of FL Data Cleaning Under Non-IID Settings. The more heterogeneous the data distributions of FL clients are, the less accurate and fair the FL model is. However, one of the key findings here is that the non-iid-ness in FL clients' data may amplify the deterioration of FL model fairness caused by data cleaning. Future research on FL data cleaning should carefully consider non-iid-ness and its negative impact on FL model fairness.

Robustness of FL Data Cleaning Under High Error Rates. Under low data error rates, the FL models resulting from the cleaned data may maintain a relatively stable utility and fairness. However, with high error rates, both FL model utility and fairness worsen drastically. Future research on FL data cleaning should investigate more robust solutions for highly erroneous FL datasets.

Towards a Multi-criteria Approach. Data cleaning often hurts FL model utility, and may have a positive and negative impact on FL model fairness. Furthermore, even when combining data cleaning with FL bias mitigation, this has a negative impact on model utility without fully achieving model fairness. This highlights the need to devise novel FL data cleaning methods that account for FL model utility and fairness by design, as well as robustness to data non-IID-ness and high error rates, in a multi-criteria approach.

Which FL Data Cleaning Techniques Better Suit a Dataset? Our comparative analysis of several data cleaning techniques applied to different datasets shows different behaviors of the cleaning techniques, as well as different impacts on different datasets. This opens interesting research paths to explore the automatic identification of data cleaning techniques that are most appropriate for a dataset based on its properties.

This work opens other interesting research perspectives, such as investigating existing FL data cleaning methods on other types of data and associated models (*e.g.*, text, audio, image, video, etc.), their specific types of errors, and related methods for cleaning them [15]–[19]. It could also be interesting to investigate the practicality of such comprehensive studies on large-scale datasets, by combining FL data cleaning methods to data subset selection methods [40], [41].

APPENDIX

A. Model Hyperparameters

TABLE VIII
EACH MODEL'S HYPERPARAMETERS WERE OPTIMIZED FOR THE UNDERLYING DATASET. BATCH SIZE IS 128 WITH ALL MODELS

Dataset	Model	Optimizer	Learning rate
Adult	LR	AdamW	0.0001
	MLP (64, 32)	RMSprop	0.0001
	SVM	Adam	0.001
ARS	LR	RMSprop	0.01
	MLP (64, 32)	Adagrad	0.01
	SVM	RMSprop	0.01
Heart	LR	RMSprop	0.001
	MLP (64, 32)	Adagrad	0.01
	SVM	RMSprop	0.03
KDD	LR	SGD	0.01
	MLP (64, 32)	SGD	0.003
	SVM	SGD	0.01
MEPS	LR	Adagrad	0.001
	MLP (64, 32)	Adagrad	0.001
	SVM	RMSprop	0.0003

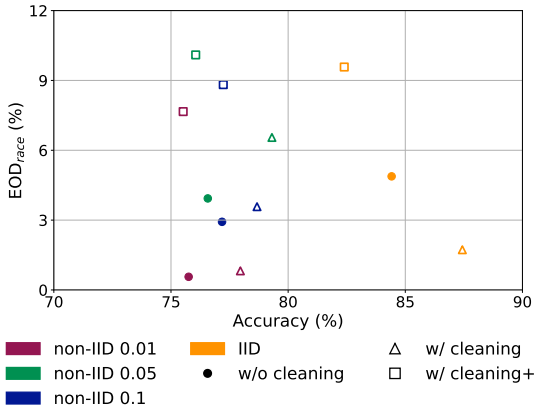
B. Data Cleaning Under Various FL Non-IID Settings

In the following, we explore the impact of data cleaning on various non-IID FL settings, where five FL clients have more or less heterogeneous statistical data distributions. We apply the classical approach for non-IID FL data using the Dirichlet function [42]–[44], to have a label distribution skew between FL clients. We consider three non-IID settings with three values of the Dirichlet α parameter: 0.01, 0.05, and 0.1. The lower α, the higher the non-IID-ness. We also include the IID setting. In all these settings, we consider three cases: one without data cleaning and two variants of data cleaning,

where data outliers, missing values, and label errors are corrected simultaneously. In the first variant of data cleaning (*w/ cleaning*), we apply the statistical methods *ol-std-md-l*, *mv-md-l*, and *le-cl*, and in the second variant (*w/ cleaning+*), we combine *ol-std-md-l* with *mv-cf* and *le-fc* methods that specifically target FL heterogeneous data. Figure 3 presents the results of Adult/LR and compares the different non-IID and IID settings in terms of FL model accuracy and fairness[2]. Overall, *w/ cleaning+* induces lower FL model accuracy than *w/ cleaning*, although the former specifically targets FL systems, however, with strong assumptions that do not hold in all cases. This is consistent with our Observation 4 in IV-C. We also observe that *w/ cleaning+* induces higher EOD_{race} bias than *w/ cleaning*, which can be explained by the fact that this metric is indirectly impacted by the model utility.

Observation 6: Under non-IID settings, statistical methods of data cleaning can provide better model accuracy and fairness than the considered FL-based cleaning methods.

Fig. 3. Data cleaning under various FL non-IID settings

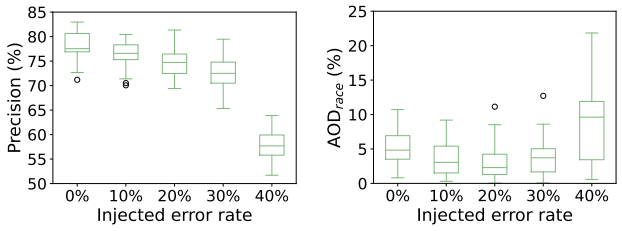


C. Data Cleaning Under Various Error Rates

To evaluate the robustness of FL data cleaning under different amounts of errors, we randomly inject synthetic errors into a dataset, in addition to the native errors in that dataset, before applying a combination of *ol-std-md-l*, *mv-md-l*, and *le-cl* cleaning methods, to simultaneously handle outliers, missing values, and label errors. Figure 4 presents the results of Adult/LR with various rates of injected label errorsfootnote[3]. We observe that under low error rates (up to 20%), the FL model resulting from the cleaned data maintains a stable utility in terms of precision, as well as a stable fairness level in terms of AOD_{race}. However, with higher error rates, both FL model utility and fairness worsen drastically, mainly because *le-cl* fails to learn correct data patterns with fewer clean data.

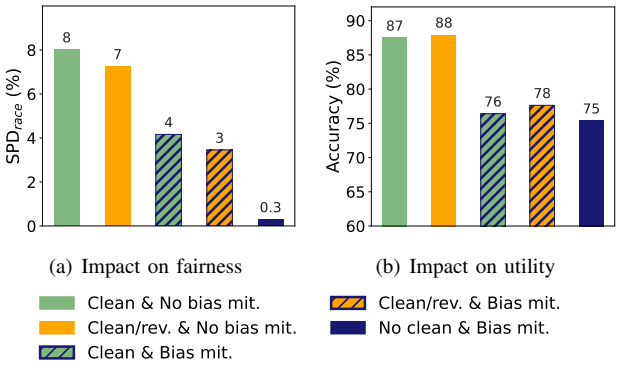
Observation 7: In the presence of high error rates, the FL models, although trained over cleaned data, are less robust.

Fig. 4. Impact of data cleaning with various rates of error injection



(a) Impact on utility (b) Impact on fairness

Fig. 5. Combining data cleaning with bias mitigation



(a) Impact on fairness (b) Impact on utility

D. Data Cleaning vs. Bias Mitigation

In the following, we analyze the impact of FL data cleaning when combined with a state-of-the-art FL bias mitigation method [45]. In Figure 5, we compare the behavior of five cases: (i) the FL model resulting from clean data after correcting its outliers, missing values, and label errors using the *ol-std-md-l*, *mv-md-l*, and *le-cl* methods, respectively (*Clean*); (ii) the FL model resulting from applying the same cleaning methods in a different order, first *le-cl*, then *mv-md-l*, and finally *ol-std-md-l* (*Clean/rev.*); (iii) the FL model from the cleaned data *Clean* to which FL bias mitigation is applied; (iv) the FL model from the cleaned data *Clean/rev.* to which FL bias mitigation is applied; and (v) as a baseline, the FL model resulting from the original corrupt data to which FL bias mitigation is applied. These results are from Adult dataset with FL logistic regression model[3]. Figure 5(a) (resp. Figure 5(b)) compares model fairness (resp. utility) in all five cases. The results show that although FL bias mitigation reduces the bias on cleaned data, the combination of data cleaning methods, whatever their order is, with bias mitigation induces a much

[3]Results of other datasets can be found in [28].

higher bias than the baseline without data cleaning. The negative impact of combining data cleaning with bias mitigation is also perceptible for model accuracy, which drops from 87%-88% to 76%-78%. Thus, the joint application of state-of-the-art FL bias mitigation and data cleaning does not allow to correctly handle FL model utility and fairness, since the underlying techniques have side-effects on each other [7], [46].

Observation 8: Combining state-of-the-art FL bias mitigation and data cleaning has a negative impact on model utility without fully achieving model fairness.

REFERENCES

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *The 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 54, Fort Lauderdale, FL, USA, 2017, pp. 1273–1282.

[2] H. Hwang, "New AI readiness report reveals insights into ML lifecycle," 2022. [Online]. Available: https://www.datacenterknowledge.com/ai-data-centers/new-ai-readiness-report-reveals-insights-into-ml-lifecycle

[3] Zeitgeist, "AI Readiness Report 2024," 2024. [Online]. Available: https://go.scale.com/hubfs/Content/Scale%20Zeitgeist%20AI%20Readiness%20Report%202024%204-29%20final.pdf

[4] N. Polyzotis, S. Roy, S. E. Whang, and M. Zinkevich, "Data Lifecycle Challenges in Production Machine Learning: A Survey," *SIGMOD Rec.*, vol. 47, no. 2, p. 17–28, 2018.

[5] S. E. Whang and J.-G. Lee, "Data Collection and Quality Challenges for Deep Learning," *Proceedings of the VLDB Endowment*, vol. 13, no. 12, pp. 3429–3432, 2020.

[6] S. Vucinich and Q. Zhu, "The Current State and Challenges of Fairness in Federated Learning," *IEEE Access*, vol. 11, pp. 80 903–80 914, 2023.

[7] N. Benarba and S. Bouchenak, "Bias in Federated Learning: A Comprehensive Survey," *ACM Computing Surveys*, vol. 57, no. 11, pp. 1–36, 2025.

[8] J. Xu, Z. Chen, T. Q. Quek, and K. F. Ernest Chong, "FedCorr: Multi-Stage Federated Learning for Label Noise Correction," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022)*, New Orleans, LA, USA, Jun. 2022, pp. 10 174–10 183.

[9] S. Min, H. Asif, X. Wang, and J. Vaidya, "Cafe: Improved Federated Data Imputation by Leveraging Missing Data Heterogeneity," *IEEE Transaction on Knowledge and Data Engineering*, vol. 37, no. 5, pp. 2266–2281, May 2025.

[10] Z. Yao and C. Zhao, "FedTMI: Knowledge Aided Federated Transfer Learning for Industrial Missing Data Imputation," *Journal of Process Control*, vol. 117, pp. 206–215, 2022.

[11] K. Itokazu, L. Wang, and S. Ozawa, "Outlier Detection by Privacy-Preserving Ensemble Decision Tree Using Homomorphic Encryption," in *International Joint Conference on Neural Networks (IJCNN 2021)*, Shenzhen, China, 2021.

[12] M. Heyden, J. Wilwer, E. Fouche, V. Arzamasov, S. Thoma, S. Matthiesen, and T. Gwosch, "Tandem Outlier Detectors for Decentralized Data," in *Proceedings of the 34th International Conference on Scientific and Statistical Database Management (SSDBM '2022)*, E. Pourabbas, Y. Zhou, Y. Li, and B. Yang, Eds. New York: Assoc Computing Machinery, 2021, p. 25.

[13] W. Du, D. Xu, X. Wu, and H. Tong, *Fairness-aware Agnostic Federated Learning*, Virtual Event, 2021, pp. 181–189.

[14] D. Y. Zhang, K. Ziyi, and W. Dong, "FairFL: A Fair Federated Learning Approach to Reducing Demographic Bias in Privacy-Sensitive Classification Models," in *The 2020 IEEE International Conference on Big Data*, Atlanta, Georgia, USA, 2020, pp. 1051–1060.

[15] F. Gröger, S. Lionetti, P. Gottfrois, A. Gonzalez-Jimenez, L. Amruthalingam, E. V. Goessinger, H. Lindemann, M. Bargiela, M. Hofbauer, O. Badri, P. Tschandl, A. Koochek, M. Groh, A. A. Navarini, and M. Pouly, "CleanPatrick: A Benchmark for Image Data Cleaning," 2025.

[16] Y. Zhang, Z. Jin, F. Liu, W. Zhu, W. Mu, and W. Wang, "ImageDC: Image Data Cleaning Framework Based on Deep Learning," in *2020 IEEE International Conference on Artificial Intelligence and Information Systems (ICAIIS)*, 2020, pp. 748–752.

[17] H. H. Malik and V. S. Bhardwaj, "Automatic Training Data Cleaning for Text Classification," in *2011 IEEE 11th International Conference on Data Mining Workshops*, 2011, pp. 442–449.

[18] M. Yang, L. Chen, and J. Tian, "A Training Data Cleaning Framework for Video Distortion Diagnosis," in *2017 12th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, 2017, pp. 214–217.

[19] M. Eugenia Fernandes, W. Geremias dos Santos, B. Masiero, L. Rodriguez Forti, and G. Moura de Holanda, "Data Cleaning applied to Ecoacoustic Datasets: A Literature Review," in *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, vol. 272, no. 3, 2025, pp. 1383–1394.

[20] R. Kohavi, "Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996, pp. 202–207.

[21] UCI Machine Learning Repository, "Census-Income (KDD)," 2000. [Online]. Available: https://doi.org/10.24432/C5N30T

[22] R. L. Shinmoto Torres, D. C. Ranasinghe, Q. Shi, and A. P. Sample, "Sensor Enabled Wearable RFID Technology for Mitigating the Risk of Falls Near Beds," in *2013 IEEE International Conference on RFID (RFID)*, 2013, pp. 191–198.

[23] S. Ulianova, "Cardiovascular Disease dataset (Heart)," 2018. [Online]. Available: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset

[24] S. B. Cohen, "Design Strategies and Innovations in the Medical Expenditure Panel Survey," *Medical Care*, vol. 41, no. 7, p. III, 2003.

[25] E. Bisong, "Logistic Regression," in *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*, 2019, pp. 243–250.

[26] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A Comprehensive Survey on Support Vector Machine Classification: Applications, Challenges and Trends," *Neurocomputing*, vol. 408, pp. 189–215, 2020.

[27] B. B. Chaudhuri and U. Bhattacharya, "Efficient Training and Improved Performance of Multilayer Perceptron in Pattern Classification," *Neurocomputing*, vol. 34, no. 1, pp. 11–27, 2000.

[28] Anonymous, "TITANIA's Repository," 2026. [Online]. Available: https://anonymous.4open.science/r/titania-4A2C/

[29] K. Hiniduma, S. Byna, and J. L. Bez, "Data Readiness for AI: A 360-Degree Survey," *ACM Computing Surveys*, vol. 57, no. 9, pp. 1–39, 2025.

[30] C. G. Northcutt, T. Wu, and I. L. Chuang, "Learning with Confident Examples: Rank Pruning for Robust Classification with Noisy Labels," in *Conference on Uncertainty in Artificial Intelligence (UAI)*, Sydney, Australia, 2017.

[31] Z. Zhao, R. Birke, R. Han, B. Robu, S. Bouchenak, S. B. Mokhtar, and L. Chen, "Enhancing Robustness of On-line Learning Models on Highly Noisy Data," *IEEE Transactions on Dependable and Secure Computing*, vol. 18, no. 1, pp. 269–282, 2021.

[32] N. M. Müller and K. Markert, "Identifying Mislabeled Instances in Classification Datasets," in *International Joint Conference on Neural Networks (IJCNN)*, Budapest, Hungary, 2019.

[33] Z. Zhao, S. Cerf, R. Birke, B. Robu, S. Bouchenak, S. Ben Mokhtar, and L. Y. Chen, "Robust Anomaly Detection on Unreliable Data," in *The 49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN 2019)*, 2019, pp. 630–637.

[34] M. Bashari, M. Sesia, and Y. Romano, "Robust Conformal Outlier Detection under Contaminated Reference Data," in *The 42nd International Conference on Machine Learning (ICML 2025)*, 2025.

[35] Oona Rainio and Jarmo Teuho and Riku Klén, "Evaluation Metrics and Statistical Tests for Machine Learning," *Scientific Reports*, vol. 14, no. 1, p. 6086, 2024.

[36] Cleanlab, "Cleanlab open-source documentation," 2025. [Online]. Available: https://docs.cleanlab.ai/

[37] A. M. Mood, F. A. Graybill, and D. C. Boes, *Introduction to the Theory of Statistics*, 3rd ed. McGraw-Hill, 1974.

[38] J. W. Tukey, *Exploratory Data Analysis*. Addison-Wesley, 1977.

[39] Y. Zhoua, S. Aryala, Mohamed, and R. Bouadjenek, "A Comprehensive Review of Handling Missing Data: Exploring Special Missing Mechanisms," 2024.

[40] KrishnaTeja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, Abir De, Rishabh K. Iyer, "GRAD-MATCH: Gradient Matching based Data Subset Selection for Efficient Deep Model Training," in *International Conference on Machine Learning*, Virtual Event, 2021, pp. 5464–5474.

[41] KrishnaTeja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, Rishabh K. Iyer, "GLISTER: Generalization Based Data Subset Selection for Efficient and Robust Learning," in *AAAI Conference on Artificial Intelligence*, Virtual Event, 2021, pp. 8110–8118.

[42] L. Gao, H. Fu, L. Li, Y. Chen, M. Xu, and C.-Z. Xu, "FedDC: Federated Learning with Non-IID Data via Local Drift Decoupling and Correction," in *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022)*, 2022, pp. 10 102–10 111.

[43] M. Yurochkin, M. Agarwal, S. Ghosh, K. Greenewald, N. Hoang, and Y. Khazaeni, "Bayesian Nonparametric Federated Learning of n-Neural Networks," in *International Conference on Machine Learning (ICML 2019)*, 2019, pp. 7252–7261.

[44] S. Kotz, N. Balakrishnan, and N. L. Johnson, *Continuous Multivariate Distributions, Volume 1: Models and Applications*. John Wiley & Sons, 2019, vol. 334.

[45] Y. Djebrouni, N. Benarba, O. Touat, P. De Rosa, S. Bouchenak, A. Bonifati, P. Felber, V. Marangozova, and V. Schiavoni, "Bias Mitigation in Federated Learning for Edge Computing," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 7, no. 4, pp. 1–35, 2023.

[46] S. Dehdashtian, B. Sadeghi, and V. N. Boddeti, "Utility-Fairness Trade-Offs and How to Find Them," in *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024)*, June 2024, pp. 12 037–12 046.