Still More Shades of Null: An Evaluation Suite for Responsible Missing Value Imputation

Falaah Arif Khan*
fa2161@nyu.edu
New York University
New York, USA

Denys Herasymuk*
herasymuk@ucu.edu.ua
Ukrainian Catholic University
Lviv, Ukraine

Nazar Protsiv
protsiv.pn@ucu.edu.ua
Ukrainian Catholic University
Lviv, Ukraine

Julia Stoyanovich
stoyanovich@nyu.edu
New York University
New York, USA

ABSTRACT

Data missingness is a practical challenge of sustained interest to the scientific community. In this paper, we present `Shades-of-Null`, an evaluation suite for responsible missing value imputation. Our work is novel in two ways (i) we model realistic and socially-salient missingness scenarios that go beyond Rubin's classic Missing Completely at Random (`MCAR`), Missing At Random (`MAR`) and Missing Not At Random (`MNAR`) settings, to include multi-mechanism missingness (when different missingness patterns co-exist in the data) and missingness shift (when the missingness mechanism changes between training and test) (ii) we evaluate imputers holistically, based on imputation quality and imputation fairness, as well as on the predictive performance, fairness and stability of the models that are trained and tested on the data post-imputation.

We use `Shades-of-Null` to conduct a large-scale empirical study involving 29,736 experimental pipelines, and find that while there is no single best-performing imputation approach for all missingness types, interesting trade-offs arise between predictive performance, fairness and stability, based on the combination of missingness scenario, imputer choice, and the architecture of the predictive model. We make `Shades-of-Null` publicly available, to enable researchers to rigorously evaluate missing value imputation methods on a wide range of metrics in plausible and socially meaningful scenarios.

PVLDB Reference Format:
Falaah Arif Khan, Denys Herasymuk, Nazar Protsiv, and Julia Stoyanovich. Still More Shades of Null: An Evaluation Suite for Responsible Missing Value Imputation. PVLDB, 18(9): 2899 - 2913, 2025.
doi:10.14778/3746405.3746416

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

*Co-first authors.

Proceedings of the VLDB Endowment, Vol. 18, No. 9 ISSN 2150-8097.
doi:10.14778/3746405.3746416

1 INTRODUCTION

As AI becomes more widely deployed into society, data — most importantly, openly accessible high quality AI-ready data — becomes a precious shared commodity. Among the factors affecting data quality is data missingness, a prevailing practical challenge of sustained interest to the data management, statistics and data science communities, and to the scientific community writ large [32, 37, 39, 55, 58, 69, 74, 79, 84, 97, 102].

Debates on handling missing values in data management date back to the field's inception, with classic discussions such as Date [15]. At the operational level, missing values are typically denoted by `null`, but hidden missing values can exist (e.g., 'AL' being selected by default in a job application). At the semantic level, `null` can have multiple meanings—unknown, inapplicable, or intentionally withheld. This paper does not engage in the semantic debate or consider hidden missing values [76]. Instead, we focus on a specific case: a dataset X (a single relation) where some features are missing, marked by `null`, indicating that the feature has a real-world value but is unobserved in X. Our goal is to use X in a machine learning (ML) setting, either for model training or inference. Since ML models cannot handle `null` directly, missing values must be imputed as part of data preprocessing.

As our starting point, we will use Rubin's missingness framework [79] that, nearly 50 years since it was proposed, still remains the most popular approach to modeling missing data. Consider a dataset X of n samples, each with p features, and an indicator R such that $R_{i,j} = 1$ when the value of the j's feature of X_i is missing: $X_{i,j}$ is null, and $R_{i,j} = 0$ when that the value is observed: $X_{i,j}$is not null. Rubin identified three data missingness scenarios:

*Missing Completely at Random (**MCAR**).* In a job applicant dataset with salary and years of experience, `MCAR` holds if salary is missing due to administrative errors, unrelated to the salary itself or work experience. That is: $P(R|X) = P(R)$.

*Missing at Random (**MAR**).* If job applicants with fewer years of experience are more likely to withhold their salary, and this can be explained by observed covariates (i.e., years of experience), then `MAR` holds. Here, missingness depends only on observed features, not the missing values themselves: $P(R|X) = P(R|X_{\text{obs}})$.

*Missing Not at Random (**MNAR**).* Consider a job applicant whose salary depends on geographic location and skills test results—neither

captured in the data—rather than years of experience. Suppose applicants with lower salaries are more likely to withhold this information, hoping for a higher offer. In this case, MNAR holds because missingness is correlated with the missing value itself and *cannot* be explained by observed covariates (i.e., years of experience): $P(R|X) \neq P(R|X_{\text{obs}})$.

*Missing value imputation (**MVI**).* Rubin's framework has shaped a vast body of work on missing value imputation, extensively reviewed in several comprehensive surveys [2, 4, 20, 30, 35, 39, 44, 56, 57, 65, 73–75, 102]. MVI methods fall into three main categories: (1) *Statistical methods*, such as median or mode imputation [81]; (2) *Learning-based impute-then-classify*, which iteratively impute missing values using k-nearest neighbors [6], clustering [28], decision trees [90], or ensembles [87]; (3) *Joint data cleaning and model training*, integrating imputation with model learning [47, 51, 52], based on Rubin's multiple imputation framework [80].

Beyond Rubin's framework: Mixing scenarios and dealing with missingness shift. Rubin's framework, while analytically clean, does not fully capture real-world missingness. First, MCAR assumptions rarely hold, and real-world data often falls on a continuum between MAR and MNAR, depending on collection methods [32]. Second, missingness mechanisms frequently *co-exist within a dataset* (affecting different features or tuples), leading to *multi-mechanism missingness* [102]. For instance, Mitra et al. [69] introduce the *data missingness life cycle*, showing how data integration from diverse sources creates *structured missingness* beyond Rubin's model. Third, in *data-centric AI*, missingness assumptions valid during training may shift post-deployment, a phenomenon termed *missingness shift*, analogous to data distribution shift [101].

Missingness as a form of bias. Consider the job applicant screening example with gender and age as features. Female applicants who suspect wage discrimination may withhold salary information more often than men, hoping to narrow the gender pay gap. This leads to more missing salary values for women, where missingness depends on the observed covariate (gender), aligning with MAR. This reflects *pre-existing bias*, where data encodes historical societal discrimination [26]. For another example, suppose disability status is included as a feature. Applicants with disabilities may be more likely to omit this information. If disability status is uncorrelated with other features, this scenario aligns with MNAR, with missingness itself acting as a proxy for disadvantage.

When handling missing values, data scientists must also address *technical bias* [26], where incorrect technical choices create disparities in predictive accuracy, often amplifying pre-existing bias. A key example is imputing missing values under incorrect assumptions, which can worsen disparities in classifier performance [34, 82, 83, 89]. For instance, if job applicant salaries are missing under MAR or MNAR (e.g., older women withhold salaries due to perceived discrimination), imputing them under MCAR could further depress salary estimates, reinforcing the gender wage gap and ageism, and leading to discriminatory outcomes.

Missing value imputation can impact model arbitrariness. Missingness is an indication of uncertainty in the data. MVI methods "resolve" this uncertainty at the tuple level, but they may induce a change in the data distribution in ways that impacts the stability of predictions of a model trained on this data. In some cases, the resulting models produce vastly different — and even *arbitrary* — predictions under small perturbations in the input [12, 13, 77, 78]. For example, if a job applicant's salary is imputed in vastly different ways upon two consecutive applications for the same position, and this, in turn, impacts the hiring decision, then the decision-making process violates the principle of process fairness (e.g., [1, 91]). Importantly, instability and accuracy are orthogonal: models can be accurate in expectation while still being unstable [61].

Research gap. Despite numerous MVI techniques being proposed each year, there has been limited systematic progress in assessing them across key performance aspects, including imputation correctness, predictive accuracy, and fairness—measured as disparities in imputation quality or model performance across groups. Moreover, while missingness signals uncertainty, there has been no comprehensive evaluation of the *stability* of models trained on cleaned data. Crucially, realistic modeling of missingness, identifying bias sources, and selecting appropriate stakeholder groups and fairness metrics must be grounded in the specific context of use [27, 53, 68, 72]. For instance, age-based discrimination is relevant in both hiring and lending, yet older applicants face disadvantages (and legal protections) in hiring, while younger applicants are disadvantaged in lending. Thus, MVI techniques must be evaluated in societally meaningful scenarios.

Summary of contributions. We implemented an experimental benchmark called Shades-of-Null to rigorously and comprehensively evaluate state-of-the-art MVI techniques on a variety of realistic missingness scenarios (including single- and multiple-mechanism missingness and missingness shift), on a suite of evaluation metrics (including fairness and stability), in the context of data preprocessing in a machine learning pipeline.

Our work is (1) *novel*: to the best of our knowledge, the settings of multi-mechanism missingness and missingness shifts have not been empirically studied before; (2) *comprehensive*: we evaluate a suite of 15 MVI techniques on 7 benchmark datasets using 6 model types, running a total of **29,736** pipelines, and is the first study of such scale in the missing data domain, to the best of our knowledge; (3) *normatively grounded*: we focus on decision-making contexts such as lending, hiring, and healthcare, where missingness is socially salient. Mitigating social harm such as algorithmic discrimination is a leading concern in these domains [5], and we evaluate the impact of MVI approaches on downstream model fairness and stability (which have been understudied in the context of missing data), in addition to classically studied imputation quality and model correctness metrics.

While developing the Shades-of-Null evaluation suite, we found and fixed several bugs in existing MVI implementations, including data leakage and omitted hyperparameter tuning. See full version of the paper for details [48]. We make Shades-of-Null publicly available and hope to enable researchers to comprehensively evaluate new MVI methods on a wide range of evaluation metrics, under plausible and socially meaningful missingness scenarios.

2 RELATED WORK

Missing value imputation techniques. Learning-based approaches have become increasingly popular, and include k-nearest neighbors, decision trees, support vector machines, clustering, and ensembles [6, 35, 57]. Zhou et al. [102] and Liu et al. [60] review deep learning-based approaches (variational auto-encoders and generative adversarial networks) and representation learning (graph neural networks and diffusion-based methods). Multiple imputation [81, 102] and expectation maximization [74, 92] are also influential, but too computationally expensive to be popular in practice [4].

MNAR-specific techniques, like not-MIWAE [40] and GINA [62], tackle the challenge of MNAR data by employing identifiable generative models that effectively account for complex missingness mechanisms. Recent methods, including NOMI [93] and TDM [100], introduce advancements like uncertainty-driven networks and transformed distribution matching, which enhance both imputation accuracy and computational efficiency.

Beyond impute-then-classify, the data management community has proposed holistic methods like CPClean [47] and ActiveClean [52], that jointly perform data cleaning and model training, deriving from the multiple imputation framework [80]. These methods detect and repair a variety of errors including outliers, mislabels, duplicates, and missing values, and hence are less directly optimized to model missingness, instead focusing on improving data quality holistically. BoostClean [51] aims to reduce the human effort in error repair by learning efficiently from a few gold standard annotations (from a human oracle).

Evaluating MVI *techniques.* We are aware of several surveys of MVI techniques, all conducted with a strong empirical focus [4, 21, 57, 67, 85]. Miao et al. [67] compare 19 MVI methods on 15 datasets, and while our results corroborate their findings (see Section 5.1), their evaluation is limited to imputation quality and overall accuracy (but not fairness or stability). Other empirical studies have been primarily focused on medical datasets, and only evaluate missingness under MCAR [4, 21, 57, 85]. Further, most proposed methods only evaluate imputation quality, using metrics such as MAE, MSE, RMSE, and AUC [35, 44], although some also evaluate overall predictor accuracy [57]. Additionally, the performance of MVI techniques under multi-mechanism missingness [102] and missingness shifts [101] remains unexplored in prior work, despite these conditions being more likely to occur in practice due to distribution shifts in production deployments [32].

Notably, overwhelming evidence in the literature indicates that there is no single "best-performing" MVI approach on accuracy [21, 30, 35, 57, 84], and that model performance (narrowly measured based on 'correctness' thus far) depends on dataset characteristics such as size and correlation between variables [4] and missingness rates in the train and test sets [57, 84].

Fairness and missingness. There has been some recent interest in studying the social harm that can come from poorly chosen MVI techniques [10, 25, 42, 63, 94, 97–99]. Most empirical studies [10, 43, 94, 98, 99] have worked with the COMPAS [54] and Adult [18] datasets, the latter of which has been "retired" from community use due to issues with provenance [16]. Further, these experimental studies employ randomly-generated missingness: usually by randomly sampling or using a fixed set of columns, and randomly picking rows in which to replace values with `null`. We critique this approach, since detecting and mitigating unfairness requires broader socio-technical thinking, such as having higher rates of missingness for minority groups and in features that are highly correlated with sensitive attributes (called *proxy* variables in the fairness literature) [10].

A notable exception is Martínez-Plumed et al. [63], who map social mechanisms such as prejudicial access and self-reporting bias to missingness categories like missing-by-design and item nonresponse. They also analyze feature correlations to study the effects of different missingness types. We adopt a similar methodology to simulate realistic missingness in this work but identify conceptual limitations in their fairness framing. The authors state: "The surprising result was to find that, [...] the examples with missing values seem to be fairer than the rest." **However, asserting that some rows of data are more or less fair is misguided**, as fairness is not a property of individual samples (e.g., job applicants) but of the model (e.g., in hiring decisions), which determines fairness through inclusion or exclusion in positive outcomes. We reinterpret their findings to suggest that excluding samples with missing values can increase model unfairness, reinforcing the case against deletion as a missing data strategy.

Zhang and Long [98] evaluate MVI methods on *imputation fairness*, defined as the difference in imputation accuracy between privileged and disadvantaged groups. They find that imputation unfairness increases with higher missingness disparity, higher overall missingness rates, and greater data imbalance across groups. Further, they find that varying missingness mechanisms for the same imputation method impacts prediction fairness. Their analysis is limited to randomly generated `null` values in COMPAS. We extend their work to additional datasets, missingness scenarios, and alternative imputation fairness definitions.

In a follow-up work, Zhang and Long [99] introduce *imputation fairness risk* and provide bounds for "correctly specified" imputation methods. While this is a commendable theoretical contribution in a largely unexplored area, we question its broader implications: imputation quality metrics do not fully capture downstream model performance [97]. In other words, a classifier can perform well despite poor imputation quality [84]. This raises a key question: Does minimizing imputation unfairness reduce model unfairness? Our empirical findings suggest it does not, as discussed in Section 4.4.

Finally, Jeong et al. [43] propose a decision tree-based method that integrates fairness into model training while handling missing values. Their approach splits only on observed values to mitigate disparities introduced by imputation. Their evaluation is limited to MCAR scenarios (with more missingness for disadvantaged groups). In contrast, we assess more advanced MVI techniques under diverse missingness scenarios (MCAR, MAR, MNAR, and missingness shift) without applying fairness interventions. Nonetheless, we share their broader motivation of assessing and mitigating unfairness holistically throughout the data lifecycle. Future work could explore different combinations of MVI and fairness interventions.

Missingness and stability. We are not aware of any work investigating the effect of missing value imputation on model stability.

3 BENCHMARK OVERVIEW

3.1 Methodology for Simulating Missingness

We start with datasets in which there are no `null` values, and then simulate missingness. We make this choice because we are interested in comparing `MVI` performance under single-mechanism versus multi-mechanism missingness, and under missingness shifts, and, to the best of our knowledge, there are no datasets with naturally-occurring documented missingness of this form.

Our methodology for simulating missingness is based on *evaluation scenarios*, defined by the missingness mechanism during training and testing, shown in Table 1: (1) single-mechanism missingness, injected similarly into train and test sets (S1 - S3); (2) single-mechanism missingness, injected differently into train and test sets (missingness shift) (S4 - S9); and (3) missingness is mixed, to include all three missingness mechanisms, and is injected similarly into train and test sets (S10).

For each dataset in our study (see Section 3.5), we designed socially-salient missingness scenarios corresponding to the three missingness mechanisms (`MCAR`, `MAR`, `MNAR`). Following [44, 63], we identified features for missing value injection, denoted by $\mathcal{F}^m$, based on their Spearman correlation with the target variable and feature importance scores computed using scikit-learn. These selected features were chosen to reflect plausible missingness patterns. For instance, in the `diabetes` dataset, while features like blood pressure or cholesterol levels are expected to be consistently observed, others, such as family history or physical activity, might be omitted or withheld due to privacy concerns or reporting biases. The remaining features, denoted by $\mathcal{F}^c$, were considered complete, with no missing values.

The three missing mechanisms share the same set of selected features ($\mathcal{F}^m$), but differ in their injection strategies. For `MCAR`, the missing values are randomly injected on $\mathcal{F}^m$. In contrast, the missingness of `MAR` is based on sensitive attributes within $\mathcal{F}^c$ to simulate pre-existing bias, as described in Section 1. Specifically, higher rates of missingness were injected to disadvantaged groups wherever possible (in some cases there were too few samples from disadvantaged groups), reflecting realistic disparities caused by unequal access, distrust, or procedural injustice [3]. Finally, for `MNAR`, the missingness is determined by missing values themselves, and the likelihood of missing values depends on the missing features.

Table 3 presents the selected columns ($\mathcal{F}^m$) and injection conditions for the `diabetes` dataset, based on the correlation coefficients and feature importance values in Figure 2. Additional information on other datasets is available in full version [48].

3.2 Missing Value Imputation (`MVI`) Techniques

As discussed in Section 2, many competitive `MVI` techniques have been proposed. We selected 15 of them, from 8 broad categories based on taxonomies presented in [20, 41, 67, 93], namely: (1) `deletion`; (2) statistical: `median-mode` and `median-dummy`; (3) machine learning-based: `miss-forest` [88] and `clustering` [28]; (4) discriminative deep learning-based: `datawig` [7] and `auto-ml` [41]; (5) generative deep learning-based: `gain` [96] and `hi-vae` [71]; (6) MNAR-specific: `not-miwae` [40] and `mnar-pvae` [62]; (7) multiple imputation: `boostclean` [51]; and (8) other recent: `nomi` [93], `tdm` [100], and `edit-gain` [66]. See full version [48] for details.

Table 1: Evaluation Scenarios

	Train			Test		
Scenario	MCAR	MAR	MNAR	MCAR	MAR	MNAR
S1	✓			✓		
S2		✓			✓	
S3			✓			✓
S4	✓				✓	
S5	✓					✓
S6		✓		✓		
S7		✓				✓
S8			✓	✓		
S9			✓		✓	
S10	✓	✓	✓	✓	✓	✓

3.3 Evaluation Metrics

Following [35, 57], we evaluate `MVI` techniques in two ways: directly using imputation quality metrics and indirectly based on downstream model performance.

3.3.1 Imputation Quality. Shadbahr *et al.* [84] report that distributional metrics capture downstream model performance better than classically-used discrepancy metrics. To confirm or refute this claim, we use a mix of both. To assess agreement with true values, we compute *Root Mean Square Error (RMSE)* for numerical features and *F1 score* for categorical features. To assess distributional alignment, we compute *KL-divergence* (i.e., the Shannon entropy) between the true and the predicted feature distributions, for both numerical and categorical features, measured for the imputed columns only as well as for the full dataset. For categorical features, we obtain the probability distributions using the `value_counts` method with normalization from `pandas`. For numerical features, we use Gaussian kernel density estimation from `scipy`, with 1000 samples. Finally, to assess imputation fairness [98, 99], we compute *F1 score difference*, *RMSE difference*, and *KL divergence difference* between privileged (*priv*) and disadvantaged (*dis*) groups.

3.3.2 Model Performance. To assess the impact of `MVI` techniques on model correctness, we report the *F1 score* because it is a more reliable metric than accuracy for imbalanced data.

For evaluating model stability, we report average *Label Stability* [14, 49] over the full test set (closely related to the self-consistency metric from Cooper et al. [11]), computed per-sample for binary classification as *Label Stability* $= \frac{|B_+ - B_-|}{B}$, where B_+ is the number of times the sample is classified into the positive class and B_- is the number of times the sample is classified into the negative class, and $B = B_+ + B_-$ models are trained by bootstrapping over the train set. We set $B = 50$ in all our experiments.

Lastly, we report model fairness based on group-specific error rates, namely *True Positive Rate Difference (TPRD)*, *True Negative Rate Difference (TNRD)*, *Selection Rate Difference* (SRD), and *Disparate Impact (DI)*. (Note that DI computes the ratio of selection rates, but we refer to it as DI as is standard in the literature [23].)

Fairness metrics based on error rates align with formal equality of opportunity, while those based on selection rates (SRD, DI) reflect substantive equality [50]. The choice of metric depends on the

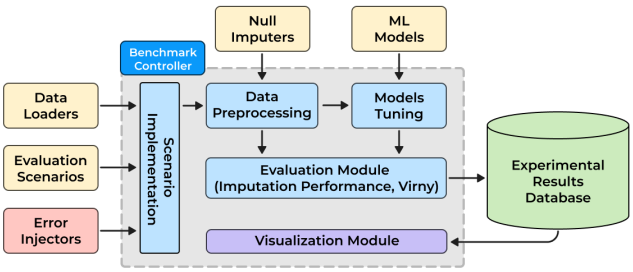


Figure 1: `Shades-of-Null` architecture

domain and stakeholder [70]. SRD captures absolute disparities (e.g., fixed quotas in college admissions), while DI measures relative disparities (e.g., the 4/5th rule in U.S. hiring). From the individual's perspective, TPRD suits opportunity allocation (e.g., hiring, loans), ensuring access to positive outcomes, whereas TNRD applies to exclusion decisions (e.g., medical diagnoses), emphasizing fairness in avoiding false negatives.

3.4 `Shades-of-Null` Architecture

The architecture of `Shades-of-Null` is shown in Figure 1. Its core component, the *benchmark controller*, executes user-specified missingness scenarios by applying error injectors to input datasets. It then imputes missing values using selected `MVI` technique(s) and preprocesses data via standard scaling (numerical) and one-hot encoding (categorical), and then trains ML models with hyperparameter tuning. The evaluation module assesses imputation quality and model performance. For comprehensive profiling, it uses Virny [36], a Python library that computes accuracy, stability, and fairness metrics across multiple sensitive attributes and their intersections.

`Shades-of-Null` incorporates two optimizations to enhance experimental efficiency. First, it decouples missing value imputation from model training, allowing imputed datasets to be stored and reused in subsequent training and evaluation stages. Second, it supports simultaneous evaluation on multiple test sets (e.g., with varying missingness rates or types), significantly reducing running time, and so executing a pipeline with one training set and multiple test sets takes about the same time as with a single test set.

3.5 Datasets and Tasks

As noted in Section 1, we focus on *socially salient* missingness. With this in mind, we selected seven datasets from diverse social decision-making contexts, including lending, hiring, marketing, admissions, and healthcare, summarized in Table 2. Each dataset involves a binary classification task, where a positive label represents access to a desirable social good (e.g., education, employment, or healthcare). We chose these datasets to ensure broad coverage of (i) social domains, (ii) dataset sizes, and (iii) numerical-to-categorical column ratios.

Table 2: Dataset Information

name	domain	# tuples	# attrs	sensitive attrs
`diabetes`	healthcare	952	17	sex
`german`	finance	1,000	21	sex, age
`folk-income`	finance	15,000	10	sex, race
`law-school`	education	20,798	11	sex, race
`bank`	marketing	40,004	13	age
`heart`	healthcare	70,000	11	sex
`folk-employment`	hiring	302,640	16	sex, race

3.6 Model Types

We evaluate predictive performance of 6 ML models: (i) decision tree (`dt_clf`) with a tuned maximum tree depth, minimum samples at a leaf node, number of features used to decide the best split, and criteria to measure the quality of a split; (ii) logistic regression (`lr_clf`) with tuned regularization penalty, regularization strength, and optimization algorithm; (iii) gradient boosted trees (`lgbm_clf`) with tuned number of boosted trees, maximum tree depth, maximum tree leaves, and minimum number of samples in a leaf; (iv) random forest (`rf_clf`) with a tuned number of trees, maximum tree depth, minimum samples required to split a node, and minimum samples at a leaf node (v) neural network, historically called the multi-layer perceptron (`mlp_clf`) with two hidden layers, each with 100 neurons, and a tuned activation function, optimization algorithm, and learning rate; (vi) a deep table-learning method called GANDALF [46] (`gandalf_clf`) with a tuned learning rate, number of layers in the feature abstraction layer, dropout rate for the feature abstraction layer, and initial percentage of features to be selected in each Gated Feature Learning Unit (GFLU) stage. Search grids of hyperparameters for all models are defined in our codebase.

4 SINGLE AND MULTI-MECHANISM MISSINGNESS

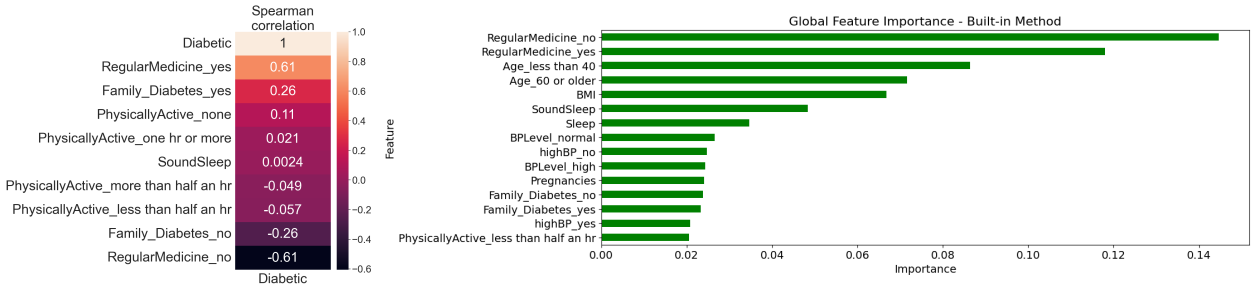
To simulate single-mechanism missingness (S1-S3 in Table 1) we inject 30% of each training and test sets with nulls, according to the missingness scenarios described in Section 3.1. For multi-mechanism or mixed missingness, when `MCAR`, `MAR` and `MNAR` coexist (S10 in Table 1), we inject 10% of nulls for each of the three mechanism into both training and test sets, for a total of 30% nulls.

To evaluate model correctness, we report results for F1, see full version [48] for accuracy results. For fairness, we use binary group definitions. For datasets with two sensitive attributes, we define the doubly-disadvantaged group as disadvantaged (*dis*) and everyone else as privileged (*priv*). For example: on the `law-school`, `folk-income` and `folk-employment` datasets, non-White women are the *dis* group, and White women, non-White men and White men are the *priv* group. We report results for TPRD, see results for other fairness metrics in the full version of the paper [48]. For stability, we used a bootstrap of 50 estimators, each seeing a random 80% of the training set [19]. Higher values of F1 and label stability are better, and values of TPRD close to zero are better.

Different models are the best-performing on different datasets. In Figures 3, 4 and 5, we report on the best-performing models (according to F1) for five most representative datasets per experiment, and compare performance against a model trained on clean data. Complete results are available in the full version [48].

Table 3: Missingness scenarios for an error rate of 30% for `diabetes`. SoundSleep is a numerical column; Family_Diabetes, PhysicallyActive and RegularMedicine are categorical columns.

Mechanism	Missing Column ($\mathcal{F}^m$)	Conditional Column (I)	Pr($\mathcal{F}^m$ | I is dis)	Pr($\mathcal{F}^m$ | I is priv)
MCAR	SoundSleep, Family_Diabetes, PhysicallyActive, RegularMedicine	N/A	0.3	0.3
MAR	Family_Diabetes, RegularMedicine PhysicallyActive, SoundSleep	Sex Age	0.2 (female) 0.2 ($\geq$ 40)	0.1 (male) 0.1 ($<$ 40)
MNAR	Family_Diabetes RegularMedicine PhysicallyActive SoundSleep	Family_Diabetes RegularMedicine PhysicallyActive SoundSleep	0.25 (yes) 0.2 (yes) 0.25 (none, $<\frac{1}{2}$ hour) 0.2 ($<$ 5)	0.05 (no) 0.1 (no) 0.05 ($>\frac{1}{2}$ hour, $>$ 1 hour) 0.1 ($\geq$ 5)



(a) Correlation with label

(b) Feature importance

Figure 2: EDA for designing missingness scenarios in `diabetes`.

4.1 Correctness of the Predictive Model

Figure 3 shows the F1 of models trained with different `MVI` techniques. We find interesting trends in `MVI` performance based on characteristics of the dataset and missingness type. All techniques are competitive for all missingness mechanisms, including mixed missingness, on `heart` and `law-school`. `boostclean`, which uses multiple imputation (MI), is otherwise only competitive on small datasets (`diabetes` and `german`), and only under `MCAR` and mixed missingness on `german`. None of the `MVI` techniques are able to match the F1 of the model trained on clean data on `folk-employment`, and this effect is strongest under `MNAR` (notably, stronger than under mixed missingness). `boostclean` shows particularly poor performance on `folk-income`, with a 0.08 decrease in F1 compared to other methods, for all missingness types. We discuss this unexpected performance of MI further in Section 7.

`auto-ml`, `datawig` and `miss-forest` are generally the best performing `MVI` techniques, with `nomi` a close second, offering an optimal balance between imputation accuracy and training time (see the full paper [48] for training time analysis). However, simpler statistical techniques (e.g., `median-mode` under `MAR`) are also competitive. This underscores the need to evaluate novel DL-based and ML-based methods holistically (e.g., on a variety of missingness scenarios) to ensure that they justify the additional training overhead and complexity they introduce compared to simple methods.

Interestingly, `not-miwae` and `mnar-pvae` do not demonstrate superior performance compared to other methods in our socially salient MNAR scenario. Instead, their performance aligns closely with other leading MVI approaches under MNAR conditions. This finding is further discussed in Section 7.

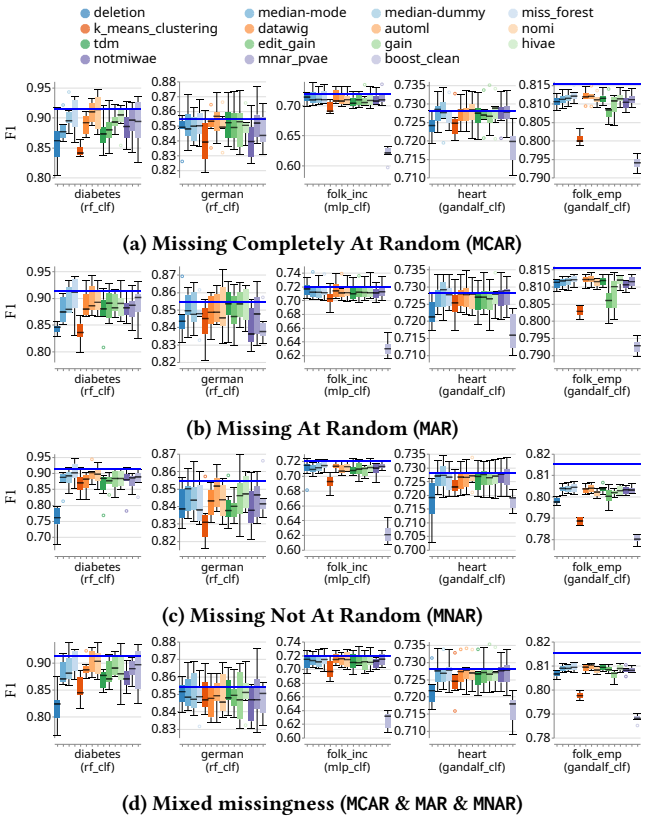
In line with conventional wisdom [44, 58, 59, 63], we find that deletion worsens predictive performance. This effect is strongest for small datasets like `diabetes`, with F1 decreasing as much as 0.1 under `MNAR`, compared to the model trained on clean data. This is due to `deletion` discarding useful information, whereas retaining rows with nulls can still provide valuable signal for model training.

The F1 score on the `bank` dataset is low (0.32), due to severe class imbalance (base rate 0.117, see full version of the paper [48]). Interestingly, models trained on imputed data can sometimes outperform those trained on clean data, as seen for `german` and `heart` under `MCAR`. We hypothesize this occurs when models trained on cleaned data learns spurious correlations (e.g., from noisy or erroneous values), while MVI methods may impute more accurate values, mitigating such artifacts and improving performance.

4.2 Fairness of the Predictive Model

Figure 4 shows the effect of `MVI` on fairness, according to TPRD. Wang and Singh [94] posit that models will exhibit more unfairness under `MAR` and `MNAR` compared to `MCAR`, but we only find weak empirical evidence towards this, even for `deletion`. A nuance here is that we designed missingness scenarios, described in Section 3.1, to be realistic — including `MAR` scenarios where people from disadvantaged groups withhold information that could hurt their chances of getting the desired outcome. Hence, dropping these rows can in fact improve fairness under `MAR` and `MNAR`, as observed on `bank`.

In contrast to Wang and Singh [94], we find that the effect of `MVI` on fairness is strongly correlated with fairness of the model trained on clean data, corroborating the findings of Guha et al. [34]. Fairness depends on two things: dataset characteristics and model



(a) Missing Completely At Random (MCAR)

(b) Missing At Random (MAR)

(c) Missing Not At Random (MNAR)

(d) Mixed missingness (MCAR & MAR & MNAR)

Figure 3: F1 of best performing models (shown in figure) for different imputation strategies (colors in the legend), datasets (x-axis), and missingness mechanisms (subplots). Datasets are in increasing order by size. Blue line shows median performance of the model trained on clean data.

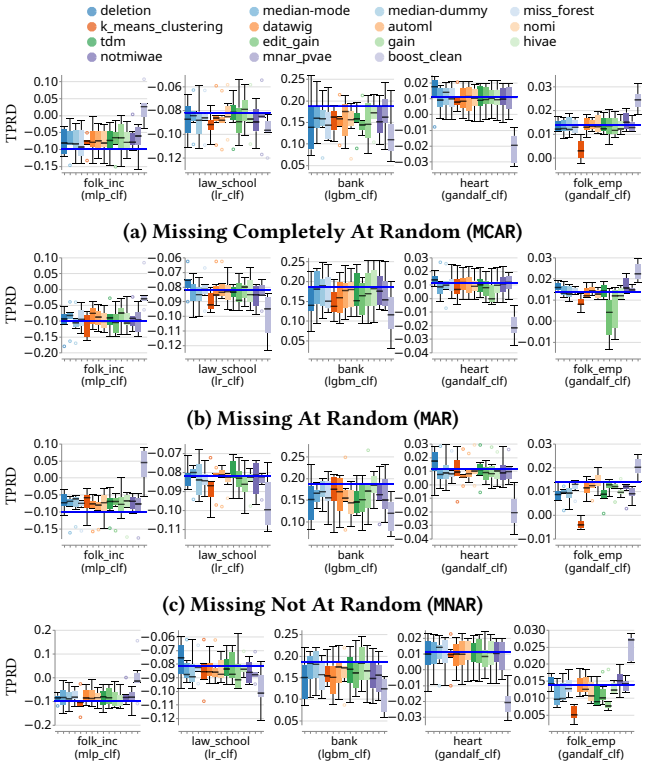
type. All MVI techniques except for boostclean have the same model type as the model trained on clean data (because they are impute-then-classify approaches) and generally preserve fairness of that model, under all missingness mechanisms. Notably, this is agnostic to whether the TPRD of the model trained on clean data is low (close to 0.01 on heart and folk-employment, and 0 on german) or high (close to -0.1 on folk-income and 0.2 on bank).

On the other hand, boostclean, is a joint data cleaning and model training approach and thereby constitutes it own model type, and shows fairness trends that deviate from the model trained on clean data. boostclean significantly improves fairness on folk-income (TPRD close to 0, compared to -0.1 for the clean model) and bank (TPRD close to 0.1, compared to 0.2 for the clean model), but marginally worsens fairness on law-school (TPRD -0.1 compared to -0.08 for the clean model) and heart (TPRD -0.02 compared to 0.01 for the clean model), for all missingness types.

4.3 Stability of the Predictive Model

Figure 5 shows label stability of models trained with different MVI techniques. In line with conventional wisdom [17], stability depends primarily on dataset characteristics (especially size) and model type. Impute-then-classify methods like miss-forest, auto-ml, and datawig, which perform best on F1, also match the stability of models trained on clean data across missingness types and dataset sizes. nomi, which performs best on accuracy and training time, shows comparable stability to these top methods across datasets. clustering, which performed poorly on F1, is likewise unstable on small datasets (diabetes and german, with 905 and 1k samples, respectively). A notable exception is german under mixed missingness, where clustering is competitive despite underperforming on MCAR, MAR, or MNAR individually. This may be because imputation accuracy affects data uncertainty, which in turn drives model uncertainty [29].

Deletion worsens stability compared to the model trained on clean data for all missingness types on diabetes, but, notably, only under MNAR on german. boostclean, which constitutes its own model type, shows a deviation from the stability of the model trained on clean data on all datasets except bank: worsening stability compared to the clean model on folk-income, law-school and folk-employment (except under MCAR), but, surprisingly, improving it on heart, even under mixed missingness.



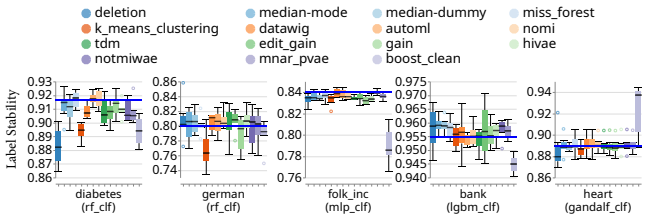
(a) Missing Completely At Random (MCAR)

(b) Missing At Random (MAR)

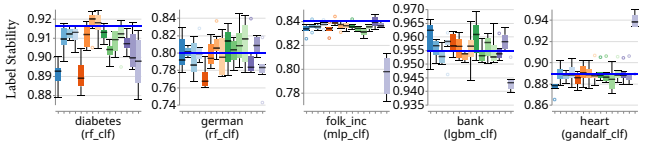
(c) Missing Not At Random (MNAR)

(d) Mixed missingness (MCAR & MAR & MNAR)

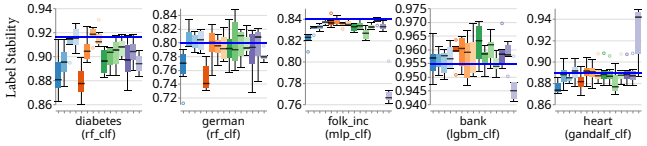
Figure 4: True Positive Rate Difference (unfairness) of best performing models (shown in figure) for different imputation strategies (colors in the legend), datasets (x-axis), and missingness mechanisms (subplots). Values close to 0 are desirable. Datasets are in increasing order by size. Blue line shows median TPRD of the model trained on clean data.



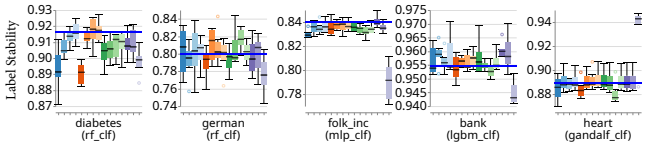
(a) Missing Completely At Random (MCAR)



(b) Missing At Random (MAR)



(c) Missing Not At Random (MNAR)



(d) Mixed missingness (MCAR & MAR & MNAR)

Figure 5: Label Stability of best performing models for different imputation strategies (colors in the legend), datasets (x-axis), and missingness mechanisms (subplots). Values close to 1 are desirable. Blue line shows median performance of the model trained on clean data.

4.4 Imputation Quality and Fairness

In Figure 6, we report the imputation quality of 10 most accurate MVI techniques per category according to F1 score (for categorical columns), RMSE (for numerical columns), and KL divergence (for both numerical and categorical columns, computed over the columns with nulls only) and compare it with the F1 of the downstream model. See full version [48] for an extended comparison of training times and accuracy across all MVI techniques.

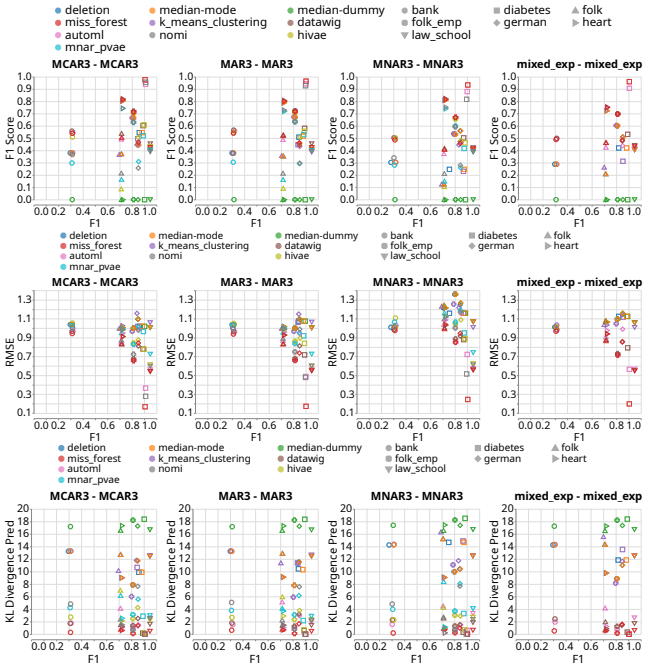
Imputation Quality. MVI techniques with widely varying imputation quality can yield models with similar F1, suggesting that imputation correctness is not a reliable predictor of downstream performance [97]. For instance, in Figure 6a on diabetes, median-dummy has imputation F1 near 0; auto-ml, miss-forest, and nomi are near 1; others fall between 0.5–0.6, yet all produce models with F1 close to 1. This pattern holds across datasets, missingness types, and for numerical columns (Figure 6b). Similar trends are observed for KL divergence (Figure 6c), reinforcing that neither discrepancy-based nor distributional metrics reliably predict downstream model performance, contradicting Shadbahr et al. [84]'s claim.

Imputation Fairness. In Figure 7, we report the imputation fairness of different MVI techniques, according to F1 score difference (for categorical columns), RMSE difference (for numerical columns), and KL divergence difference (for both numerical and categorical columns, computed over the columns with nulls only) and compare it with the fairness of the downstream model, according to TPRD. We find that, while model fairness is generally agnostic to missingness type (as discussed in Section 4.2), **imputation fairness is highly sensitive to missingness mechanism**. For example, in Figure 7c, miss-forest has good imputation fairness on german under MAR (KL difference of -0.4) but significant imputation unfairness under MCAR (KL difference of 2.25), MNAR (KL difference of 1.1) and mixed missingness (KL difference of 1.4).

Further, **imputation fairness is insufficiently predictive of model fairness**. For example, on german under mixed missingness, median-dummy has KL difference close to 0, datawig and miss-forest have KL difference between 1 and 1.5, clustering and median-mode have KL difference between -1 and -1.5, but the models trained using all five of these techniques have a TPRD close to -0.02. Conversely, on diabetes under MAR, auto-ml, clustering, miss-forest, and median-dummy all have near perfect imputation fairness (KL difference close to 0), but different model fairness (TPRD between 0.04 and 0.12). We see similar trends for other imputation fairness metrics such as F1 score difference (Figure 7a) and RMSE difference (Figure 7b).

5 MISSINGNESS SHIFT

Next, we evaluate the correctness, fairness, and stability of 10 most effective MVI techniques from various categories under missingness



(c) KL divergence (imputation) vs F1 (model)

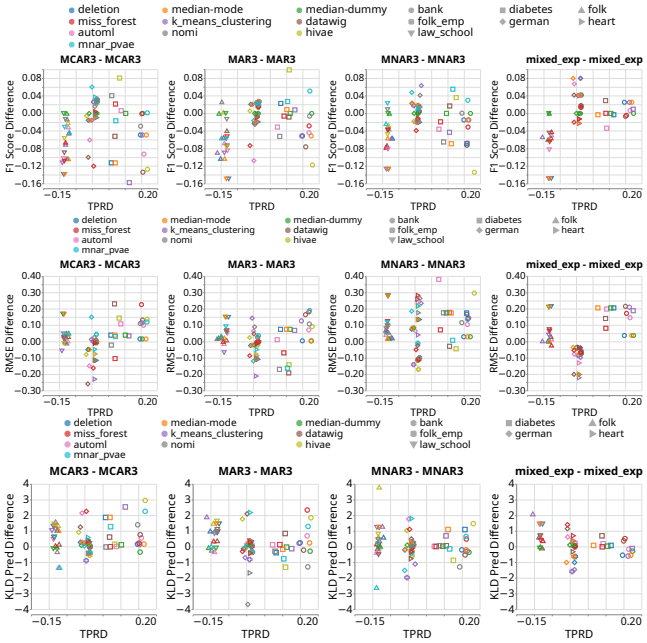
Figure 6: Imputation quality vs. model performance: imputation correctness (F1, RMSE and KL divergence) may not be indicative of model correctness (F1).

shift. We simulate missingness shift in two ways: (i) by varying the missingness mechanism between training and test (S4-9 in Table 1); and (ii) by varying the missingness rates between training and test. First, we hold the fraction of nulls in the test set constant (at 30%) and vary the fraction of nulls in the training set (10%, 30% and 50%). Then, we hold the fraction of nulls in the training set constant (at 30%) and vary the fraction of nulls in the test set (10%, 20%, 30%, 40% and 50%). Note that we have fewer settings for training missingness rates because varying the test set is less computationally demanding (as discussed in Section 3.4). We discuss results on `diabetes`, and defer results on other datasets, with fixed and variable training and test missingness rates, to the full version of the paper [48].

5.1 Correctness of the Predictive Model

Training set missingness. Figure 8 shows the F1 of the Random Forest model on `diabetes` as a function of training missingness rate. Of all `MVI` techniques, `deletion` is most strongly affected by missingness: F1 degrades with increasing missingness rate, and this effect is strongest under `MNAR`. This includes when `MNAR` is encountered both during training (the bottom row in Figure 8 shows the steepest decline in F1 compared to other rows—training missingness) *and* during testing (the right-most column in Figure 8 has the lowest F1 compared to other columns —test missingness).

All other `MVI` techniques, including `boostclean`, are generally robust to higher training missingness rates, and only show a 5% decrease in F1 (compared to 10% decrease with `deletion`), even at rates as high as 50%. This is because even imperfect imputation



(c) KL divergence difference (imputation) vs TPRD (model)

Figure 7: Imputation fairness vs. model fairness: imputation fairness (F1 difference, RMSE difference and KL divergence difference) may not be indicative of model fairness (TPRD).

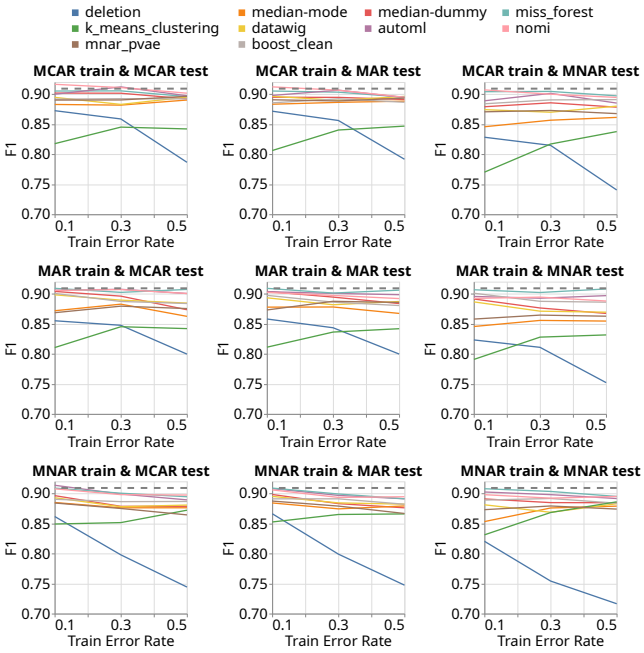


Figure 8: F1 of the Random Forest model on `diabetes`, as a function of training set missingness rate. Dashed line shows performance of the model trained on clean data.

provides valuable insights for the model, making deletion a less favorable choice. A notable exception is `clustering`, which, surprisingly and somewhat counter-intuitively, has higher F1 at higher training missingness rates, and is actually better under `MNAR` than under `MAR` and `MCAR`, for all missingness rates and scenarios.

Test set missingness. We find that F1 generally decreases with an increase in test missingness, corroborating the findings of Shadbahr et al. [84] and Miao et al. [67]. A notable exception is `miss-forest`, which is robust to both forms of missingness shift such as changing missingness rates (as observed in [67]) and missingness mechanisms. As for training set missingness, F1 decreases with an increase in test missingness most steeply under `MNAR` (both during training and test), further supporting the findings of Miao et al. [67]. And, once again, `clustering` is an exception to this trend, instead showing invariance to test missingness rates under `MNAR` train (irrespective of test missingness) but not under `MCAR` and `MAR` train. See full version of the paper [48] for complete results.

5.2 Fairness of the Predictive Model

Training set missingness. Figure 9 shows TPRD of Random Forest on `diabetes` as a function of training set missingness rate. While we previously found that fairness is generally agnostic to missingness type when it is the same between training and test sets (see Section 4.2), we find that **model fairness is highly sensitive to missingness shift** — in terms of both different missingness rates and different missingness mechanisms between training and test.

Worryingly, there is no consistent trend across `MVI` technique, missingness type and training test missingness rate. For example, consider `miss-forest`, which was the most robust to missingness

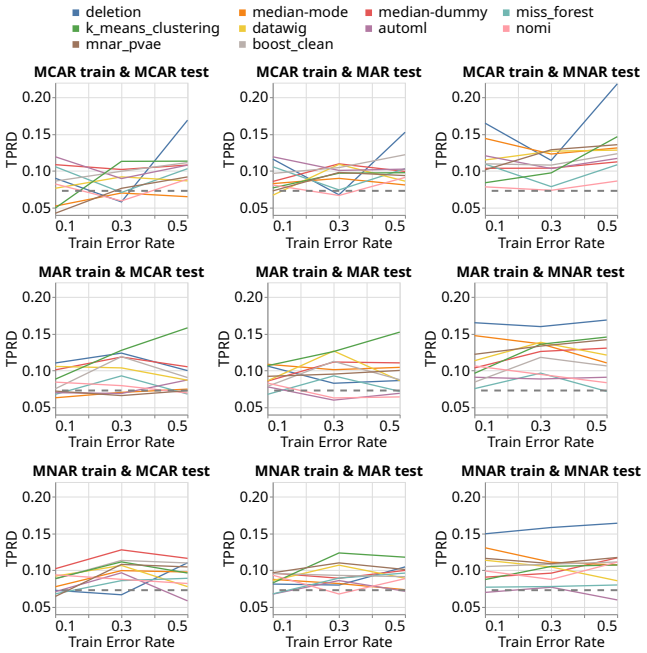


Figure 9: True Positive Rate Difference of Random Forest on `diabetes`, as a function of training missingness rate. Dashed line shows performance of the model trained on clean data.

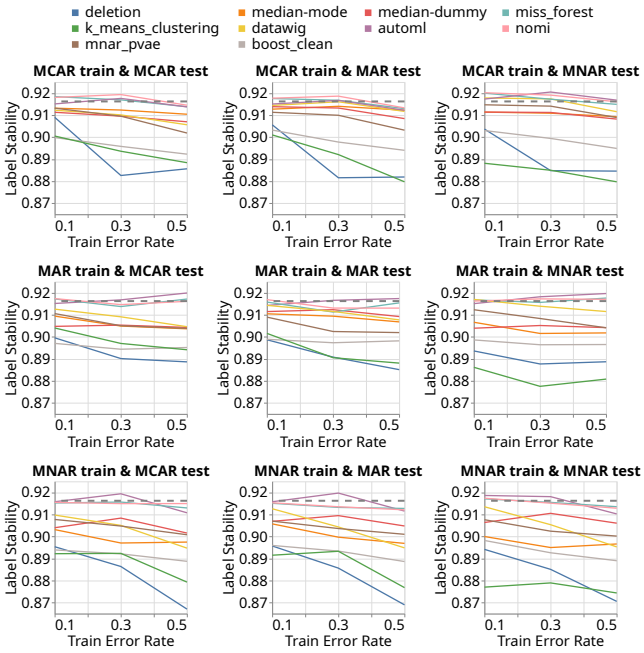


Figure 10: Label Stability of Random Forest on `diabetes`, as a function of training set missingness rate. Dashed line shows performance of the model trained on clean data.

shift according to F1. Under MCAR training, `miss-forest` preserves fairness of the model trained on clean data (shown with a dashed grey line) when training and test missingness rates match (at 30% training error rate, fixed at 30% for this experiment), but worsens fairness (higher TPRD) when they are different (at 10% and 50% training error rates). Under MAR training, however, we see the opposite behavior, with `miss-forest` preserving clean model fairness at 10% and 50% train missingness rates, but worsening fairness when training and test missingness rates are equal (at 30%). Under MNAR training, TPRD increases with an increase in training missingness rate under MCAR and MAR test, and remains constant when there is no shift in missingness mechanism (under MNAR test).

Test set missingness. We measured the impact of test missingness rate on fairness and found that fairness is highly sensitive to such shifts. For `boostclean`, `datawig`, and `mnar-pvae`, TPRD generally increases (fairness worsens) as test error rate rises, though not always monotonically. `miss-forest`, `auto-ml`, and `nomi` are robust to increases in test missingness rate under all scenarios. Simpler methods such as `deletion`, `clustering`, `median-mode` and `median-dummy` show no consistent trend, even when missingness types remain the same and only missingness rates change between train and test. For example, with `deletion` and `clustering`, TPRD increases with test missingness in scenarios S1 (MCAR train, MCAR test) and S3 (MNAR train, MNAR test), but decreases in S2 (MAR train, MAR test). See full version for details [48].

In summary, our results corroborate the findings of Guha et al. [34], and are a cause for concern as they indicate that the MVI techniques that perform best during development may not preserve fairness post-deployment, where shifts are likely to occur.

5.3 Stability of the Predictive Model

Training set missingness. Figure 10 shows Label Stability of the Random Forest model on `diabetes` as a function of training set missingness rate. Under MCAR and MAR missingness, most MVI techniques with the exception of `boostclean`, `deletion`, and `clustering` show good stability (comparable to the model trained on clean data), and are generally insensitive to missingness rates. `deletion` and `clustering` are the least stable methods and show a monotonic decrease in stability with increase in missingness rate, with strongest effect under MNAR. `miss-forest`, `auto-ml`, and `nomi` are the most robust MVI techniques and preserve stability of the clean model under all missingness settings and error rates. This further highlights how imputation quality directly influences data uncertainty, ultimately impacting the overall uncertainty of the final model, as discussed in Section 4.3.

Test set missingness. We find that test missingness rate has little effect on model stability under all settings except `clustering`, which shows lower label stability at higher missingness rates, most pronounced in scenario S7 (MAR train, MNAR test). See full version of the paper [48] for complete results.

6 RUNNING TIME

Table 4 presents the training time of MVI techniques for each dataset, averaged across all unique training scenarios (single- and multi-mechanism S1-S3, S10). Our time efficiency analysis approach aligns with prior work [66, 93], who also focused on training time as inference times are comparably fast across all techniques.

Table 4: Training time (in seconds) of MVI techniques averaged across single- and multi-mechanism scenarios (S1-3, S10). Imputers are sorted by running time on the `folk_emp` dataset, and datasets are ordered by the number of rows. Dataset shapes reflect training sets with 30% rows with nulls. Values represent mean running times across seeds, with standard deviations.

Imputer	diabetes (633, 17)	german (700, 21)	folk_inc (12000, 10)	law_school (16638, 11)	bank (32003, 13)	heart (56000, 11)	folk_emp (242112, 16)
median-dummy	0.013 ± 0.000	0.014 ± 0.001	0.021 ± 0.001	0.024 ± 0.001	0.027 ± 0.000	0.053 ± 0.001	0.773 ± 0.024
median-mode	0.012 ± 0.000	0.014 ± 0.001	0.023 ± 0.001	0.025 ± 0.001	0.031 ± 0.001	0.067 ± 0.003	0.933 ± 0.019
deletion	0.013 ± 0.000	0.013 ± 0.001	0.024 ± 0.001	0.025 ± 0.000	0.046 ± 0.001	0.087 ± 0.004	1 ± 0.049
mnar_pvae	8 ± 0.705	14 ± 11	14 ± 1	22 ± 11	55 ± 30	42 ± 6	206 ± 7
edit_gain	2 ± 0.119	2 ± 0.141	13 ± 0.221	21 ± 2	30 ± 1	62 ± 7	215 ± 4
nomi	11 ± 3	14 ± 7	22 ± 2	22 ± 2	29 ± 1	38 ± 2	356 ± 20
tdm	932 ± 74	1023 ± 11	1297 ± 22	1172 ± 92	1412 ± 34	1310 ± 42	1449 ± 31
notmiwae	161 ± 101	217 ± 82	555 ± 2	804 ± 8	665 ± 284	1393 ± 535	2944 ± 725
gain	261 ± 12	298 ± 5	1115 ± 38	1484 ± 30	1964 ± 36	2892 ± 48	6148 ± 178
hivae	68 ± 0.784	95 ± 1	745 ± 15	1073 ± 11	2450 ± 130	3794 ± 129	7163 ± 317
k_means_clustering	10 ± 0.402	13 ± 0.442	27 ± 0.640	323 ± 7	998 ± 49	1037 ± 64	7427 ± 778
miss_forest	111 ± 17	244 ± 86	1758 ± 526	2530 ± 851	4307 ± 1310	6337 ± 1601	20358 ± 4934
datawig	596 ± 185	277 ± 59	604 ± 46	2361 ± 492	5089 ± 651	7592 ± 854	31060 ± 3743
automl	1953 ± 195	1805 ± 212	5559 ± 581	6803 ± 565	13893 ± 1687	19055 ± 2710	104476 ± 14743

Our results reveal that statistical imputers `deletion`, `median-mode`, and `median-dummy` are the fastest, while still delivering competitive accuracy for larger datasets like `heart` and `folk_emp`. In contrast, `miss-forest`, `datawig`, and `auto-ml` exhibit the longest training times, with at least one of these methods achieving the highest imputation accuracy in most cases. Interestingly, `auto-ml` requires three times more training time than `datawig`, the second most computationally intensive technique. This difference is due to the auto-ML nature of `auto-ml`, which involves extensive hyperparameter and network architecture tuning.

Among non-statistical techniques, `mnar-pvae`, `edit-gain`, and `nomi` are the most efficient. Notably, `nomi` delivers accuracy on par with `miss-forest`, `datawig`, and `auto-ml`, successfully balancing imputation accuracy and training time. A key comparison is between `gain` and `edit-gain`. As explained in the full version of the paper [48], `edit-gain` achieves a 28x speedup on `folk_emp`, with even greater improvements for smaller datasets, as shown in Table 4, while maintaining comparable accuracy to `gain`.

7 SUMMARY OF EXPERIMENTAL FINDINGS

Do not drop your nulls! Building on prior evidence [44, 58, 59, 63], we confirm that deletion is the least effective strategy for model accuracy, fairness, and stability—especially when each row holds valuable information. While deletion leads to data loss by design, its suitability depends more on data quality than quantity. If rows are duplicates or contain errors, deletion may be warranted.

Multiple imputation shows mixed results. There is conflicting evidence on the performance of multiple imputation (MI) [45], and our empirical findings are similarly mixed and somewhat unexpected. Feng [24] and Le Morvan et al. [55] argue that MI outperforms impute-then-classify approaches in predictive performance, while Graham [32] and McNeish [64] find MI effective even with limited data and small error rates. In contrast, we find that MI (specifically, `boostclean`) is *only* competitive on small datasets and is less stable than simpler `MVI` techniques. This likely stems from the complexity-stability trade-off [9, 17]: MI employs a more complex model class that minimizes empirical loss but exhibits greater prediction variance under small training set perturbations.

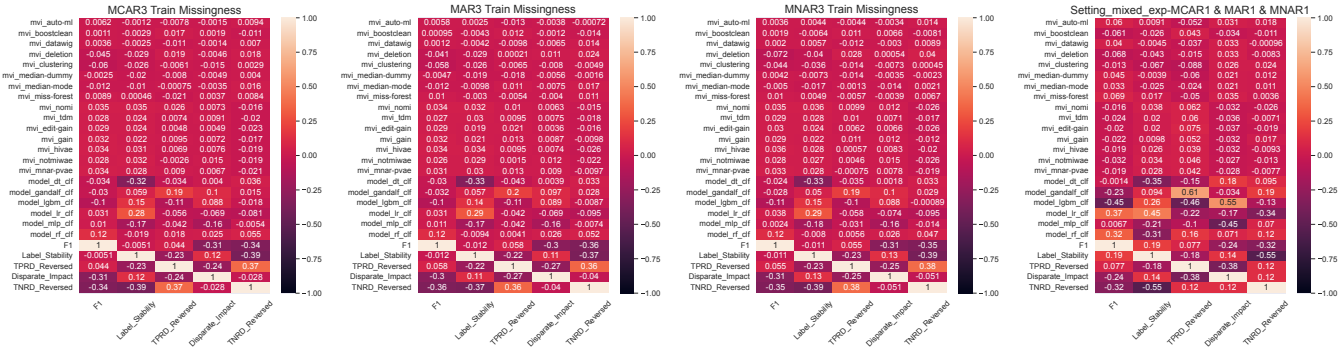
Fairness is highly missingness-specific. We find that no `MVI` technique is consistently fairness-preserving, corroborating the findings of Zhang and Long [98] and Guha et al. [34]. Further, we find that fairness is highly sensitive to changes in missingness rates and missingness mechanisms between training and test sets, which are likely to occur in practice, and therefore a cause for ethical concern.

Imputation quality and fairness metrics often fail to predict the correctness or fairness of downstream models. Shadbahr et al. [84] argue that distributional imputation quality metrics better predict model performance than discrepancy metrics. However, we find that neither reliably predicts downstream performance, as strong learners can compensate for poor imputations. Moreover, imputation fairness does not predict model fairness: fair imputers can still yield unfair models, while models trained with fairness-poor `MVI` techniques can achieve good downstream fairness.

Model stability depends more on the dataset size and `MVI` *technique than on the missingness scenario.* We find that for large datasets even simple statistical imputers can preserve stability. In contrast, for small datasets, only a few `MVI` techniques do so, while deletion, statistical imputation and complex ML-based `MVI` all worsen stability.

Sensitivity to train and test missingness rates. Model performance (F1) is more affected by test missingness than training missingness. Fairness, however, is highly sensitive to both. Model stability is largely unaffected by test missingness, but most `MVI` techniques become more unstable with higher training missingness.

Existing MNAR-specific methods are insufficient. `MNAR` is theoretically the hardest setting to model. `MVI` techniques, including `MNAR`-specific `not-miwae` and `mnar-pvae`, perform poorly under `MNAR`. For example, F1 and stability are more sensitive to missingness rates under `MNAR` than under `MCAR` or `MAR`. This performance gap stems from unrealistic assumptions and limited evaluations,



(a) MCAR missingness **(b) MAR missingness** **(c) MNAR missingness** **(d) Mixed missingness**

Figure 11: Spearman correlation (ρ) between MVI technique, model type, and performance metrics (F1, fairness and stability) for different train missingness mechanisms (subplots). TPRD and TNRD values close to 0 are ideal (fair), so we compute correlations using $TPRD\_Reversed = 1 - |TPRD|$ and $TNRD\_Reversed = 1 - |TNRD|$. Supplementary plots are in the full version [48].

typically on large datasets and a single scenario, lacking the diversity of our benchmark. Advancing MNAR-specific MVI methods and building comprehensive benchmarks are key future directions.

Deep learning outperforms tree-based models on large datasets. Interest in deep learning for tabular data has surged [8, 22, 31, 38, 95]. Recall that Figures 3, 4, and 5 show only the best-performing models (by F1) for clarity. Tree-based methods like random forests and gradient-boosted trees outperform deep table-learning models like GANDALF [46] on small datasets (`diabetes`, `german`), aligning with prior findings [33, 86]. However, GANDALF excels on larger datasets (`heart`, `folk-employment`), highlighting the increasing effectiveness of deep learning across diverse data modalities.

Best-performing approaches under single- and multi-mechanism missingness can differ. A key contribution of our study is the evaluation under mixed (multi-mechanism) missingness — a complex pattern likely to arise in practice. We find that performance trends differ between single- and multi-mechanism settings. Figure 11 shows that model types like `gandalf_clf` and `lgbm_clf` are uncorrelated with performance metrics under single-mechanism missingness, but become correlated under mixed missingness: F1 correlations are $\rho = -0.24$ for `gandalf_clf` and $\rho = -0.37$ for `lgbm_clf`, while fairness (TPRD) correlations are $\rho = 0.55$ and $\rho = -0.4$, respectively.

It's complicated! Prior studies [30, 35, 44, 57, 84] have found no universally best missing value imputation technique for predictive performance. Our findings reinforce this, revealing trade-offs between F1, fairness, and stability that depend on the predictive model's architecture. For instance, under multi-mechanism missingness (Figure 11), logistic regression (`lr_clf`) correlates positively with F1 (ρ=0.23) and stability (ρ=0.38) but negatively with fairness (TPRD, ρ=-0.23). Random forest (`rf_clf`) shows weak positive correlations with fairness (TPRD, ρ=0.16) and F1 (ρ=0.25) but a negative correlation with stability (ρ=-0.24). Deep table-learning (`gandalf_clf`) is strongly correlated with fairness (TPRD, ρ=0.55), weakly with stability (ρ=0.12), and negatively with F1 (ρ=-0.24). This is not bad news, but a recognition of the complexity of learning from incomplete data, and of the need for rigorous, holistic evaluation protocols, like those used here, to identify the best imputation method and model architecture for a given task.

8 CONCLUSIONS, LIMITATIONS AND FUTURE WORK

Conclusions. We introduced `Shades-of-Null`, an evaluation suite for responsible missing value imputation. A key contribution is evaluating fairness and stability alongside predictive performance. Additionally, we model realistic missingness scenarios beyond Rubin's MCAR, MAR, and MNAR, incorporating multi-mechanism missingness and missingness shift. Through 29,736 experimental pipelines, we assessed various MVI methods under realistic missingness conditions, revealing key trends and trade-offs across evaluation metrics.

Limitations. We began with clean datasets, designed meaningful missingness scenarios, and simulated them using error injectors. A key limitation is that we evaluate performance on synthetically generated—rather than naturally occurring—missingness. This is common in the field: clean ground truth is rarely available, so we assume that injecting synthetic errors only worsens the effects of existing (unknown) ones. Under this assumption, any observed performance degradation likely underestimates the true degradation relative to an ideal ground truth.

Future work. Creating research datasets with naturally occurring missingness shifts is a promising direction. While predictive performance, fairness, and stability are often treated as orthogonal, we uncover trade-offs driven by missingness, imputation choice, and, in some cases, model architecture. The absence of a universally best imputation method highlights the need for holistic metrics, evaluation procedures, and techniques grounded in the data lifecycle. Missing value imputation remains a critical aspect of responsible data engineering, to which our work makes a contribution.

9 ACKNOWLEDGMENTS

This research was conducted as part of the RAI for Ukraine program, led by the Center for Responsible AI at New York University in collaboration with Ukrainian Catholic University in Lviv. It was supported in part by a grant from the Simons Foundation (SFARI award #1280457, JS) and by NSF Awards No. 2326193 and 2312930.

REFERENCES

[1] 1983. Motor Vehicle Manufacturers Ass'n v. State Farm Mutual Auto. Ins. Co. https://supreme.justia.com/cases/federal/us/463/29/

[2] Mohamed Abdelaal, Christian Hammacher, and Harald Schoening. 2023. Rein: A comprehensive benchmark framework for data cleaning methods in ml pipelines. *arXiv preprint arXiv:2302.04702* (2023).

[3] Nil-Jana Akpinar, Zachary Lipton, and Alexandra Chouldechova. 2024. The Impact of Differential Feature Under-reporting on Algorithmic Fairness. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1355–1382.

[4] Mustafa Alabadla, Fatimah Sidi, Iskandar Ishak, Hamidah Ibrahim, Lilly Affendey, Zafienas Che Ani, Marzanah Jabar, Umar Bukar, Navin Kumar Devaraj, Ahmad Muda, Anas Tharek, Noritah Omar, and Izham Jaya. 2022. Systematic Review of Using Machine Learning in Imputing Missing Values. *IEEE Access* 10 (01 2022), 1–1. https://doi.org/10.1109/ACCESS.2022.3160841

[5] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.* 104 (2016), 671.

[6] Gustavo EAPA Batista and Maria Carolina Monard. 2003. An analysis of four missing data treatment methods for supervised learning. *Applied artificial intelligence* 17, 5-6 (2003), 519–533.

[7] Felix Biessmann, Tammo Rukat, Phillipp Schmidt, Prathik Naidu, Sebastian Schelter, Andrey Taptunov, Dustin Lange, and David Salinas. 2019. DataWig: Missing value imputation for tables. *Journal of Machine Learning Research* 20, 175 (2019), 1–6.

[8] Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. 2022. Deep neural networks and tabular data: A survey. *IEEE transactions on neural networks and learning systems* (2022).

[9] Leo Breiman. 2001. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statist. Sci.* 16, 3 (2001), 199 – 231. https://doi.org/10.1214/ss/1009213726

[10] Simon Caton, Saiteja Malisetty, and Christian Haas. 2022. Impact of Imputation Strategies on Fairness in Machine Learning. *J. Artif. Int. Res.* 74 (sep 2022), 25. https://doi.org/10.1613/jair.1.13197

[11] A. Feder Cooper, Katherine Lee, Madiha Choksi, Solon Barocas, Christopher De Sa, James Grimmelmann, Jon Kleinberg, Siddhartha Sen, and Baobao Zhang. 2023. Arbitrariness and Prediction: The Confounding Role of Variance in Fair Classification. arXiv:2301.11562 [cs.LG]

[12] Amanda Coston, Anna Kawakami, Haiyi Zhu, Ken Holstein, and Hoda Heidari. 2023. A validity perspective on evaluating the justified use of data-driven decision-making algorithms. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. IEEE, 690–704.

[13] Kathleen Creel and Deborah Hellman. 2022. The algorithmic leviathan: Arbitrariness, fairness, and opportunity in algorithmic decision-making systems. *Canadian Journal of Philosophy* 52, 1 (2022), 26–43.

[14] Michael Christopher Darling and David John Stracuzzi. 2018. *Toward uncertainty quantification for supervised classification.* Technical Report. Sandia National Lab.(SNL-NM), Albuquerque, NM (United States).

[15] C. J. Date. 1984. A Critique of the SQL Database Language. *SIGMOD Rec.* 14, 3 (1984), 8–54. https://doi.org/10.1145/984549.984551

[16] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. Retiring adult: New datasets for fair machine learning. 34 (2021), 6478–6490.

[17] Pedro Domingos. 2000. A Unified Bias-Variance Decomposition and its Applications. 231–238.

[18] Dheeru Dua and Casey Graff. 2019. UCI Machine Learning Repository - Adult Dataset. https://archive.ics.uci.edu/dataset/2/adult. https://archive.ics.uci.edu/dataset/2/adult

[19] Bradley Efron and Robert J Tibshirani. 1994. *An introduction to the bootstrap*. CRC press.

[20] Tlamelo Emmanuel, Thabiso Maupong, Dimane Mpoeleng, Thabo Semong, Banyatsang Mphago, and Oteng Tabona. 2021. A survey on missing data in machine learning. *Journal of Big data* 8 (2021), 1–37.

[21] Alireza Farhangfar, Lukasz Kurgan, and Jennifer Dy. 2008. Impact of imputation of missing values on classification error for discrete data. *Pattern Recognition* 41, 12 (2008), 3692–3705.

[22] Sheikh Amir Fayaz, Majid Zaman, Sameer Kaul, and Muheet Ahmed Butt. 2022. Is deep learning on tabular data enough? An assessment. *International Journal of Advanced Computer Science and Applications* 13, 4 (2022).

[23] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*, Longbing Cao, Chengqi Zhang, Thorsten Joachims, Geoffrey I. Webb, Dragos D. Margineantu, and Graham Williams (Eds.). ACM, 259–268. https://doi.org/10.1145/2783258.2783311

[24] Raymond Feng. 2023. *Adapting Fairness-Intervention Algorithms to Missing Data*. Ph.D. Dissertation. Harvard College.

[25] Raymond Feng, Flavio Calmon, and Hao Wang. 2024. Adapting Fairness Interventions to Missing Values. *Advances in Neural Information Processing Systems* 36 (2024).

[26] Batya Friedman and Helen Nissenbaum. 1996. Bias in Computer Systems. *ACM Trans. Inf. Syst.* 14, 3 (1996), 330–347. https://doi.org/10.1145/230538.230561

[27] Andreas Fuster, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. 2022. Predictably Unequal? The Effects of Machine Learning on Credit Markets. *The Journal of Finance* 77, 1 (2022), 5–47. https://doi.org/10.1111/jofi.13092

[28] Satish Gajawada and Durga Toshniwal. 2012. Missing value imputation method based on clustering and nearest neighbours. *International Journal of Future Computer and Communication* 1, 2 (2012), 206–208.

[29] Yarin Gal et al. 2016. Uncertainty in deep learning. (2016).

[30] Pedro García Laencina, José Luis Sancho-Gómez, and Aníbal Figueiras-Vidal. 2010. Pattern classification with missing data: A review. *Neural Computing and Applications* 19 (03 2010), 263–282. https://doi.org/10.1007/s00521-009-0295-6

[31] 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.

[32] John W Graham. 2009. Missing data analysis: Making it work in the real world. *Annual review of psychology* 60 (2009), 549–576.

[33] 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.

[34] 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* (2024).

[35] Md. Kamrul Hasan, Md. Ashraful Alam, Shidhartho Roy, Aishwariya Dutta, Md. Tasnim Jawad, and Sunanda Das. 2021. Missing value imputation affects the performance of machine learning: A review and analysis of the literature (2010–2021). *Informatics in Medicine Unlocked* 27 (2021), 100799. https://doi.org/10.1016/j.imu.2021.100799

[36] Denys Herasymuk, Falaah Arif Khan, and Julia Stoyanovich. 2024. Responsible Model Selection with Virny and VirnyView. In *Companion of the 2024 International Conference on Management of Data, SIGMOD/PODS 2024, Santiago AA, Chile, June 9-15, 2024*. ACM, 488–491. https://doi.org/10.1145/3626246.3654738

[37] David C. Howell. 2007. The Treatment of Missing Data. https://api.semanticscholar.org/CorpusID:63503512

[38] Yejin Hwang and Jongwoo Song. 2023. Recent deep learning methods for tabular data. *Communications for Statistical Applications and Methods* 30, 2 (2023), 215–226.

[39] Ihab F. Ilyas and Xu Chu. 2019. *Data Cleaning*. ACM Books, Vol. 28. ACM. https://doi.org/10.1145/3310205

[40] Niels Bruun Ipsen, Pierre-Alexandre Mattei, and Jes Frellsen. 2020. not-MIWAE: Deep generative modelling with missing not at random data. *arXiv preprint arXiv:2006.12871* (2020).

[41] Sebastian Jäger, Arndt Allhorn, and Felix Bießmann. 2021. A benchmark for data imputation methods. *Frontiers in big Data* 4 (2021), 693674.

[42] Haewon Jeong, Hao Wang, and Flavio Calmon. 2022. Fairness without Imputation: A Decision Tree Approach for Fair Prediction with Missing Values. *Proceedings of the AAAI Conference on Artificial Intelligence* 36 (06 2022), 9558–9566. https://doi.org/10.1609/aaai.v36i9.21189

[43] Haewon Jeong, Hao Wang, and Flavio P Calmon. 2022. Fairness without imputation: A decision tree approach for fair prediction with missing values. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 9558–9566.

[44] Luke Joel and Wesley Doorsamy. 2022. A Review of Missing Data Handling Techniques for Machine Learning. (09 2022). https://doi.org/10.15157/IJITIS.2022.5.3.971-1005

[45] Ralph C. A. Rippe Joost R. van Ginkel, Marielle Linting and Anja van der Voort. 2020. Rebutting Existing Misconceptions About Multiple Imputation as a Method for Handling Missing Data. *Journal of Personality Assessment* 102, 3 (2020), 297–308. https://doi.org/10.1080/00223891.2018.1530680 arXiv:https://doi.org/10.1080/00223891.2018.1530680 PMID: 30657714.

[46] Manu Joseph and Harsh Raj. 2022. GANDALF: gated adaptive network for deep automated learning of features. *arXiv preprint arXiv:2207.08548* (2022).

[47] Bojan Karlaš, Peng Li, Renzhi Wu, Nezihe Merve Gürel, Xu Chu, Wentao Wu, and Ce Zhang. 2020. Nearest neighbor classifiers over incomplete information: From certain answers to certain predictions. *arXiv preprint arXiv:2005.05117* (2020).

[48] Falaah Arif Khan, Denys Herasymuk, Nazar Protsiv, and Julia Stoyanovich. 2024. Still More Shades of Null: An Evaluation Suite for Responsible Missing Value Imputation. *CoRR* abs/2409.07510 (2024). https://doi.org/10.48550/ARXIV.2409.07510 arXiv:2409.07510

[49] Falaah Arif Khan, Denys Herasymuk, and Julia Stoyanovich. 2023. On Fairness and Stability: Is Estimator Variance a Friend or a Foe? *arXiv preprint arXiv:2302.04525* (2023).

[50] Falaah Arif Khan, Eleni Manis, and Julia Stoyanovich. 2022. Towards Substantive Conceptions of Algorithmic Fairness: Normative Guidance from Equal Opportunity Doctrines. In *Equity and Access in Algorithms, Mechanisms, and Optimization, EAAMO 2022, Arlington, VA, USA, October 6-9, 2022*. ACM, 18:1–18:10.

https://doi.org/10.1145/3551624.3555303
[51] Sanjay Krishnan, Michael J Franklin, Ken Goldberg, and Eugene Wu. 2017. Boostclean: Automated error detection and repair for machine learning. *arXiv preprint arXiv:1711.01299* (2017).
[52] Sanjay Krishnan, Jiannan Wang, Eugene Wu, Michael J Franklin, and Ken Goldberg. 2016. Activeclean: Interactive data cleaning for statistical modeling. *Proceedings of the VLDB Endowment* 9, 12 (2016), 948–959.
[53] Andreas Köchling and Marlen C. Wehner. 2020. Discriminated by an Algorithm: A Systematic Review of Discrimination and Fairness by Algorithmic Decision-Making in the Context of HR Recruitment and HR Development. *Business Research* 13, 3 (2020), 795–848. https://doi.org/10.1007/s40685-020-00134-w
[54] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. *How We Analyzed the COMPAS Recidivism Algorithm.* Technical Report. ProPublica. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm
[55] Marine Le Morvan, Julie Josse, Erwan Scornet, and Gaël Varoquaux. 2021. What's a good imputation to predict with missing values? *Advances in Neural Information Processing Systems* 34 (2021), 11530–11540.
[56] Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. 2022. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 12, 3 (2022), e1452.
[57] Wei-Chao Lin and Chih-Fong Tsai. 2020. Missing value imputation: a review and analysis of the literature (2006–2017). *Artificial Intelligence Review* 53 (02 2020), 1487–1509. https://doi.org/10.1007/s10462-019-09709-4
[58] R.J.A. Little and D.B. Rubin. 1987. *Statistical Analysis with Missing Data.* John Wiley & Sons, Inc., New York.
[59] Todd D Little, Kai U Schnabel, and et al. 2015. *Modeling Longitudinal and Multilevel Data: Practical Issues, Applied Approaches, and Specific Examples.* Psychology Press.
[60] Mingxuan Liu, Siqi Li, Han Yuan, Marcus Eng Hock Ong, Yilin Ning, Feng Xie, Seyed Ehsan Saffari, Yuqing Shang, Victor Volovici, Bibhas Chakraborty, et al. 2023. Handling missing values in healthcare data: A systematic review of deep learning-based imputation techniques. *Artificial intelligence in medicine* 142 (2023), 102587.
[61] Carol Xuan Long, Hsiang Hsu, Wael Alghamdi, and Flavio P Calmon. 2023. Arbitrariness Lies Beyond the Fairness-Accuracy Frontier. *arXiv preprint arXiv:2306.09425* (2023).
[62] Chao Ma and Cheng Zhang. 2021. Identifiable generative models for missing not at random data imputation. *Advances in Neural Information Processing Systems* 34 (2021), 27645–27658.
[63] Fernando Martínez-Plumed, Cèsar Ferri, David Nieves, and José Hernández-Orallo. 2019. Fairness and missing values. *arXiv preprint arXiv:1905.12728* (2019).
[64] Daniel McNeish. 2017. Missing data methods for arbitrary missingness with small samples. *Journal of Applied Statistics* 44, 1 (2017), 24–39. https://doi.org/10.1080/02664763.2016.1158246 arXiv:https://doi.org/10.1080/02664763.2016.1158246
[65] José Mena, Oriol Pujol, and Jordi Vitrià. 2022. A Survey on Uncertainty Estimation in Deep Learning Classification Systems from a Bayesian Perspective. *ACM Computing Surveys (CSUR)* 54 (2022), 1 – 35.
[66] Xiaoye Miao, Yangyang Wu, Lu Chen, Yunjun Gao, Jun Wang, and Jianwei Yin. 2021. Efficient and effective data imputation with influence functions. *Proceedings of the VLDB Endowment* 15, 3 (2021), 624–632.
[67] Xiaoye Miao, Yangyang Wu, Lu Chen, Yunjun Gao, and Jianwei Yin. 2022. An experimental survey of missing data imputation algorithms. *IEEE Transactions on Knowledge and Data Engineering* 35, 7 (2022), 6630–6650.
[68] Sam Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application* 8, 1 (2021), 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902
[69] Rajarshi Mitra, Stephen F. McGough, Tamoghna Chakraborti, and et al. 2023. Learning from data with structured missingness. *Nat Mach Intell* 5 (2023), 13–23. https://doi.org/10.1038/s42256-022-00596-z
[70] Arvind Narayanan. 2018. 21 Fairness Definitions and Their Politics. Available at https://www.cs.princeton.edu/~arvindn/talks.
[71] Alfredo Nazabal, Pablo M Olmos, Zoubin Ghahramani, and Isabel Valera. 2020. Handling incomplete heterogeneous data using vaes. *Pattern Recognition* 107 (2020), 107501.
[72] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations. *Science* 366, 6464 (2019), 447–453. https://doi.org/10.1126/science.aax2342
[73] Benjamin Peherstorfer, Karen Willcox, and Max Gunzburger. 2018. Survey of multifidelity methods in uncertainty propagation, inference, and optimization. *Siam Review* 60, 3 (2018), 550–591.
[74] Therese Pigott. 2010. A Review of Methods for Missing Data. *Educational Research and Evaluation: An International Journal on Theory and Practice* 7 (08 2010), 353–383. https://doi.org/10.1076/edre.7.4.353.8937
[75] Maximilian Pintz, Joachim Sicking, Maximilian Poretschkin, and Maram Akila. 2022. A Survey on Uncertainty Toolkits for Deep Learning. *arXiv preprint arXiv:2205.01040* (2022).
[76] Abdulhakim Qahtan, Ahmed Elmagarmid, Raul Fernandez, Mourad Ouzzani, and Nan Tang. 2018. FAHES: A Robust Disguised Missing Values Detector. https://doi.org/10.1145/3219819.3220109
[77] Alene K. Rhea, Kelsey Markey, Lauren D'Arinzo, Hilke Schellmann, Mona Sloane, Paul Squires, Falaah Arif Khan, and Julia Stoyanovich. 2022. An external stability audit framework to test the validity of personality prediction in AI hiring. *Data Min. Knowl. Discov.* 36, 6 (2022), 2153–2193. https://doi.org/10.1007/S10618-022-00861-0
[78] Alene K. Rhea, Kelsey Markey, Lauren D'Arinzo, Hilke Schellmann, Mona Sloane, Paul Squires, and Julia Stoyanovich. 2022. Resume Format, LinkedIn URLs and Other Unexpected Influences on AI Personality Prediction in Hiring: Results of an Audit. In *AIES '22: AAAI/ACM Conference on AI, Ethics, and Society, Oxford, United Kingdom, May 19 - 21, 2021*, Vincent Conitzer, John Tasioulas, Matthias Scheutz, Ryan Calo, Martina Mara, and Annette Zimmermann (Eds.). ACM, 572–587. https://doi.org/10.1145/3514094.3534189
[79] D.B. Rubin. 1976. Inference and Missing Data. *Biometrika* 63 (1976), 581–592.
[80] Donald B. Rubin. 1987. *Multiple Imputation for Nonresponse in Surveys.* John Wiley & Sons Inc., New York. https://doi.org/10.1002/9780470316696
[81] Joseph L Schafer and John W Graham. 2002. Missing data: our view of the state of the art. *Psychological methods* 7, 2 (2002), 147.
[82] Sebastian Schelter, Yuxuan He, Jatin Khilnani, and Julia Stoyanovich. 2019. Fairprep: Promoting data to a first-class citizen in studies on fairness-enhancing interventions. *EDBT* (2019).
[83] Sebastian Schelter and Julia Stoyanovich. 2020. Taming Technical Bias in Machine Learning Pipelines. *IEEE Data Eng. Bull.* 43, 4 (2020), 39–50. http://sites.computer.org/debull/A20dec/p39.pdf
[84] Tolou Shadbahr, Michael Roberts, Jan Stanczuk, Julian Gilbey, Philip Teare, Sören Dittmer, Matthew Thorpe, Ramon Viñas Torné, Evis Sala, Pietro Lio, Mishal Patel, Jacobus Preller, James Rudd, Tuomas Mirtti, Antti Rannikko, John Aston, Jing Tang, and Carola-Bibiane Schönlieb. 2023. The impact of imputation quality on machine learning classifiers for datasets with missing values. *Communications medicine* 3 (10 2023), 139. https://doi.org/10.1038/s43856-023-00356-z
[85] Reza Shahbazian and Sergio Greco. 2023. Generative Adversarial Networks Assist Missing Data Imputation: A Comprehensive Survey & Evaluation. *IEEE Access* (2023).
[86] Ravid Shwartz-Ziv and Amitai Armon. 2022. Tabular data: Deep learning is not all you need. *Information Fusion* 81 (2022), 84–90.
[87] Daniel J Stekhoven and Peter Bühlmann. 2012. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* 28, 1 (2012), 112–118.
[88] Daniel J. Stekhoven and Peter Bühlmann. 2011. MissForest—nonparametric missing value imputation for mixed-type data. *Bioinformatics* 28, 1 (10 2011), 112–118. https://doi.org/10.1093/bioinformatics/btr597 arXiv:https://academic.oup.com/bioinformatics/article-pdf/28/1/112/50568519/bioinformatics_28_1_112.pdf
[89] Julia Stoyanovich, Serge Abiteboul, Bill Howe, H. V. Jagadish, and Sebastian Schelter. 2022. Responsible data management. *Commun. ACM* 65, 6 (2022), 64–74. https://doi.org/10.1145/3488717
[90] Bhekisipho Twala. 2009. An empirical comparison of techniques for handling incomplete data using decision trees. *Applied Artificial Intelligence* 23, 5 (2009), 373–405.
[91] Tom R. Tyler. 2006. *Why People Obey the Law.* Princeton University Press, Princeton, NJ.
[92] Jianwei Wang, Ying Zhang, Kai Wang, Xuemin Lin, and Wenjie Zhang. 2024. Missing Data Imputation with Uncertainty-Driven Network. *Proc. ACM Manag. Data* 2, 3, Article 117 (May 2024), 25 pages. https://doi.org/10.1145/3654920
[93] Jianwei Wang, Ying Zhang, Kai Wang, Xuemin Lin, and Wenjie Zhang. 2024. Missing Data Imputation with Uncertainty-Driven Network. *Proceedings of the ACM on Management of Data* 2, 3 (2024), 1–25.
[94] Yanchen Wang and Lisa Singh. 2021. Analyzing the impact of missing values and selection bias on fairness. *International Journal of Data Science and Analytics* 12 (08 2021), 1–19. https://doi.org/10.1007/s41060-021-00259-z
[95] Han-Jia Ye, Si-Yang Liu, Hao-Run Cai, Qi-Le Zhou, and De-Chuan Zhan. 2024. A closer look at deep learning on tabular data. *arXiv preprint arXiv:2407.00956* (2024).
[96] Jinsung Yoon, James Jordon, and Mihaela Schaar. 2018. Gain: Missing data imputation using generative adversarial nets. In *International conference on machine learning*. PMLR, 5689–5698.
[97] Yiliang Zhang and Qinqin Long. 2021. Assessing Fairness in the Presence of Missing Data. *Advances in neural information processing systems* 34 (2021), 16007–16019. https://api.semanticscholar.org/CorpusID:245006257
[98] Yiliang Zhang and Qi Long. 2021. Fairness in missing data imputation. *arXiv preprint arXiv:2110.12002* (2021).

[99] Yiliang Zhang and Qi Long. 2022. Fairness-aware missing data imputation. In *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*.

[100] He Zhao, Ke Sun, Amir Dezfouli, and Edwin V Bonilla. 2023. Transformed distribution matching for missing value imputation. In *International Conference on Machine Learning*. PMLR, 42159–42186.

[101] Helen Zhou, Sivaraman Balakrishnan, and Zachary Lipton. 2023. Domain adaptation under missingness shift. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 9577–9606.

[102] Youran Zhou, Sunil Aryal, and Mohamed Reda Bouadjenek. 2024. A Comprehensive Review of Handling Missing Data: Exploring Special Missing Mechanisms. *arXiv preprint arXiv:2404.04905* (2024).