Fault Lines: Benchmarking the Impact of Label Data Quality on ML Robustness and Fairness

David Jackson*
University of Amsterdam
Amsterdam, Netherlands
d.i.jackson@uva.nl

Paul Groth
University of Amsterdam
Amsterdam, Netherlands
p.t.groth@uva.nl

Hazar Harmouch
University of Amsterdam
Amsterdam, Netherlands
h.harmouch@uva.nl

ABSTRACT

Artificial intelligence systems depend critically on high-quality data, yet real-world datasets are often imperfect. Label noise, such as incorrect or biased labels, can lead to suboptimal model decisions. While label noise has garnered increasing attention, existing research primarily examines random noise, employs simpler models, or relies on limited evaluation criteria. To address this, we introduce Fault Lines, a comprehensive, model-agnostic benchmark comprising 15 datasets systematically corrupted with diverse types of label noise, paired with an evaluation framework. This resource supports the evaluation of data cleaning pipelines and guides the design of models that are robust, in both performance and fairness, to label noise. We benchmark the robustness to label noise of 22 state-of-the-art classification models, including gradient boosting, transformers, and fairness-oriented models. Our findings show that many models maintain strong performance under high random noise (e.g., up to 40% noise leads to only a modest reduction in Robust GBDT performance). However, these models are significantly less robust to even small amounts of biased noise (<10%), which can cause substantial performance drops (e.g., 7% noise reduces ResNet's AUC by 4.4% on average) or maintain apparent stability at the expense of severe fairness degradation (e.g., MLP's Predictive Parity difference increases by 700% under 30% biased noise in the ACS Unemployment dataset). We investigate how different model architectures handle the impact of biased noise. Notably, transformer-based models appear more robust than boosting models when handling biased noise, though this advantage depends on tuning and comes with higher variance. Finally, we identify key factors for ML practitioners to mitigate the effects of label noise, including model selection, dataset analysis, and preprocessing.

PVLDB Reference Format:
David Jackson, Paul Groth, and Hazar Harmouch. Fault Lines: Benchmarking the Impact of Label Data Quality on ML Robustness and Fairness. PVLDB, 19(4): 670-683, 2025.
doi:10.14778/3785297.3785308

PVLDB Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://github.com/fault-lines/fault_lines_benchmark.

Proceedings of the VLDB Endowment, Vol. 19, No. 4 ISSN 2150-8097.
doi:10.14778/3785297.3785308

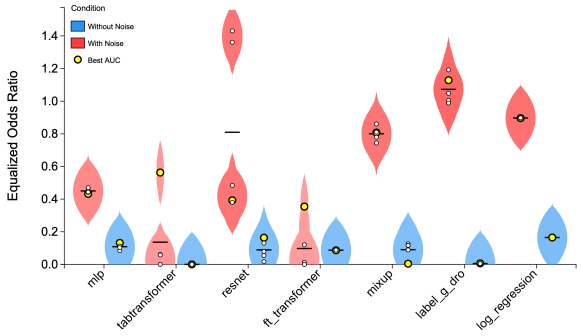


Figure 1: Equalized odds of different classifiers trained using clean (blue) versus biased data (red). Higher values equal higher unfairness.

1 THE RIPPLE EFFECTS OF LABEL NOISE

Artificial intelligence (AI) has revolutionized decision-making across industries, from healthcare and finance to criminal justice and education. At the heart of this transformation lies data — the fuel of modern AI systems [95]. Given data's central role, its quality is a critical factor in determining the reliability [59], fairness [7], and societal impact [11] of AI applications. In real-world applications, however, data is seldom pristine. Label errors, arising from human oversight, ambiguous labeling, or inconsistent collection practices [32], introduce noise that can compromise machine learning (ML) models. Studies reveal that up to 30% of labels in large-scale healthcare datasets may be inaccurate, reflecting significant challenges in acquiring reliable annotations [65, 69]. Beyond training data, pervasive label errors in test sets of widely used benchmark datasets, such as ImageNet and CIFAR-10, destabilize model evaluations, with studies estimating error rates up to 10%, undermining the reliability of the reported accuracies [64].

In high-stakes domains like healthcare and criminal justice, label noise in AI systems can have profound consequences. Inaccurate labels in medical datasets, which studies estimate affect up to 30% of large-scale healthcare data [65], can lead to misdiagnoses, jeopardizing patient safety and eroding trust in AI-driven systems. Similarly, noisy or biased data in criminal justice applications can amplify systemic inequities, perpetuating unfair outcomes for marginalized groups [61]. As Figure 1 illustrates, biased training data can significantly skew fairness metrics across different classifiers. These errors, if unaddressed, risk automating and magnifying human biases, undermining the promise of equitable AI. The data management community plays a central role in responsible AI [82]. Given that prior work has shown that automated data cleaning can

inadvertently harm model performance and fairness [33, 62], and modern cleaning methods are increasingly designed with machine learning tasks in mind [60], understanding the robustness of ML models becomes crucial to develop better data cleaning methods.

Addressing label noise requires robust technical solutions and evaluation frameworks. Techniques like noise-robust loss functions or data cleaning can mitigate the impact of erroneous labels, but their effectiveness varies across domains [35] or comes with additional risks [33]. Moreover, evaluating model fairness under noisy conditions demands metrics that account for disparate impacts across demographic groups. As AI systems increasingly inform high-stakes decisions, integrating these strategies is essential for ensuring both performance and equity.

Despite increasing awareness, significant gaps persist in the literature. Prior studies (see Section 2) have extensively explored label noise's impact on model accuracy, often relying on random perturbations to simulate errors. However, these approaches overlook patterns reflective of real-world biases tied to data collection or societal inequities and its interplay with fairness, particularly in tabular data. While more recent work [91] begins to address group-dependent noise in fairness contexts, it remains limited to random flips within groups, neglecting correlated or feature-driven noise structures. Furthermore, most label noise research focuses solely on training deep learning models on image or text data [44, 53, 71]. Beyond this, there is a lack of comprehensive evaluations of how such realistic noise affects a broad range of state-of-the-art models, as well as analyses grounded in diverse, reproducible datasets that reflect real-world complexity.

In this study, we address these shortcomings by systematically benchmarking the effect of label noise on model robustness across 15 diverse datasets, using a range of modern ML models. We discuss six noise strategies, including random, conditional, and correlated, designed to emulate real-world label imperfections in binary classification tasks. Our analysis spans both performance and fairness, assessing how label noise affects accuracy, AUC, and equity across subgroups defined by socioeconomic status, age, and others. We evaluate a suite of models, including gradient boosting methods, transformers, and models tailored for unfair or out-of-domain data. Under varying noise conditions, we test their robustness on both in-distribution and out-of-distribution data, while employing multiple fairness metrics to capture trade-offs. Our findings reveal the varied impacts of label noise and underscore the need for robust algorithms, as well as effective data preparation and wrangling methods to mitigate such errors.

The main contributions of this work are as follows:

- **Realistic Benchmark Dataset**: A collection of 15 diverse tabular datasets subjected to different label noise strategies motivated by real-world data quality issues encountered in binary classification tasks, such as those arising from human oversight or inconsistent labeling practices (see Section 3.2).
- **Evaluation Framework**: A reproducible and extendable experimental setup, available on GitHub, designed to assess the impact of label noise on ML models, with a comprehensive inclusion of multiple fairness metrics to evaluate trade-offs across demographic groups (see Section 3.3).
- **Benchmarking 20 SOTA Models**: An assessment of robustness under varying noise levels across a diverse suite of 20 modern ML models, including gradient boosting methods, neural networks, transformer-based architectures, and fairness-aware and domain robustness optimization techniques (see Section 4).
- **Practical Insights**: A set of actionable lessons for data and ML practitioners coupled with suggestions for future research directions (see Section 5).

By tackling the aforementioned challenges, we highlight the critical role of data quality in building robust ML systems and lay the groundwork for more trustworthy AI pipelines. Before detailing our methodology and results, we briefly review related work.

2 RELATED WORK

We position our work within the data quality for machine learning line of research that emphasizes the critical role of data management in ensuring robust and fair ML systems [59, 73, 82].

Label Noise. The impact of data quality on ML has garnered significant attention, including label noise and its effects on model performance and fairness. Frénay and Verleysen [26] offer a detailed survey on label noise sources and impacts on performance of classification, proposing a taxonomy (Noisy Completely at Random (NCAR), Noisy at Random (NAR), Noisy Not at Random (NNAR)). Mitigation strategies are also discussed; however, the success of models designed to handle label noise is typically confined to less complex noise patterns [20, 89]. Algan and Ulusoy [3] investigate the effects of uniform, class-dependent, and feature-dependent label noise on deep learning, revealing its severe impact on test accuracy. Most label noise research focuses on training deep learning models on images [13, 93] or textual data [71, 94]. Tools like k-nearest neighbor (k-NN) [9] and support vector machines (SVMs) [75] have been shown to be sensitive to label noise [57]. Mohammed et al. [59] research the effects of six data quality dimensions, including label accuracy, on ML for tabular data, focusing on simpler models and random noise.

Fairness. The impact of label noise on model fairness is often-neglected. An emerging and rapidly growing field [77], fairness in ML, seeks to prevent biases in data and inaccuracies in models from resulting in unjust treatment of individuals based on specific traits [66]. Wang et al. [91] explore the detrimental effects of group-dependent label noise on fair classification, demonstrating how naive parity constraints can harm accuracy and fairness. Liao and Naghizadeh [55] investigate the interplay of social and data biases in fair ML, offering practical guidelines for selecting fairness criteria or de-biasing strategies based on their robustness to label noise and feature errors.

Similar Benchmarks. Several benchmark datasets exist for studying data quality issues for AI systems, but they either lack comprehensiveness [52] or consist of image and text classification datasets [94]. Some benchmark datasets have a different focus, such as Shah et al. [76] who study the impact of categorical duplicates in AI, Liang et al. [54] for federated learning, or Hirzel and Feffer [39], which emphasizes dataset realism over fairness under noise. Our work builds upon TableShift [28], a benchmark for distribution shifts in tabular data, a related but separate task. We leverage its data, which includes

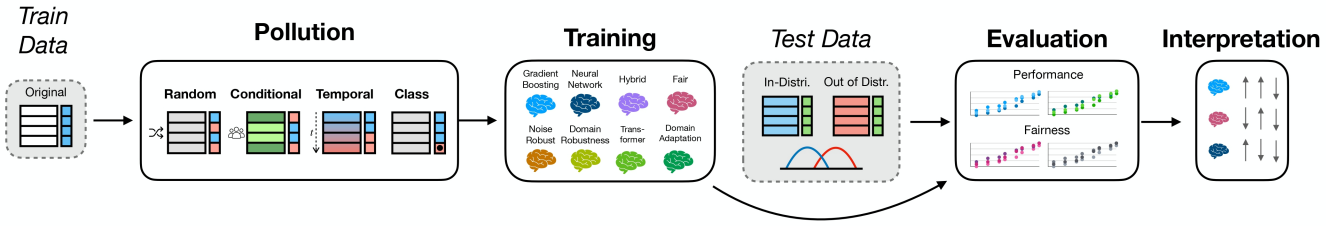


Figure 2: Visualization of the Fault Lines benchmark pipeline for evaluating label noise effects on ML models.

real-world tabular datasets with diverse domains and sensitive attributes, to systematically introduce group-dependent label noise and assess fairness implications. Furthermore, the benchmark's encompassed distribution shifts make it possible for us to test how ML models trained on noisy/biased data perform on unseen data, two challenges that appear unexplored in tandem within the current literature. This intersection distinguishes our study, bridging the gap between label noise, robustness, and fairness-aware evaluation using a comprehensive list of models and tabular datasets.

3 THE FAULT LINES BENCHMARK

The Fault Lines benchmark systematically evaluates label noise's impact on ML models through a novel taxonomy, diverse datasets, and comprehensive fairness metrics. This section outlines our experimental framework, detailed in three parts: label noise taxonomy, dataset description, and robustness metrics. Figure 2 gives an overview of the Fault Lines workflow.

3.1 Label Noise Taxonomy

In this paper, we focus on binary classification in the presence of label noise, examining its impact on model performance and fairness. Here we introduce key concepts of label noise and fairness metrics essential to understanding our study's methodology and findings. We employ six label noise strategies, adapted from established frameworks such as Random Noise and Class-Conditional Noise as described in prior work (e.g., [26], [81]), and further refined by us to capture diverse real-world noise patterns. While drawing on these standard approaches, we introduce novel formulations, particularly in Correlated Feature Noise, Concatenated Feature Noise, Temporal Noise, and Combined Noise, tailored to model more nuanced label perturbations specific to binary classification experiments. These strategies vary in the degree of determinism and randomness, allowing us to explore a wide range of noise behaviors and their impact on model performance. Our approach specifically targets label pollution without altering the feature distributions, as we aim to isolate the effects of label noise on binary classification tasks.

Random Noise. In random label noise, the labels $y \in 0, 1$ are flipped independently of the feature values, aligning with NCAR [26]. Each row i has a probability p of having its label flipped. This noise process can be represented as:

$$p(y_i) = \begin{cases} y_i & \text{with probability } (1-p), \\ 1-y_i & \text{with probability } p. \end{cases}$$

Example: In the FICO HELOC dataset (Table 1), where the task is to predict if an applicant is likely to repay a line of credit, random noise could occur due to data entry errors by staff [32]. For instance, a clerical worker might accidentally flip a customer's true label due to a typo or misclick during manual data entry.

Conditional Feature Noise. In conditional feature noise, the label is flipped only when a specific condition on a feature or features is met. For instance, when a particular feature f_j equals a certain value (e.g., $f_j = x$), the label flips with probability p. Mathematically, this can be expressed as:

$$\hat{y}_i = \begin{cases} y_i & \text{if } f_j \neq x, \\ \begin{cases} y_i & \text{with probability } (1-p), \\ 1-y_i & \text{with probability } p \end{cases} & \text{if } f_j = x. \end{cases}$$

Here, f_j represents a specific feature, and the label is perturbed only when this feature equals x.

Example: In the Hypertension dataset (Table 1), where the task is to predict hypertension, conditional feature noise might occur if a specific feature, such as poverty, triggers a labeling error. For example, individuals in low-income areas could be misclassified due to limited healthcare access, affecting the accuracy of recorded diagnoses.

Correlated Feature Noise. It involves label flipping based on the interaction between multiple features. For example, labels are flipped if both $f_j = x$ and $f_k = y$. This type of noise is contingent on the simultaneous satisfaction of multiple conditions:

$$\hat{y}_i = \begin{cases} y_i & \text{if } f_j \neq x \text{ or } f_k \neq y, \\ \begin{cases} y_i & \text{with probability } (1-p), \\ 1-y_i & \text{with probability } p \end{cases} & \text{if } f_j = x \text{ and } f_k = y. \end{cases}$$

Here, f_j and f_k represent features that, when equal to x and y respectively, trigger a label flip with a specified probability.

Example: In the Hypertension dataset, a label could flip with probability p if an individual is elderly and from a low-income area due to intersectional errors from limited healthcare access and age-related assumptions.

Concatenated Feature Noise. Concatenated feature noise occurs when labels are flipped based on the presence of either one of multiple conditions. For instance, if $f_j = x$ or $f_k = y$, the label is

Table 1: The 15 training datasets included. Subgroup refers to the condition chosen for noise, and P_S denotes the proportion of entries in the training dataset that fulfill this condition (e.g., 24.9% of rows in the Hypertension are labeled as poor).

Name	Rows	Imbalance Ratio	Subgroup	P_S	Direction*
FICO HELOC [24]	2,220	0.3175	Delinquency Status	21.3%	$0 \to 1$
ANES Voting [84]	4,159	0.4452	No Answer	33.3%	$0 \to 1$
ICU Hospital Mortality [92]	7,116	0.0816	Age	30.4%	$1 \to 0$
ICU Length of Stay [92]	8,634	0.6785	Age	29.4%	$0 \to 1$
Nhanes Lead [16]	11,807	0.0271	Missing Demographic	26.0%	$0 \to 1$
Hospital Readmission [83]	34,288	0.7368	Race	20.2%	$0 \to 1$
College Scorecard [88]	98,556	0.1428	Veteran	25.5%	$0 \to 1$
Hypertension [15]	216,411	0.67	Poverty	24.9%	$1 \to 0$
Food Stamps [21]	629,018	0.2347	Race	28.2%	$1 \to 0$
Diabetes [15]	969,229	0.1425	High BMI	30.9%	$0 \to 1$
Sepsis [72]	1,122,299	0.012	High Heart Rate	19.0%	$0 \to 1$
Income [21]	1,264,123	0.4763	Race	28.9%	$1 \to 0$
Unemployment [21]	1,290,914	0.0353	Race	23.4%	$0 \to 1$
Assistments [23]	2,132,526	0.4412	High Attempt Usage	19.4%	$0 \to 1$
Public Coverage [21]	4,006,249	0.2885	Race	28.6%	$0 \to 1$

*Direction indicates the predominant label flip applied during noise injection (e.g., $0 \to 1$ means class 0 labels are more likely to be flipped to class 1).

flipped with probability p. Mathematically:

$$\hat{y}_i = \begin{cases} y_i & \text{if } f_j \neq x \text{ and } f_k \neq y, \\ \begin{cases} y_i & \text{with probability } (1-p), \\ 1-y_i & \text{with probability } p \end{cases} & \text{if } f_j = x \text{ or } f_k = y. \end{cases}$$

Example: In the College Scorecard dataset (Table 1), where the goal is to predict if an educational institution has a low completion rate, a high percentage of non-traditional students such as veterans or extreme admission rates could lead to misreporting.

Temporal or Contextual Noise. In temporal or contextual noise, the probability of label corruption depends on the time t_i or the context of the data, such as geographical location. The probability $p(t_i)$ of flipping the label varies along the time axis, where either the oldest or the newest data can have the highest flip probability, depending on the context. We define a maximum flip probability $p_{\max}$ and a minimum flip probability $p_{\min}$, and these can be assigned to either end of the time spectrum (e.g., oldest or newest data). The probability $p(t_i)$ can increase or decrease linearly, exponentially, or logarithmically depending on the temporal distance from a reference point, such as the most recent data t_{ref}:

$$p(t_i) = p_{\min} + \left(\frac{|t_i - t_{\text{ref}}|}{t_{\max} - t_{\min}} \right) \cdot (p_{\max} - p_{\min})$$

Example: In the FICO HELOC dataset, a loan repayment label might flip (e.g., from "will repay" to "won't repay") with probability (p_{t_i}), where older records from 10 years ago have a higher p_{t_i} due to outdated manual entry errors, while recent data has a lower p_{t_i} from automated systems.

Class-Conditional Noise. In class-conditional noise, the probability of label noise depends on the class itself. For example, instances of class 1 might have a higher probability of label noise than those of class 0, which have a probability p_0. This can be formalized as:

$$\hat{y}_i = \begin{cases} y_i & \text{with probability } (1 - p_{y_i}), \\ 1 - y_i & \text{with probability } p_{y_i}, \end{cases}$$

Example: In the ACS Food Stamps dataset (Table 1), "not receiving" (class 0) labels might flip to "receiving" with low p_0, while "receiving" (class 1) flips to "not receiving" with higher p_1, due to stricter recertification processes for current recipients.

Combining Multiple Noise Types. In addition to using individual noise types, we can combine multiple noise strategies to better capture real-world complexities. For example, we can combine Conditional Feature Noise with Class-Conditional Noise to apply different probabilities of label flipping based on both feature conditions and class membership. Consider the scenario where labels are flipped for rows where $f_j = x$. If the true label is 0, the flip probability is higher compared to cases where $y_i = 1$. This can be formalized as follows:

$$p(y_i, f_j) = \begin{cases} p_{\text{flip0}} & \text{if } f_j = x \text{ and } y_i = 0, \\ p_{\text{flip1}} & \text{if } f_j = x \text{ and } y_i = 1, \end{cases}$$

where $p_{\text{flip0}} > p_{\text{flip1}}$.

Example: In the Diabetes dataset, when High BMI is "true", the flip probability depends on the class. If the true label is "no diabetes", the flip to "diabetes" has a higher probability because clinicians might overassume diabetes risk in overweight individuals, while if the true label is "diabetes", the flip to "no diabetes" has a lower probability since confirmed diagnoses are rarely reversed.

3.2 Datasets and realistic label noise

To systematically investigate label noise's effects on ML models, we leverage datasets from [28], encompassing 15 diverse tabular datasets tailored for binary classification tasks summarized in Table 1. These datasets span domains like healthcare (e.g., Hospital Readmision, Diabetes), finance (e.g., FICO Heloc), and social outcomes (e.g., Food Stamps, Unemployment). Furthermore, the datasets exhibit variation in key attributes, including size, imbalance ratio, and feature complexity, enabling a comprehensive analysis of how these factors modulate model resilience and fairness outcomes.

While Fault Lines includes all six noise strategies outlined in Section 3.1, we conduct in-depth experiments focusing on two specific types: (1) random noise and (2) a combination of conditional feature and class-conditional noise, which we here refer to as biased noise. *Random noise* serves as a baseline, its simplicity and well-studied effects on accuracy [3, 26] providing a reference point for comparing more complex, systematic noise types. In contrast, *biased noise*, the hybrid of conditional feature and class-conditional noise, where labels flip based on feature values (e.g., Poverty = 1) and vary by class probability (e.g., higher for $Y = 1$ than $Y = 0$), aims to capture real-world imperfections more effectively. This combination simulates systematic biases, such as those tied to ethnicity, age, temporary work status, or socioeconomic factors (see Section 3.1). We prioritize this kind of noise because prior research highlights its relevance to fairness, capturing interactions between feature-driven errors and class imbalances that exacerbate inequities in high-stakes domains like healthcare and justice, unlike the studied random noise [4, 6, 45, 63].

To introduce a *realistic biased noise*, we ground our choices of the conditions in a rationale supported by empirical research, reflecting real-world disparities and biases, as follows: age [63, 67], race and ethnicity [6, 46], work status/poverty [14, 68], veteran status [42], high BMI [50], and missing responses or demographics [12, 58, 90]. More details for each of the conditions is available on our GitHub.

3.3 Robustness Metrics

As previously mentioned, we evaluate model robustness based on two criteria: performance and fairness. We assess performance using accuracy and the AUC. Accuracy provides a straightforward measure of overall correctness, while AUC is particularly valuable for handling imbalanced datasets, as it captures the trade-off between true positive and false positive rates across various classification thresholds.

Fairness is a multifaceted concept, and no single metric can capture all its dimensions. Different fairness metrics measure different aspects of fairness, such as equal outcomes (Demographic Parity), balanced error rates (equalized odds), or trustworthiness of predictions (Predictive Parity). Moreover, many fairness metrics are mutually incompatible, meaning that optimizing for one metric can worsen performance on another [29]. For example, achieving Demographic Parity might require sacrificing equalized odds or accuracy parity. By evaluating multiple fairness metrics, we ensure a more comprehensive understanding of the model's behavior and identify potential trade-offs. This approach aligns with recent work in fairness-aware ML, which emphasizes the importance of context-dependent fairness definitions and the limitations of relying on a single metric [51].

Equality of Opportunity requires the true positive rate (TPR) of a model to be equal across demographic groups, denoted $A = a$ and $A = b$ (e.g., different societal categories). It is formalized as:

$$\Pr(\hat{Y} = 1 \mid Y = 1, A = a) = \Pr(\hat{Y} = 1 \mid Y = 1, A = b)$$

where $\hat{Y}$ is the predicted label and Y is the true label [37]. This metric ensures that all groups have an equal chance of receiving a correct positive prediction, making it particularly relevant in contexts where equitable access to opportunities—such as job offers or medical treatment—is a priority. It is sensitive to noise that disproportionately affects the correct identification of positive cases across groups.

Demographic Parity, also known as Statistical Parity, demands that the probability of a positive prediction ($\hat{Y} = 1$) be equal across groups, irrespective of underlying differences [37]:

$$\Pr(\hat{Y} = 1 \mid A = a) = \Pr(\hat{Y} = 1 \mid A = b).$$

This metric prioritizes equal outcomes, making it relevant in contexts like hiring or lending where parity in opportunity is a societal goal. However, it ignores base rates (e.g., true prevalence of $Y = 1$), which can mask unfairness if noise alters these rates unevenly across groups—a limitation we explore in our experiments.

Equalized Odds, sometimes called Equality of Opportunity, balances both true positive rates (TPR) and false positive rates (FPR) across groups [37]. It is expressed as:

$$\Pr(\hat{Y} = 1 \mid Y = 1, A = a) = \Pr(\hat{Y} = 1 \mid Y = 1, A = b)$$

for TPR equality, and:

$$\Pr(\hat{Y} = 1 \mid Y = 0, A = a) = \Pr(\hat{Y} = 1 \mid Y = 0, A = b)$$

for FPR equality. By ensuring that correct predictions and errors occur at similar rates, Equalized Odds suits high-stakes settings—such as medical diagnoses—where noise-driven disparities in error types (e.g., false positives) can have severe consequences. Its dual conditions make it a stringent test of fairness under our six noise strategies.

Predictive Parity focuses on the precision of positive predictions, requiring that the likelihood of a true positive given a positive prediction be uniform [18]:

$$\Pr(Y = 1 \mid \hat{Y} = 1, A = a) = \Pr(Y = 1 \mid \hat{Y} = 1, A = b)$$

This metric emphasizes the trustworthiness of affirmative outcomes, critical in scenarios like criminal justice where a prediction's reliability directly impacts individuals. Noise that skews positive predictions (e.g., via correlated feature errors) can undermine Predictive Parity.

We apply the trained models to both in-distribution and out-of-distribution test data (see Figure 2), simulating a scenario where models trained on noisy data are applied to unseen data, thereby imitating two concurrent data quality issues. To assess the fairness of ML models, we opt to evaluate the top five trials with respect to the AUC for each model, rather than focusing solely on the single highest performing trial. Given that fairness metrics such as Demographic Parity or equalized odds may not align closely with AUC, relying on a single top-performing trial risks highlighting a configuration that achieves strong predictive performance at the expense of significant fairness deficiencies. By broadening our analysis to include the top five trials, we capture a more representative range of near-optimal hyperparameter settings, enabling us to examine the trade-offs between performance and fairness more thoroughly.

4 APPLYING FAULT LINES

In this section, we describe the results of benchmarking 22 ML classifiers (see Section 4.1) using Fault Lines, with the goal of answering key questions about model behavior under label noise. Specifically, we investigate the following research questions: How

Table 2: Evaluated ML Models.

Model	Category	Key Characteristics	Reference
SVM	Linear Baseline	Margin maximization, kernel-based separation	[75]
Logistic Regression	Linear Baseline	Linear feature weighting, probabilistic output	[40]
XGBoost	Gradient Boosting	Iterative tree boosting, gradient descent optimization	[17]
LightGBM	Gradient Boosting	Leaf-wise tree growth, histogram-based splits	[47]
CatBoost	Gradient Boosting	Categorical feature handling, ordered boosting	[36]
MLP	Feedforward Neural	Multi-layer perceptrons, dense feature connections	[86]
ResNet	Feedforward Neural	Residual layers, deep feature hierarchies	[38]
FT Transformer	Attention-Based Neural	Self-attention over tokenized features	[30]
TabTransformer	Attention-Based Neural	Attention on embedded categorical features	[41]
SAINT	Attention-Based Neural	Self- and intersample attention layers	[80]
NODE	Hybrid Neural	Oblivious decision trees in neural framework	[70]
FairEG	Fairness-Aware Optimization	Randomized classifier via cost-sensitive reduction	[1]
FairGS	Fairness-Aware Optimization	Grid search over constraint multipliers	[1]
FairRS	Fairness-Aware Optimization	Random search for constraint multipliers	[1]
Robust-GBDT	Robust Optimization	Boosting with noise-aware loss adjustments	[56]
DRO	Robust Optimization	Minimax optimization over uncertainty sets	[74]
Group DRO	Robust Optimization	Group-specific minimax loss optimization	[74]
IRM	Robust Optimization	Invariant feature learning across environments	[5]
DANN	Adaptation/Augmentation	Adversarial domain-invariant feature learning	[27]
MixUp	Adaptation/Augmentation	Interpolation-based training data mixing	[96]
Label Group DRO	Robust Optimization	Group-specific loss optimization with label focus	[74]
VREX	Adaptation/Augmentation	Variance-regularized risk extrapolation	[49]

do different classification models respond to varying levels of random label noise? (Section 4.2); How do dataset characteristics (e.g., size, class imbalance, feature dimensionality) influence model robustness to label noise? (Section 4.2); How does biased noise that targets specific subgroups affect model robustness compared to random noise? (Section 4.3); What are the trade-offs between performance and fairness under biased label noise across different model architectures? (Sections 4.3 and 4.5); and how do more complex noise patterns (e.g., intersectional, temporal) impact model behavior? (Section 4.4). The complete set of results is available on our GitHub repository. Throughout this section, we use numbers (❶-⓬) to reference the empirical findings that support each of our recommendations in Table 5 in Section 5. Before diving into the analysis, we give a brief description of the investigated models.

4.1 Models

To understand the robustness of different architectures towards label noise, we evaluate a diverse set of ML models, ranging from simple, general-purpose approaches to complex, purpose-specific designs as shown in Table 2. Our initial selection of models includes gradient boosting models (e.g., XGBoost, LightGBM, CatBoost), neural network architectures (e.g., MLP, ResNet, FT Transformer), and distribution robustness models (e.g., DRO, IRM, MixUp) from the benchmarking framework of [28], supplemented by noise-robust models like Robust-GBDT (with noise-aware loss functions) and fairness-aware models like FairEG (using cost-sensitive reductions). These additional models were selected to specifically evaluate performance under label noise conditions with fairness considerations, which is the primary focus of our study. This model set spans a wide range of design principles, allowing us to assess how architectural choices affect robustness to label noise, with robustness insights detailed in Sections 4.2 and 4.4.

We include *linear models* (e.g., linear SVM) for their simplicity and interpretability, serving as a baseline to assess noise impact on basic classifiers. *Gradient boosting models* (e.g., XGBoost) are chosen due to their proven robustness and superior performance on tabular data, as highlighted by [31], which notes tree-based methods often outperform deep learning in such settings due to effective handling of structured features. As shown in our results (Section 4.2), we indeed observed this superiority in the clean and random noise settings, with boosting models like CatBoost achieving top performance. However, to our surprise, transformer-based models, among others, can demonstrate greater resilience to biased noise (see Section 4.3). For *deep learning*, we adopt feedforward neural and attention-based neural models, motivated by their widespread use in noise-robust learning, as surveyed in [81], where deep architectures excel with complex noise patterns like feature-dependent noise. Hybrid neural models combine tabular and neural strengths, offering a bridge between GBDT and deep learning for noise resilience. Robust Optimization models (e.g., those with noise-tolerant loss functions) are included to directly address label noise, leveraging techniques like those in [81] to mitigate overfitting to corrupted labels.

Adaptation/Augmentation models (e.g., domain-adaptive or data-augmented approaches) are incorporated despite their primary design for distribution shifts [28], as their ability to adapt to data variations may enhance robustness to group-dependent noise and

Table 3: Model performances (accuracy/AUC) under varying random noise rates. N_{medium} is the performance average for 10% and 20% noise, while N_{high} is the performance average for 30% and 40% noise.

Model	N_{none}	N_{medium}	N_{high}
Robust-GBDT	85.35/82.96	84.91/81.38	84.26/78.97
CatBoost	85.06/83.95	84.88/81.92	83.61/78.77
FT Transformer*	85.05/82.44	85.02/81.08	84.01/80.98
MLP	83.24/78.16	83.04/76.64	82.16/73.94
ResNet	83.77/80.95	83.41/80.02	81.98/76.93
SVM	83.43/78.26	81.15/76.39	78.11/71.83
Group DRO+	82.02/78.49	80.51/76.14	78.31/70.62
MixUp+	77.11/77.13	75.74/75.12	73.51/72.74

*2 datasets missing due to large feature dimensionality [28].

+6 datasets missing as domain generalization models cannot be trained here [28].

fairness under unseen conditions. Finally, *fairness-aware optimization models* (e.g., Fairlearn variants) are added to explicitly tackle fairness under biased labels, using LightGBM as a base to optimize fairness-performance trade-offs.

In our experiments, we adopt the training setup and hyperparameter configurations from Gardner et al. [28], which involve standardized train-validation-test splits and model-specific hyperparameter tuning. For all models included in their benchmark, we reuse their pre-optimized hyperparameters directly. In contrast, for models not covered in their work, such as Robust-GBDT, we perform our own hyperparameter tuning using Optuna, following a similar cross-validation-based strategy. We include two fairness-aware reduction methods from Agarwal et al. [1]: FairEG (exponentiated gradient) and FairGS (grid search). Both optimize for fairness constraints using LightGBM as the base classifier. We also include FairRS, a randomized reweighting approach in which fairness constraint multipliers (λ) are selected via random search. All fairness-aware models are tested under two common fairness constraints: Demographic Parity and Equalized Odds. These are applied to data with feature-level label noise, assuming prior knowledge of the noise type. Full configuration details for all models, including code and parameter settings, are available on our GitHub repository.

4.2 Random Noise

We evaluate each model under varying levels of random label noise (0%, 5%, 10%, 20%, 30%, 40%, and 50%) across all datasets. To maintain clarity, we present a subset of representative models that capture the broader trends observed across all evaluated architectures. Full results for all models, including those exhibiting similar patterns due to architectural similarity, are available on our GitHub repository. We confirmed normality of accuracy distributions at each noise level using the Shapiro–Wilk test [78]. One-way ANOVA [25] tested differences across noise groups, followed by Tukey's HSD [87] for pairwise comparisons. Normality (Shapiro–Wilk, $p > 0.05$) and equal variances (Levene's test, $p > 0.05$) were confirmed, then ANOVA and Tukey's HSD ($\alpha = 0.05$) were applied.

Table 3 summarizes the performance of six representative classifiers, averaged across all datasets. For clarity, we group label noise levels into none (0%), medium (10%–20%), and high (30%–40%). SOTA models such as CatBoost, Robust-GBDT, and FT Transformer achieve strong performance on clean data and are able to maintain high accuracy despite high label noise, without requiring specialized preprocessing ❶. Statistical analysis confirms their robustness to random noise up to 40%, with no significant degradation (ANOVA: $p = 0.670$, Tukey's HSD: mean difference ≤ -0.0591, $p_{\text{adj}} > 0.698$). Performance deteriorates significantly at 50% noise, corresponding to a 1:1 signal-to-noise ratio (ANOVA: $p < 0.01$). At low noise levels (e.g., 10%), models like CatBoost occasionally demonstrated marginal performance gains, which may be attributable to noise-induced regularization; however, these effects did not reach statistical significance.

Models such as FT Transformer exhibit a more pronounced degradation in average performance across all hyperparameter trials as noise increases (e.g., a 2.22% accuracy drop for FT Transformer from 0% to 30% noise) compared to the best-trial performance (e.g., a 0.8% drop). This suggests that FT Transformer's robustness to label noise, like that of other deep learning models, more heavily depends on careful hyperparameter tuning, particularly for parameters like regularization strength and learning rate. In contrast, boosting models exhibit more stable performance across hyperparameter settings, with minimal degradation in average performance. These findings align with the current consensus that boosting models often outperform deep learning approaches in tabular data settings [79]. The reliance on extensive hyperparameter optimization for deep learning models introduces practical trade-offs, as real-world scenarios often involve limited computational resources or noisy validation data, potentially widening the performance gap with boosting models.

Simple models, such as SVMs, not only achieve lower baseline performance but also exhibit greater sensitivity to random label noise. For example, SVMs experience a 5.32% drop in best accuracy from the clean to high-noise setting, consistent with prior findings [57]. Interestingly, models with higher baseline performance do not necessarily exhibit greater degradation under random noise. Robust-GBDT, for instance, maintains strong performance even under high noise (84.26% accuracy), outperforming SVMs trained on clean data (83.43% accuracy). Given the continued use of models like SVMs in sensitive domains such as healthcare [34], practitioners should be cautious when deploying them in noisy data scenarios ❷.

Adaptation- and augmentation-based models such as MixUp or VREX demonstrate robustness to noise (see Table 3), but begin with substantially lower baseline performance (MixUp baseline performance starts at 77.11% accuracy), limiting their practical effectiveness.

In addition to model performance, we analyze the fairness implications of random label noise across key subgroups (e.g., individuals above versus below the poverty line in the Hypertension dataset, or racial subgroups in the Hospital Readmission dataset, as seen in Table 1). Although label noise was introduced uniformly at random and independently of group membership, we observe that fairness disparities can nonetheless shift as the noise rate increases. However, the direction and magnitude of these shifts vary substantially across datasets. For example, in the Hypertension dataset, a 30% random flip led to a modest increase in equal false negative rates between groups, while in the Hospital Readmission dataset, the same noise rate slightly reduced group-disparity.

Table 4: Pearson Correlation Coefficients of dataset characteristics with random noise-induced performance drops for ResNet, CatBoost, SVM, and FT Transformer.

Characteristic	ResNet (Acc / AUC)	CatBoost (Acc / AUC)	SVM (Acc / AUC)	FT Transformer (Acc / AUC)
Imbalance Ratio	**-0.440** / **0.445**	**-0.408** / 0.372	0.282 / 0.344	**-0.446** / 0.224
Rows	0.408 / 0.303	0.272 / 0.354	0.013 / 0.109	0.088 / -0.051
Columns	-0.031 / **-0.616**†	**-0.749**† / **-0.501**	**-0.541**† / **-0.489**	**-0.729**† / 0.073
Class Separability (LDA)	0.278 / -0.158	0.117 / 0.151	-0.385 / -0.337	0.032 / 0.002
Feature-to-Row Ratio	-0.012 / **-0.663**†	**-0.679**† / **-0.520**†	**-0.571**† / **-0.519**†	0.051 / 0.045

$^{\dagger}p < 0.05$; bold values indicate $p < 0.10$. Sample size: $n = 15$ datasets for ResNet, CatBoost, SVM; $n = 13$ datasets for FT Transformer.

One factor that appears to modulate the impact of random label noise on fairness is the degree of class imbalance within and between subgroups. In datasets where subgroups differ substantially in their base rates (e.g., Hypertension), even symmetric, group-agnostic noise can disproportionately corrupt the minority class within each subgroup. This uneven degradation of signal quality may shift the learned decision boundary and amplify group-level fairness disparities. The effect is particularly pronounced when a subgroup has low prevalence of the positive class, making it more vulnerable to label flipping. For example, in the College Scorecard dataset, the SVM classifier's Equalized Odds disparity between veterans and non-veterans nearly doubled, from 0.065 at 0% noise to 0.124 at 30% noise, despite the noise being agnostic to group membership. Notably, models that maintain robust performance under noise (e.g., Robust-GBDT, CatBoost) also tend to exhibit more stable fairness metrics across noise levels.

Taken together, these results provide a more nuanced understanding of model robustness under random label noise. While SOTA models demonstrate impressive stability under high noise levels, this robustness is not universal across model types, nor is it guaranteed in practical settings. The observation that average performance across hyperparameter trials is more sensitive to noise than best-trial performance in certain models highlights that robustness often depends on ideal validation conditions and carefully tuned hyperparameters. In real-world scenarios, where validation data may also be noisy or limited, these performance margins could widen substantially. Furthermore, even when noise is group-agnostic, its fairness effects may not be uniformly benign. Instead, they are mediated by structural properties of the data such as subgroup class prevalence and model sensitivity to label noise, which can lead to unintended disparities in predictive performance between subgroups. Beyond benchmarking clean vs. noisy accuracy, understanding how tuning, data conditions, and model assumptions interact with noise is key to designing deployable systems ❸.

Impact of dataset characteristics. As we include datasets of varying domains and characteristics, such as size, class imbalance ratio, and feature dimensionality (see Section 3.2), we now analyze how these characteristics affect the impact of noise on ML classifiers. Understanding dataset characteristics may help design models that better handle noise-related performance degradation in varied real-world settings. Table 4 presents correlation coefficients between key dataset features and performance drops (from 0% to 30% noise) for four representative models of each category: ResNet, CatBoost, SVM, and FT Transformer. Distribution robust models were excluded as they could not be trained on 6 datasets.

Several patterns emerge from this analysis:

- **Class Balance:** ResNet, CatBoost, and FT Transformer, more balanced datasets (higher Imbalance Ratios) correlate with larger accuracy drops under noise (ResNet: $r = -0.440$, $p = 0.101$; CatBoost: $r = -0.408$, $p = 0.131$), while SVM shows the opposite trend. AUC robustness improves with balance across all models, suggesting that noise in imbalanced datasets particularly disrupts minority class ranking.
- **Feature Complexity:** The number of columns and feature-to-row ratio show the strongest and most consistent correlations with noise sensitivity ❹. CatBoost and SVM exhibit significantly larger accuracy drops with increasing feature complexity (columns: $r = -0.749$ and $r = -0.541$, $p < 0.05$), with FT Transformer also showing a strong negative correlation with column count. ResNet's AUC is particularly affected by higher feature-to-row ratios ($r = -0.663$, $p < 0.05$).
- **Dataset Size:** Larger datasets (more rows) show weak, non-significant correlations with reduced performance drops, suggesting that increased data volume may provide some protection against noise effects, though our sample size ($n = 15$) limits statistical power.

This shows that model robustness cannot be evaluated in isolation from dataset properties. Practitioners working with high-dimensional or imbalanced datasets should be especially cautious when deploying models in noisy environments, as performance may degrade more severely than expected. Moreover, the contrasting responses across model types (e.g., SVM vs. CatBoost under class imbalance) imply that robustness is not solely an inherent property of the model, but a function of model–data interaction.

4.3 Biased Noise

Having established the robustness of SOTA models towards random label noise, we now focus on biased noise, where specific groups in the data are targeted. As discussed in Section 1 and 3.1, this type of noise aims to capture systematic biases in real-world data more closely ❺.

In our biased noise experiments, we introduce label noise selectively to predefined sensitive subgroups within each dataset (see Table 1). Specifically, a *subgroup noise rate* of 30% is applied within the targeted subgroup, while all other data points remain unaffected. This level of subgroup-specific noise was chosen to ensure a sufficiently strong signal for evaluating robustness and fairness impacts, without overwhelming the underlying data distribution. Because only a subset of the data is modified, the *overall noise rate* at the

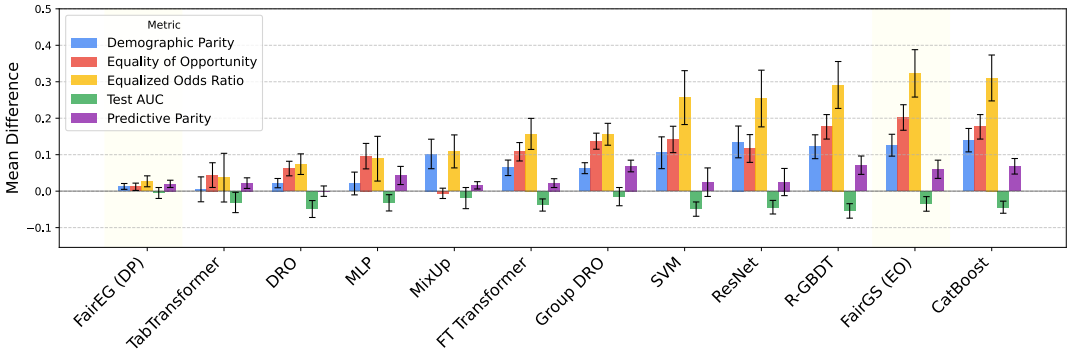


Figure 3: Impact of label noise (30% subgroup noise) on model performance and fairness metrics across different classifiers. Bar charts show mean differences between noise-free and noisy conditions for Test AUC and fairness metrics. Error bars represent pooled standard deviations. Classifiers are ordered from left to right by overall robustness to noise.

dataset level is considerably lower (7%), ranging from 5.7% (Sepsis dataset) to 9.99% (Voting dataset), depending on the prevalence of the targeted subgroup. For example, in the Hypertension dataset, we introduce noise only in instances where the poverty feature equals 1 and the label is positive (1), leaving other cases unchanged.

Overall effects of biased noise. Figure 3 presents results for 12 classifiers, chosen to represent each model category from Table 2, highlighting the diversity in robustness across architectures under biased label noise. Biased noise consistently worsened fairness across models and metrics and models, with significant increases observed in Equality of Opportunity (mean diff = 0.216, $p < 0.05$), Demographic Parity (mean diff up to 0.143, $p < 0.05$), and Equalized Odds Ratio (mean diff up to 0.348, $p < 0.05$). Predictive Parity varied widely: some models experiencing substantial shifts (mean diff = 0.102, $p < 0.05$) while others remained nearly unaffected (mean diff = 0.0002). Generally, all classifiers exhibit weaker robustness compared to the prior Section (4.2) for the aforementioned smaller noise rates ($\sim$ 7% on average).

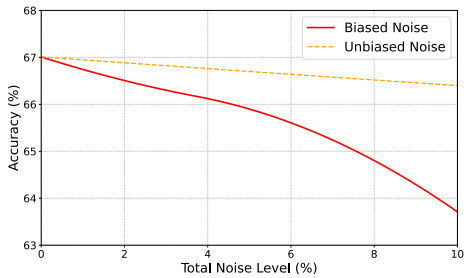


Figure 4: CatBoost accuracy for increasing noise introduced in the Hypertension Dataset using random and biased noise.

Generally, we observe that biased noise can cause significantly greater performance degradation than random noise, even when applied at equivalent overall rates. Figure 4 shows CatBoost accuracy on the Hypertension dataset, where biased noise leads to a notably sharper decline in performance compared to random noise.

Model family differences. Certain models, such as ResNet and CatBoost, exhibit substantial differences (i.e., ResNet: 0.254 for Equalized Odds ratio) in all fairness metrics between clean and noisy data, indicating high sensitivity to biased label noise. In contrast, models like TabTransformer and MixUp demonstrate greater robustness to biased noise, with smaller shifts in metrics (0.037 for TabTransformer's Equalized Odds ratio). However, their relatively low baseline accuracy (e.g., MixUp at 77.11% on clean data) limits the practical significance of this robustness.

For *fairness-aware models*, FairEG balances fairness and performance best, with modest drops under biased noise: AUC decreases by 0.8–3.6% and accuracy by about 0.3–0.8%. It's worth noting that this stability comes at the cost of a substantially lower baseline AUC compared to other classifiers (e.g., 23.97% lower baseline AUC than CatBoost), which is not evident from Figure 3, where baseline AUC is not directly shown ❻. FairGS, particularly with Equalized Odds constraints, shows significantly lower robustness, with Equality of Opportunity increasing by 169% and Equalized Odds ratio by 173%, indicating higher sensitivity to data perturbations. FairRS performs worst, with fairness metrics like Equalized Odds ratio increasing by over 300% and AUC falling by 4.6%, often underperforming standard boosting models. These results highlight that fairness-aware models also remain vulnerable to biased noise, even with prior knowledge of the noisified features. Among the fairness constraints, Demographic Parity generally yielded more stable performance than Equalized Odds.

Building on the prior analysis of *boosting and transformer models* under random label noise (Section 4.2), we now examine their performance in the presence of biased label noise. Specifically, we compare FT Transformer, the strongest-performing transformer baseline, and XGBoost, the most robust boosting model. Under biased noise, FT Transformer achieves higher peak AUC than XGBoost on 10 out of 13 datasets and exhibits a 33% smaller average AUC degradation compared to XGBoost across biased-noise settings. This contrasts with the pattern observed under random noise, where boosting models showed more consistent robustness (Section 4.2). However, this robustness comes with greater sensitivity to hyperparameter settings. FT Transformer exhibits higher variance across trials (e.g., 6.2% standard deviation in accuracy vs. 2.03%

for XGBoost), and its average performance across hyperparameter trials degrades more than that of XGBoost. Nonetheless, with appropriate hyperparameter tuning (e.g., regularization strength), FT Transformer consistently delivers strong results, indicating that its robustness is achievable but less stable ❼. Boosting models tend to offer stable performance across conditions, while transformer-based models can yield superior results under biased noise when well-tuned. While the current study focuses on empirical evaluation rather than causal explanation, we hypothesize that the advantage of FT Transformer may stem from its attention mechanism, which dynamically re-weights features and may mitigate the impact of corrupted inputs. In contrast, XGBoost's reliance on greedy, tree-based splits could make it more susceptible to biased noise when critical features are systematically corrupted.

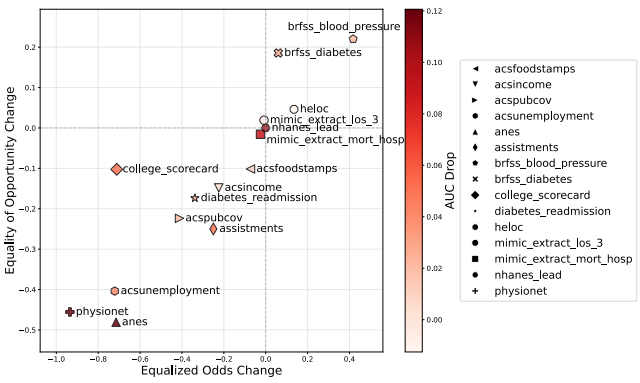


Figure 5: Impact of biased noise on fairness metrics and model performance (AUC Drop) across datasets using the ResNet model. Color indicates magnitude of the degradation.

Dataset-level effects. Figure 5 visualizes the effects of biased label noise on the ResNet model across all datasets, showing changes in fairness metrics (Equalized Odds and Equality of Opportunity) alongside AUC degradation. We use ResNet as a representative model here due to its moderate baseline performance and sensitivity to label noise, which makes fairness shifts more visible across datasets. In datasets such as Voting and Sepsis, biased noise leads to significant drops in both predictive performance and fairness, indicating broad vulnerability. In contrast, datasets like Public Coverage and Income exhibit substantial fairness degradation with minimal change in AUC, suggesting that performance metrics alone may mask growing disparities across subgroups. The reverse occurs in the Hospital Mortality dataset, where AUC drops sharply while fairness metrics remain relatively stable.

Our experiments demonstrate a fundamental asymmetry between random and biased noise: while models can exhibit stability under random noise levels up to 40%, even small amounts of biased noise (5-10%) cause disproportionate damage to robustness ❽. This asymmetry highlights a critical vulnerability in ML systems deployed in real-world settings, where systematic biases in data collection or annotation could be more likely than purely random errors. The mechanisms underlying this vulnerability differ by architecture type: boosting models suffer from bias amplification as they focus on corrupted features, while transformer models show relative resilience through their attention mechanisms that may dynamically down-weight systematically corrupted inputs.

4.4 Complex Noise Patterns: Correlated, Concatenated, and Temporal

Across 15 datasets, we find that *correlated noise* tends to have only minor effects on overall performance, yet it can produce sharp fairness degradation in small but predictive subgroups, where even minimal overlaps can trigger disproportionate bias. In the College Scorecard dataset, where the intersecting minority group constitutes 6.49% of the data, the Equalized Odds ratio for the condition veteran status increases by over 200% as subgroup noise rates rise from 10% to 40%. *Concatenated noise*, by contrast, affects broader segments of the data and therefore results in visible performance declines and more diffuse fairness harms, especially when the union of affected subgroups covers more than half of the dataset. This is the case for the Income dataset, where a 68.98% union coverage causes a ∼4% drop in accuracy across models. *Temporal noise* introduces non-stationarity that steadily reduces performance while simultaneously producing fairness changes in specific time- or age-dependent cohorts. These effects may often be hidden if evaluation ignores temporal structure, and they can manifest in counterintuitive ways. Experiments across models reveal consistent relative performance dynamics similar to those observed under biased noise (Section 4.3). We added the full results and a detailed analysis on our GitHub repository.

Together, these findings show that accuracy alone is an unreliable proxy under complex noise and that fairness audits must explicitly test intersecting demographic conditions and temporal slices. To illustrate these dynamics more concretely, Figure 6 presents results for the Hospital Readmission dataset, which involves multiple sensitive attributes including race, gender, and age. Under correlated noise, race-based fairness disparities increase sharply due to overlap between the affected group (P_S = 12.2%) and racial minority prevalence (20.1%). In contrast, concatenated noise, which affects a broader segment of the dataset (P_S = 61.3%), leads to greater fairness degradation by gender. Finally, in the temporal noise setup, noise increases linearly with age. Counterintuitively, the younger age group (30–40) shows higher Equalized Odds disparity despite receiving less noise, due to miscalibrated false positive rates induced by label inflation among older patients.

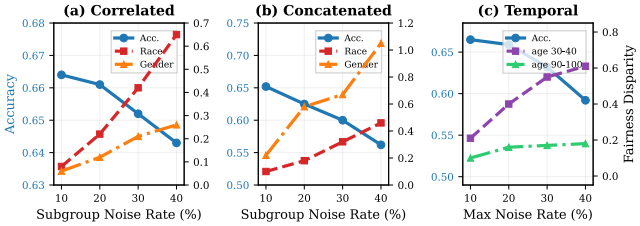
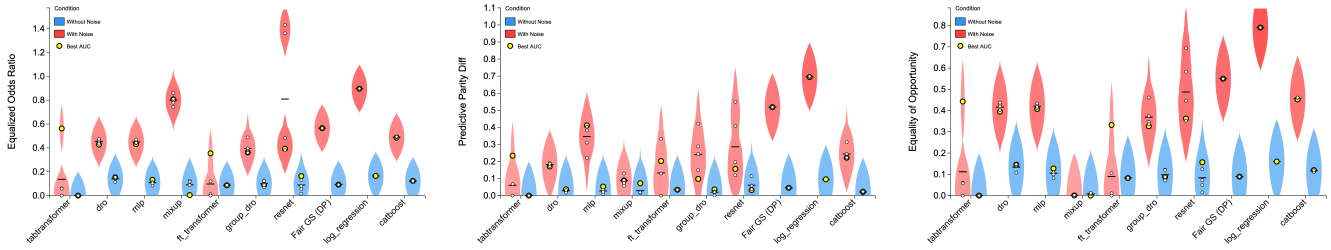


Figure 6: The impact of increasing noise on fairness and performance for (a) correlated noise (P_S = 12.2%), (b) concatenated noise (P_S = 61.3%), and (c) temporal noise in the Hospital Readmission dataset.



(a) Equalized Odds **(b) Predictive Parity** **(c) Equality of Opportunity**

Figure 7: Comparison of three fairness metrics across classifiers using the ACS Unemployment dataset. Each dot indicates one of the five top-performing trials, with yellow dots highlighting the best trial for each classifier.

While random noise can often be mitigated with robust loss functions or ensemble methods, complex noise patterns demand more targeted strategies [9]. Intersectional noise such as correlated or concatenated corruption may not be detectable through data cleaning approaches that treat protected attributes independently, underscoring the importance of fairness audits that explicitly consider intersecting demographic conditions. Temporal noise introduces another challenge: conventional cross-validation can mask fairness degradation if it ignores time structure, making time-aware validation and monitoring essential. Although fairness-aware models provide partial protection, our results indicate that they are less effective when the corruption mechanism itself is more complex. Addressing these scenarios requires fairness interventions that explicitly account for both the underlying noise process and the fairness constraints [10].

4.5 A closer look: ACS unemployment dataset

To illustrate the trends discussed in previous subsections, we examine fairness metrics for classifiers on the ACS Unemployment dataset. We use migration status as a second sensitive condition for complex noise patterns (4.39% coverage for correlated noise and 35.50% coverage for concatenated noise). For temporal noise, corruption increases with age. In clean conditions, transformer-based models (TabTransformer and FT Transformer) achieved near-perfect fairness (Equality of Opportunity ~ 0.0), while MixUp and Label Group DRO maintained low disparities (Equality of Opportunity $\sim 0.00 - 0.013$). MLP and ResNet demonstrated moderate fairness levels (Equalized Odds Ratio $\sim 0.084 - 0.164$), and Logistic Regression consistently showed higher baseline disparities ($\sim 0.160 - 0.165$). Under biased noise (30%) as shown in Figure 7, fairness degraded dramatically for most models. Logistic Regression and ResNet exhibited particularly severe deterioration (Equality of Opportunity increasing to $\sim 0.353 - 0.693$, with Logistic Regression reaching ~ 0.790). While TabTransformer and FT Transformer lost their perfect fairness, they maintained relatively moderate disparities (Demographic Parity Difference $\sim 0.0 - 0.144$). Label Group DRO, despite being designed for distributional robustness, experienced severe fairness degradation under biased noise (increases of $\sim 0.897 - 0.938$). Despite minimal change in accuracy ($\sim 97\%$ across conditions), fairness metrics diverged sharply under biased noise, revealing a hidden vulnerability in many models [11].

Under correlated noise, performance barely changes (e.g., −0.12% accuracy for CatBoost) but fairness collapses, with Predictive Parity and Equality of Opportunity for minority groups rising by up to 700%. Robust optimization models such as DRO and VREX exacerbate these issues (VREX reaches 47.9% Equalized Odds). Concatenated noise produces both performance loss (~7% AUC drop) and subgroup harms: FT Transformer maintains moderate disparities, while MixUp performs worst. Fairness-aware models help only when constraints match the corrupted attribute; otherwise, bias persists in unconstrained groups (e.g., +40% Equalized Odds for minority race) [12]. Temporal noise hides uneven fairness: older cohorts show doubled Equalized Odds, while younger cohorts see Predictive Parity drop to zero.

5 DISCUSSION

This section discusses the implications of our findings and highlights their utility for researchers and practitioners.

5.1 Utility of Fault Lines

Testing data cleaning pipelines. Our findings emphasize the need for tailored data management practices to address label noise, particularly when noise correlates with biases. Recent work by Guha et al. [33] shows that automated data cleaning can inadvertently compromise fairness, a risk possibly amplified by the biased noise types in our benchmark. Interestingly, Ni et al. [62] found that many data repair algorithms may inadvertently introduce more errors than they fix. We observe a similar counterintuitive result: in some cases, models trained on data with low levels of noise (e.g., <10%) show marginal performance gains over those trained on clean data (Section 4.2). Although these gains are not statistically significant, they challenge conventional assumptions about data quality and reinforces the need for careful evaluation of cleaning methods. Furthermore, model retraining workflows in AutoML pipelines may silently learn biased or brittle decision boundaries without any overt failure signal. Integrating our benchmark into such pipelines could help automatically surface when model performance or fairness metrics degrade disproportionately under simulated or detected noise patterns.

We observed systematic differences between random and biased noise, underscoring the limitations of generic cleaning approaches. While random noise can often be reduced with standard methods

(e.g., outlier removal), biased noise requires subgroup-aware interventions. Fault Lines offers a testbed to explore these dynamics rigorously.

Developing noise- and bias-robust models. Our results help identify model vulnerabilities across noise types, informing the design of more robust architectures. We identified specific vulnerabilities, such as high-capacity models' sensitivity to biased noise and fairness degradation under conditional corruption, which can inform model design priorities.

The benchmark facilitates systematic evaluation of novel robustness techniques against standardized noise patterns and fairness metrics. As researchers develop new methods for label noise robustness, our framework provides a consistent basis for comparing effectiveness across real-world scenarios. This is particularly valuable as the field shifts toward jointly optimizing performance and fairness under data corruption in data-centric AI.

5.2 Multi-class and Regression Tasks

We focused on binary classification to ensure methodological rigor in fairness evaluation, given the maturity of fairness metrics in this setting. However, the fundamental insights likely extend to other task types, albeit with increased complexity. Recent studies suggest that multi-class classification can be more susceptible to label noise due to increased class confusion from noise distribution across classes [22, 43]. In contrast, Mohammed et al. [59] show that regression models can be more robust to certain types of noise, including label noise. Further studies on noisy-label regression confirm that additive label noise drives error inflation [48, 85]. Fairness evaluation in multi-class classification typically relies on per-class or macro-averaged extensions of fairness metrics [10, 19]. Fairness in regression tasks is less mature due to continuous outcomes, often using residual-based measures like conditional demographic disparity [2, 8]. We provide a detailed guidance on how Fault Lines could be extended to multi-class classification and regression tasks, leveraging our existing noise taxonomy and evaluation framework in our GitHub repository. This extension redefines how noise is applied, adapts fairness metrics, and updates evaluation metrics to suit these task types.

5.3 Implications for practitioners

We show that data quality, model robustness, and fairness are tightly linked, especially in real-world AI deployments. As noted in Section 1, data quality profoundly influences AI systems: label errors can undermine reliability and amplify societal biases, particularly in high-stakes domains like healthcare and criminal justice.

Our results confirm these concerns, showing that noise type, dataset characteristics, and model choice fundamentally shape outcomes. While models demonstrate resilience to random noise, biased noise targeting specific subgroups significantly degrades robustness, with Equality of Opportunity showing mean differences of 0.216 ($p < 0.05$) and Demographic Parity differences up to 0.143 ($p < 0.05$). Dataset characteristics, such as feature complexity and class balance, further modulate these effects, with model architecture dictating trade-offs between robustness and equity. Group-agnostic noise may still create fairness disparities when class imbalance leads to unequal signal degradation.

The disproportionate impact of biased noise aligns with prior work on data quality's role in ML outcomes. Biased noise particularly harms fairness, amplifying Equalized Odds violations (mean differences up to 0.348, $p < 0.05$) for vulnerable subgroups. Unlike earlier studies, such as Northcutt et al. [64], which focused primarily on accuracy, our work bridges data quality to fairness, showing how subgroup-specific noise exacerbates disparities beyond random noise effects. For practitioners, these findings translate into concrete recommendations. In addition to motivating future research and methods, these findings provide actionable insights for practitioners, as summarized in Table 5. Crucially, we demonstrate that model-based solutions, while powerful, have inherent limitations.

Table 5: Recommendations for Practitioners Across Data Science Lifecycle Stages.

Stage	Recommendation	Ref.
Data Preparation	Validate labels with domain experts in high-stakes domains.	❷
	Iterate data cleaning and model tuning based on feedback.	❸
	Assess feature selection / dimensionality reduction for high-dimensional data.	❹
	Sample and review labels from sensitive subgroups to perform targeted subgroup label audits.	❽
	Prioritize data quality as model improvement cannot fully compensate for poor labels.	❾
	Understand your data and ensure ethical workflows by early stakeholder involvement.	⓫
Model Selection & Training	Mitigate random data entry errors using robust boosting variants with moderate regularization or robust loss functions.	❶
	Consider fairness-aware models with constraints, but expect trade-offs.	❻
	Evaluate transformer-based models, but monitor run-to-run variance.	❼
Evaluation & Deployment	Stress-test pipelines with simulated noise (e.g., using Fault Lines).	❺
	Monitor subgroup performance continuously.	❿
	Report predictive performance and multiple fairness metrics.	⓬

6 CONCLUSION

Our study advances fair data management by providing a comprehensive benchmark across 15 datasets and 22 models, revealing that while state-of-the-art ML models may demonstrate remarkable resilience to high levels of random label noise (up to 40%), they can exhibit pronounced vulnerability to even modest amounts of biased noise (5-10%). This framework integrates noise and fairness analysis to establish a foundation for developing noise-aware, fairness-preserving ML pipelines, providing actionable guidance for practitioners. Future work should integrate these findings into database management systems through automated noise detection mechanisms, fairness-aware optimization, and data quality monitoring pipelines that can proactively identify and mitigate diverse noise patterns during data ingestion and model training workflows.

REFERENCES

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. In *International Conference on Machine Learning (ICML '18)*. 60–69.

[2] Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. 2019. Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning (ICML '19)*. 120–129.

[3] Görkem Algan and Ilkay Ulusoy. 2020. Label noise types and their effects on deep learning. *arXiv:2003.10471* (2020).

[4] A. Ali, W. N. W. Ahmad, A. M. Mohamad, A. M. Zaki, and N. N. Hidayah. 2022. Issues, challenges and strategies in obtaining reliable and quality livelihood and wealth data across B40 community in Malaysia. *International Journal of Academic Research in Business and Social Sciences* 12, 8 (2022), 899–912.

[5] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv:1907.02893* (2019).

[6] Kathleen Barnes. 2019. Populations from under-represented backgrounds are not adequately represented in clinical databases. *The FASEB Journal* 33, S1 (2019), 217–1.

[7] Aki Barry, Lei Han, and Gianluca Demartini. 2023. On the impact of data quality on image classification fairness. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2225–2229.

[8] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. A convex framework for fair regression. *arXiv:1706.02409* (2017).

[9] Gérard Biau and Luc Devroye. 2015. *Lectures on the nearest neighbor method*. Vol. 246. Springer.

[10] Cody Blakeney, Gentry Atkinson, Nathaniel Huish, Yan Yan, Vangelis Metsis, and Ziliang Zong. 2022. Measuring bias and fairness in multiclass classification. In *2022 IEEE International Conference on Networking, Architecture and Storage (NAS)*. 1–6.

[11] Lou Therese Brandner, Philipp Mahlow, Anna Wilken, Annika Wölke, Hazar Harmouch, and Simon David Hirsbrunner. 2023. How data quality determines AI fairness: The case of automated interviewing. In *Proceedings of the European Workshop on Algorithmic Fairness*.

[12] J Michael Brick and Graham Kalton. 1996. Handling missing data in survey research. *Statistical methods in medical research* 5, 3 (1996), 215–238.

[13] Tom Burgert, Mahdyar Ravanbakhsh, and Begüm Demir. 2022. On the effects of different types of label noise in multi-label remote sensing image classification. *IEEE Trans. Geosci. Remote Sens.* 60 (2022), 1–13.

[14] Peter Byass. 2009. The unequal world of health data. *PLoS Medicine* 6, 11 (2009), e1000155.

[15] Centers for Disease Control and Prevention. 2021. Behavioral risk factor surveillance system (BRFSS) survey data. https://www.cdc.gov/brfss/annual_data/annual_data.htm Accessed August 23, 2025.

[16] Centers for Disease Control and Prevention (CDC) and National Center for Health Statistics (NCHS). 2018. National Health and Nutrition Examination Survey Data: Blood Lead Levels (2017–2018). U.S. Department of Health and Human Services, Centers for Disease Control and Prevention. Hyattsville, MD. Accessed August 23, 2025, from https://www.cdc.gov/nchs/nhanes/index.html.

[17] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. 785–794.

[18] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*. 797–806.

[19] Christophe Denis, Romuald Elie, Mohamed Hebiri, and François Hu. 2024. Fairness guarantees in multi-class classification with demographic parity. *Journal of Machine Learning Research* 25, 130 (2024), 1–46.

[20] Thomas G Dietterich. 2000. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning* 40 (2000), 139–157.

[21] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems* 34 (2021), 6478–6490.

[22] Yongliang Ding, Tao Zhou, Chuang Zhang, Yijing Luo, Juan Tang, and Chen Gong. 2022. Multi-class label noise learning via loss decomposition and centroid estimation. In *Proceedings of the 2022 SIAM International Conference on Data Mining*. 253–261.

[23] Mingyu Feng, Neil Heffernan, and Kenneth Koedinger. 2009. Addressing the assessment challenge with an online system that tutors as it assesses. *User modeling and user-adapted interaction* 19, 3 (2009), 243–266.

[24] FICO. 2019. The explainable machine learning challenge. https://community.fico.com/s/explainable-machine-learning-challenge. Accessed August 23, 2025.

[25] Ronald Aylmer Fisher. 1925. Theory of statistical estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, Vol. 22. Cambridge University Press, 700–725.

[26] Benoît Frénay and Michel Verleysen. 2013. Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems* 25, 5 (2013), 845–869.

[27] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of Machine Learning Research* 17, 59 (2016), 1–35.

[28] Josh Gardner, Zoran Popovic, and Ludwig Schmidt. 2023. Benchmarking distribution shift in tabular data with tableshift. *Advances in Neural Information Processing Systems* 36 (2023), 53385–53432.

[29] Pratyush Garg, John Villasenor, and Virginia Foggo. 2020. Fairness metrics: A comparative analysis. In *IEEE International Conference on Big Data (Big Data)*. 3662–3666.

[30] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. 2021. Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems* 34 (2021), 18932–18943.

[31] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. 2022. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems* 35 (2022), 507–520.

[32] Christoph Gröger. 2021. There is no AI without data. *Commun. ACM* 64, 11 (2021), 98–108.

[33] Shubha Guha, Falaah Arif Khan, Julia Stoyanovich, and Sebastian Schelter. 2024. Automated data cleaning can hurt fairness in machine learning-based decision making. *IEEE Transactions on Knowledge and Data Engineering* 36, 12 (2024), 7368–7379.

[34] Rosita Guido, Stefania Ferrisi, Danilo Lofaro, and Domenico Conforti. 2024. An overview on the advancements of support vector machine models in healthcare applications: A review. *Information* 15, 4 (2024), 235.

[35] Bo Han, Quanming Yao, Tongliang Liu, Gang Niu, Ivor W Tsang, James T Kwok, and Masashi Sugiyama. 2020. A survey of label-noise representation learning: Past, present and future. *arXiv:2011.04406* (2020).

[36] John T Hancock and Taghi M Khoshgoftaar. 2020. CatBoost for big data: An interdisciplinary review. *Journal of Big Data* 7, 1 (2020), 94.

[37] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems* 29 (2016), 3323–3331.

[38] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '16)*. 770–778.

[39] Martin Hirzel and Michael Feffer. 2023. A suite of fairness datasets for tabular classification. *arXiv:2308.00133* (2023).

[40] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. 2013. *Applied logistic regression*. John Wiley & Sons.

[41] Xin Huang, Ashish Khetan, Milan Cvitkovic, and Zohar Karnin. 2020. Tabtransformer: Tabular data modeling using contextual embeddings. *arXiv:2012.06678* (2020).

[42] Yvonne Hunter-Johnson, Tingting Liu, Kayon Murray, Yuanlu Niu, and Malinda Suprise. 2021. Higher education as a tool for veterans in transition: Battling the challenges. *The Journal of Continuing Higher Education* 69, 1 (2021), 1–18.

[43] Gaoxia Jiang, Jia Zhang, Xuefei Bai, Wenjian Wang, and Deyu Meng. 2024. Which is more effective in label noise cleaning, correction or filtering?. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 12866–12873.

[44] Ishan Jindal, Matthew Nokleby, and Xuewen Chen. 2016. Learning deep networks from noisy labels with dropout regularization. In *IEEE International Conference on Data Mining (ICDM)*. 967–972.

[45] David Johnson and Rosanna Scutella. 2003. *Understanding and improving data quality relating to low-income households*. Vol. 3. Melbourne Institute of Applied Economic and Social Research.

[46] Josh A Johnson, Brandon Moore, Eun Kyeong Hwang, Andy Hickner, and Heather Yeo. 2023. The accuracy of race & ethnicity data in US based healthcare databases: A systematic review. *The American Journal of Surgery* 226, 4 (2023), 463–470.

[47] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems* 30 (2017).

[48] Chris Dongjoo Kim, Sangwoo Moon, Jihwan Moon, Dongyeon Woo, and Gunhee Kim. 2024. Sample selection via contrastive fragmentation for noisy label regression. *Advances in Neural Information Processing Systems* 37 (2024), 127561–127609.

[49] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. 2021. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning (ICML '21)*. 5815–5826.

[50] Blake J Lawrence, Deborah Kerr, Christina M Pollard, Mary Theophilus, Elise Alexander, Darren Haywood, and Moira O'Connor. 2021. Weight bias among health care professionals: A systematic review and meta-analysis. *Obesity* 29, 11 (2021), 1802–1812.

[51] Michelle Seng Ah Lee. 2019. Context-conscious fairness in using machine learning to make decisions. *AI Matters* 5, 2 (2019), 23–29.

[52] Daphne Lenders and Toon Calders. 2023. Real-life performance of fairness interventions-introducing a new benchmarking dataset for fair ML. In *Proceedings of the ACM/SIGAPP Symposium on Applied Computing (SAC '23)*. 350–357.

[53] Junnan Li, Richard Socher, and Steven CH Hoi. 2020. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv:2002.07394* (2020).

[54] Siqi Liang, Jintao Huang, Junyuan Hong, Dun Zeng, Jiayu Zhou, and Zenglin Xu. 2023. FedNoisy: Federated noisy label learning benchmark. *arXiv:2306.11650* (2023).

[55] Yiqiao Liao and Parinaz Naghizadeh. 2023. Social bias meets data bias: The impacts of labeling and measurement errors on fairness criteria. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI '23)*, Vol. 37. 8764–8772.

[56] Jiaqi Luo, Yuedong Quan, and Shixin Xu. 2025. Robust-GBDT: leveraging robust loss for noisy and imbalanced classification with GBDT. *Knowledge and Information Systems* (2025), 1–21.

[57] Juan Martín, José A Sáez, and Emilio Corchado. 2021. On the regressand noise problem: Model robustness and synergy with regression-adapted noise filters. *IEEE Access* 9 (2021), 145800–145816.

[58] Douglas S Massey and Roger Tourangeau. 2013. Where do we go from here? Nonresponse and social measurement. *The Annals of the American Academy of Political and Social Science* 645, 1 (2013), 222–236.

[59] Sedir Mohammed, Lukas Budach, Moritz Feuerpfeil, Nina Ihde, Andrea Nathansen, Nele Noack, Hendrik Patzlaff, Felix Naumann, and Hazar Harmouch. 2025. The effects of data quality on machine learning performance on tabular data. *Information Systems* 132 (2025), 102549.

[60] Sedir Mohammed, Hazar Harmouch, and Felix Naumann. 2025. Step-by-step data cleaning recommendations to improve ML prediction accuracy. In *Proceedings of the International Conference on Extending Database Technology (EDBT '25)*. 542–554.

[61] Sendhil Mullainathan and Ziad Obermeyer. 2017. Does machine learning automate moral hazard and error? *American Economic Review* 107, 5 (2017), 476–480.

[62] Wei Ni, Xiaoye Miao, Xiangyu Zhao, Yangyang Wu, Shuwei Liang, and Jianwei Yin. 2024. Automatic data repair: Are we ready to deploy? *Proceedings of the VLDB Endowment* 17, 10 (2024), 2617–2630.

[63] Ginah Nightingale, Emily Skonecki, and Manpreet K Boparai. 2017. The impact of polypharmacy on patient outcomes in older adults with cancer. *The Cancer Journal* 23, 4 (2017), 211–218.

[64] Curtis G Northcutt, Anish Athalye, and Jonas Mueller. 2021. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv:2103.14749* (2021).

[65] Luke Oakden-Rayner. 2020. Exploring large-scale public medical image datasets. *Academic Radiology* 27, 1 (2020), 106–112.

[66] Luca Oneto and Silvia Chiappa. 2020. Fairness in machine learning. In *Recent Trends in Learning From Data*. Studies in Computational Intelligence, Vol. 896. 155–196.

[67] Joon Sung Park, Michael S Bernstein, Robin N Brewer, Ece Kamar, and Meredith Ringel Morris. 2021. Understanding the representation and representativeness of age in AI data sets. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*. 834–842.

[68] Dimitris Pavlopoulos and Jeroen K Vermunt. 2015. Measuring temporary employment. Do survey or register data tell the truth? *Survey Methodology* 41, 1 (2015), 197–214.

[69] John W Peabody, Jeff Luck, Sharad Jain, Dan Bertenthal, and Peter Glassman. 2004. Assessing the accuracy of administrative data in health information systems. *Medical Care* 42, 11 (2004), 1066–1072.

[70] Sergei Popov, Stanislav Morozov, and Artem Babenko. 2019. Neural oblivious decision ensembles for deep learning on tabular data. *arXiv:1909.06312* (2019).

[71] Alicja Rączkowska, Aleksandra Osowska-Kurczab, Jacek Szczerbiński, Kalina Jasinska-Kobus, and Klaudia Nazarko. 2024. AlleNoise: Large-scale text classification benchmark dataset with real-world label noise. *arXiv:2407.10992* (2024).

[72] Matthew A Reyna, Christopher S Josef, Russell Jeter, Supreeth P Shashikumar, M Brandon Westover, Shamim Nemati, Gari D Clifford, and Ashish Sharma. 2020. Early prediction of sepsis from clinical data: The PhysioNet/Computing in Cardiology Challenge 2019. *Critical care medicine* 48, 2 (2020), 210–217.

[73] Yuji Roh, Geon Heo, and Steven Euijong Whang. 2019. A survey on data collection for machine learning: A big data-ai integration perspective. *IEEE Transactions on Knowledge and Data Engineering* 33, 4 (2019), 1328–1347.

[74] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. 2019. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv:1911.08731* (2019).

[75] Bernhard Schölkopf and Alexander J Smola. 2002. *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT press.

[76] Vraj Shah, Thomas Parashos, and Arun Kumar. 2024. How do categorical duplicates affect ML? A new benchmark and empirical analyses. *Proceedings of the VLDB Endowment* 17, 6 (2024), 1391–1404.

[77] Nima Shahbazi, Yin Lin, Abolfazl Asudeh, and HV Jagadish. 2023. Representation bias in data: A survey on identification and resolution techniques. *Comput. Surveys* 55, 13s (2023), 1–39.

[78] Samuel Sanford Shapiro and Martin B Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika* 52, 3-4 (1965), 591–611.

[79] Ravid Shwartz-Ziv and Amitai Armon. 2022. Tabular data: Deep learning is not all you need. *Information Fusion* 81 (2022), 84–90.

[80] Gowthami Somepalli, Micah Goldblum, Avi Schwarzschild, C Bayan Bruss, and Tom Goldstein. 2021. Saint: Improved neural networks for tabular data via row attention and contrastive pre-training. *arXiv:2106.01342* (2021).

[81] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. 2022. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems* 34, 11 (2022), 8135–8153.

[82] Julia Stoyanovich, Bill Howe, and H. V. Jagadish. 2020. Responsible data management. *Proceedings of the VLDB Endowment* 13, 12 (2020), 3474–3488.

[83] Beata Strack, Jonathan P DeShazo, Chris Gennings, Juan L Olmo, Sebastian Ventura, Krzysztof J Cios, and John N Clore. 2014. Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. *BioMed research international* 2014, 1 (2014), 781670.

[84] American National Election Studies. 2020. ANES time series cumulative data file, 1948-2020. https://electionstudies.org Accessed August 23, 2025.

[85] Qiang Sun, Wen-Xin Zhou, and Jianqing Fan. 2020. Adaptive huber regression. *J. Amer. Statist. Assoc.* 115, 529 (2020), 254–265.

[86] Hind Taud and Jean-Francois Mas. 2017. Multilayer perceptron (MLP). In *Geomatic approaches for modeling land change scenarios*. 451–455.

[87] John W Tukey. 1949. Comparing individual means in the analysis of variance. *Biometrics* (1949), 99–114.

[88] U.S. Department of Education. 2025. College scorecard data. https://collegescorecard.ed.gov. Institution- and field-of-study-level datasets; Accessed August 23, 2025.

[89] Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge Belongie. 2017. Learning from noisy large-scale datasets with minimal supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '17)*. 839–847.

[90] Carlos-José Villagrá-Arnedo, Francisco J Gallego-Durán, Patricia Compañ, Faraón Llorens Largo, Rafael Molina-Carmona, et al. 2016. Predicting academic performance from behavioural and learning data. *International Journal of Design & Nature and Ecodynamics* 11, 3 (2016), 239–249.

[91] Jialu Wang, Yang Liu, and Caleb Levy. 2021. Fair classification with group-dependent label noise. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. 526–536.

[92] Shirly Wang, Matthew BA McDermott, Geeticka Chauhan, Marzyeh Ghassemi, Michael C Hughes, and Tristan Naumann. 2020. Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii. In *Proceedings of the ACM Conference on Health, Inference, and Learning (CHIL '20)*. 222–235.

[93] Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. 2021. Learning with noisy labels revisited: A study using real-world human annotations. *arXiv:2110.12088* (2021).

[94] Tingting Wu, Xiao Ding, Minji Tang, Hao Zhang, Bing Qin, and Ting Liu. 2023. NoisywikiHow: A benchmark for learning with real-world noisy labels in natural language processing. In *Findings of the Association for Computational Linguistics (ACL '23)*. 4856–4873.

[95] Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. 2025. Data-centric artificial intelligence: A survey. *Comput. Surveys* 57, 5 (2025), 1–42.

[96] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2017. mixup: Beyond empirical risk minimization. *arXiv:1710.09412* (2017).