Honda Research Institute EU
Honda Research Institute Europe GmbH
https://www.honda-ri.de/

Evaluating and Mitigating Fairness Risks in Machine Learning under Structured Missing Data

Tim Kochs, Andrea Castellani, Felix Lanfermann, Barbara Hammer

2025

Preprint:

This is an accepted article published in AEQUITAS 2025 . The final authenticated version is available online at: https://doi.org/[DOI not available]

Evaluating and Mitigating Fairness Risks in Machine Learning under Structured Missing Data

Tim Kochs[1,2], Andrea Castellani[2], Felix Lanfermann[2] and Barbara Hammer[1]

[1]AG Machine Learning, Bielefeld University, 33615, Bielefeld, Germany
[2] Honda Research Institute Europe GmbH, Carl-Legien-Str. 30, 63073, Offenbach am Main, Germany

Abstract

Handling missing values in machine learning presents significant challenges, impacting both model predictive performance and fairness. While unstructured missingness, where values are randomly absent, has been extensively studied, structured missingness remains relatively underexplored. Structured missingness occurs when the absence of one or more features influences the absence of others. In this paper, we investigate the impact of both structured and unstructured missingness on fairness. We introduce a novel approach for simulating structured missingness and propose a training methodology designed to enhance model robustness under these conditions. We evaluate how different state-of-the-art imputation methods for handling missing data affect fairness across multiple benchmark tabular datasets. Our empirical results demonstrate that structured missingness leads to a degradation in model fairness, particularly when the missingness mechanism is conditioned on protected attributes such as race or gender. In these cases, minority groups experience higher error rates, contributing to disparities in metrics like equalized odds. We propose a preprocessing step, Missingness Robustness Augmentation, that is shown to increase model robustness towards the presence of missing values.

Keywords

Machine Learning, Missing Values, Structured Missingness, Fairness, Imputation

1. Introduction

As machine learning systems increasingly influence high-stakes decisions in healthcare [1, 2], finance [3], and criminal justice [4, 5, 6], concerns about fairness and bias have become central to their responsible deployment [7, 8, 9]. Data, the foundational element of these systems, is often incomplete, leading to challenges in achieving fair and unbiased outcomes [10, 11, 12, 13]. In particular medical application data sets often encounter missing values due to high costs, limited availability, and a high variety of practices among medical entities, to name just a few mechanisms [14, 15, 16]. Conventionally, missing data is classified as missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR) [12, 11]. However, these frameworks often overlook a more nuanced form of incompleteness known as Structured Missingness (SM), recently introduced by [15, 17]. SM refers to scenarios where the absence of data is not merely random but exhibits complex, systematic patterns, often reflecting structural inequalities within the data collection pipeline.

This missingness mechanism is particularly prevalent in the medical domain [18, 19, 16]. To name just one example: in oncology, decisions regarding neoadjuvant therapya pre-surgical intervention such as chemotherapy or radiationdepend on a range of patient-specific factors. For instance, in rectal cancer, tumors located in anatomically challenging regions may require preoperative radiation to enable less invasive surgery. However, patients with certain comorbidities (e.g., cardiovascular disease or connective tissue disorders) may not be eligible for radiation, resulting in deviations from standard treatment protocols. These clinically informed decisions can lead to systematic and non-random patterns of missing data, particularly in treatment and outcomes records, which may introduce bias if not appropriately modeled.

AEQUITAS 2025: Workshop on Fairness and Bias in AI | co-located with ECAI 2025, Bologna, Italy
tkochs@techfak.uni-bielefeld.de (T. Kochs); andrea.castellani@honda-ri.de (A. Castellani); felix.lanfermann@honda-ri.de (F. Lanfermann); bhammer@techfak.uni-bielefeld.de (B. Hammer)
© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Current state-of-the-art methods for handling missing data are largely impute-then-classify approaches [20]. However, imputation methods are rarely evaluated on downstream-task-performance, but mostly on the reconstruction error or variations thereof [21, 22]. It is still an active area of research how imputation impacts machine learning models [23, 24, 25, 26], and there is even evidence, that switching imputation methods negatively impacts model performance [23].

Besides model performance, fairness concerns towards a machine learning model play an increasingly important role. It has already been extensively studied how imputation impacts fairness [27, 28, 8, 29]. Yet these approaches do not target SM, albeit it might play an important role in this context since minority groups might be more inclined to not provide an answer to certain questions or affected by a bias in the data collections process [15, 8]. Therefore, in this contribution, we address the widely open research question of how SM impacts the fairness of ML-models. Our main contributions are threefold:

1. We propose a novel procedure to introduce SM in data, which allows extensive evaluation of its effect across multiple datasets.
2. We empirically evaluate the impact of SM on the fairness of state-of-the-art imputation methods and show that fairness can deteriorate even when SM is present only during inference.
3. We present a novel preprocessing method, Missingness Robustness Augmentation, designed to improve robustness to test-time SM. Our method is model-agnostic and demonstrates consistent improvements in fairness metrics without degrading predictive performance.

2. Related Work

Types of missingness. The fundamental theoretical background of missingness mechanisms has been defined in [11, 12], coining the mechanism missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). In the recent work [15], the concern was raised that these definitions do not form a complete characterization of missingness mechanisms. This observation led to the definition of structured missingness (SM) [17], which extends the definition of missingness to also incorporate weak structures (WS) and strong structures (SS), i.e., probabilistic or deterministic relationships of features with missing values. These additional structures can occur with each of the missingness types in [11] i.e., MAR, MNAR.

Missing data imputation. Ignoring missing data can introduce a bias into statistical analyses. As demonstrated in [11], valid inference generally requires that the missing data mechanism is MAR and the missingness process is independent of the parameter of interest. Consequently, imputation techniques are commonly employed to address missing data [30, 20]. There is a wide variety of imputation methods in the literature [31, 32, 33, 34, 35, 36, 21, 26, 22], which can be broadly classified into statistical methods, general machine learning (ML) methods, or deep learning (DL) methods. In the statistical method, simple imputation, all missing values of a feature are replaced by a single value such as the mean or mode [30]. This is often used as a baseline as it might introduce bias in high dimensional datasets [37]. Methods like Multivariate Imputation by Chained Equations (MICE) [32] offer state-of-the-art statistical methods. Referring to ML, popular technologies include k-nearest-neighbors (KNN) imputation [38], MissForest [33], and SVM imputation [39]. DL methods are becoming increasingly popular at present, yet they require large data sets in order to outperform statistical imputation methods [40, 34, 41, 22, 42, 43].

Albeit imputation offers a versatile technique of dealing with missing values, there is some concern about bias that might be introduced when used in conjunction with a machine learning pipeline [23, 44]. Furthermore many imputation methods focus on imputing numerical values, such that their application to categorical values might introduce additional bias [45]. There are

also few approaches that do not require a separate imputation step, but are capable of dealing with missing values natively [46, 47, 48]. Those are just mentioned for completeness, since these methods are not within the scope of our contribution.

Fairness. There exist various (often mutually exclusive) formalizations of fairness in machine learning which account for different normative intuitions and practicability [49]. A common distinction refers to group fairness metrics (e.g., demographic parity, equalized odds) [50], which refers to distinct groups of persons as characterized by a specific sensitive attribute such as gender or ethnicity, and individual fairness metrics (e.g., counterfactual fairness, consistency) [51]), which focus on a fair treatment of individual persons in comparisons to the whole group. Due to their efficient evaluation, group fairness methods such as equalized odds which essentially refer to the distribution of errors among the vulnerable group as compared to the reference group, enjoy a wide popularity. In contrast, individual fairness measures which rely on semantically meaningful modeling of the scenario such as a latent causal model as used within counterfactual fairness, are sometimes praised as particularly meaningful, but might suffer from lack of causal information, or even reduce to statistical group methods in practice [52].

Fairness and missing values. As pointed out in [53], missing values are inherently related to the topic of algorithmic fairness. Currently, it is common practice within the fairness literature to simply remove missing values [8], albeit it likely introduces bias into the data [12]. A growing body of literature is concerned with missing values in fair machine learning [54, 28, 29, 44, 8]. Some fairness measures have already been adapted for their use with missing values [55], and first fairness-aware imputation methods have been proposed [56, 28].

Yet, there is a lack of work evaluating common imputation methods on downstream task performance, with recent work providing explorations of different missingness mechanism [20]. In this contribution, we extend this effort by exploring downstream task performance and fairness, whereby we put a particular focus on SM and the impact of missing values during both training and test-time. Moreover, as a first remedy, we offer insights into robust training under the assumption of structured missingness.

3. Proposed Methods

3.1. Problem Statement and Notation

Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ denote a dataset with n samples and d features. The missingness in the data is represented by a binary indicator matrix $\mathbf{M} \in \{0,1\}^{n \times d}$, where $M_{ij} = 1$ if X_{ij} is missing and 0 otherwise. The primary missing data mechanisms are the following:

Missing Completely at Random (MCAR): The probability of missingness is independent of both observed and unobserved data. Formally,

$$P(\mathbf{M} \mid \mathbf{X}) = P(\mathbf{M})$$

Under MCAR, the missingness does not introduce bias in parameter estimates, allowing for unbiased analysis using complete cases [11].

Missing at Random (MAR): The probability of missingness may depend on observed data but not on the missing data itself:

$$P(\mathbf{M} \mid \mathbf{X}) = P(\mathbf{M} \mid \mathbf{X}_{\text{obs}})$$

This assumption permits the use of methods like multiple imputation to obtain unbiased estimates, provided that the model for the missingness mechanism is correctly specified [12].

Missing Not at Random (MNAR): The probability of missingness depends on unobserved data:

$$P(\mathbf{M} \mid \mathbf{X}) \neq P(\mathbf{M} \mid \mathbf{X}_{\text{obs}})$$

MNAR mechanisms require modeling the missingness process for unbiased estimates, as the missingness is informative and cannot be ignored [12].

Structured missingness (SM) refers to scenarios where the pattern of missing data exhibits dependencies among variables or across observations. Each version of structure can be applied to each missingness mechanism, as example we state the definition of MAR. The two notable structure types are:

MAR with Weak Structure (MAR-WS): Missingness in one variable depends on observed data and the missingness indicators of other variables:

$$P(\mathbf{M}_j \mid \mathbf{X}_{\text{obs}}, \mathbf{M}_{-j})$$

where $\mathbf{M}_j$ is the missingness indicator for variable j, and $\mathbf{M}_{-j}$ represents missingness indicators for other variables [17].

MAR with Strong Structure (MAR-SS): Missingness in one variable depends on observed data, missingness indicators, and the missing values of other variables:

$$P(\mathbf{M}_j \mid \mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}}, \mathbf{M}_{-j})$$

This scenario is more complex and requires advanced modeling techniques to handle the dependencies effectively [17].

3.2. Generation of Benchmarks with Structured Missingness

Having SM in the data can have a significant impact on the fairness of a predictive model, especially if the protected attribute is related to the missingness structure. In order to systematically analyze SM in a controlled environment, we introduce Structured Feature Dropout, detailed in algorithm 1, a novel method to induce SM within datasets. This approach centers on a sensitive attribute to simulate realistic missingness patterns pertinent to algorithmic fairness studies. While tailored for fairness evaluations, the methodology is adaptable for broader applications involving structured missingness.

First of all, to induce a structure in the missingness mechanism, a domain knowledge of the problem is needed. We propose to construct a Bayesian Network (BN), a probabilistic graphical model that represents attributes as random variables and their conditional dependencies. This approach aims to maximize the likelihood of generating the observed dataset. Specifically, we utilize the Bayesian networks previously developed by [53]. Although BNs are used here for their flexibility in capturing feature dependencies, other graph-based approaches may also be suitable. We would like to stress that it is not necessary to construct a complete causal model to implement this procedure, but any knowledge of protected features would suffice.

To induce missingness, first, we generate an unstructured missingness mask by randomly masking an γ fraction of features correlated with the sensitive attribute. Next, we induce structured missingness by masking an additional θ fraction of features directly related to the target label, reflecting the worst-case scenario in which the missingness mechanism directly impacts predictive features.

Figure 1 provides a visual illustration of the procedure described in algorithm 1. By incorporating domain knowledge and data-driven insights, we construct a graphical model that captures feature dependencies. Missingness is then introduced according to a specified mechanism; for instance, in the depicted example, a Missing At Random (MAR) strategy is employed,

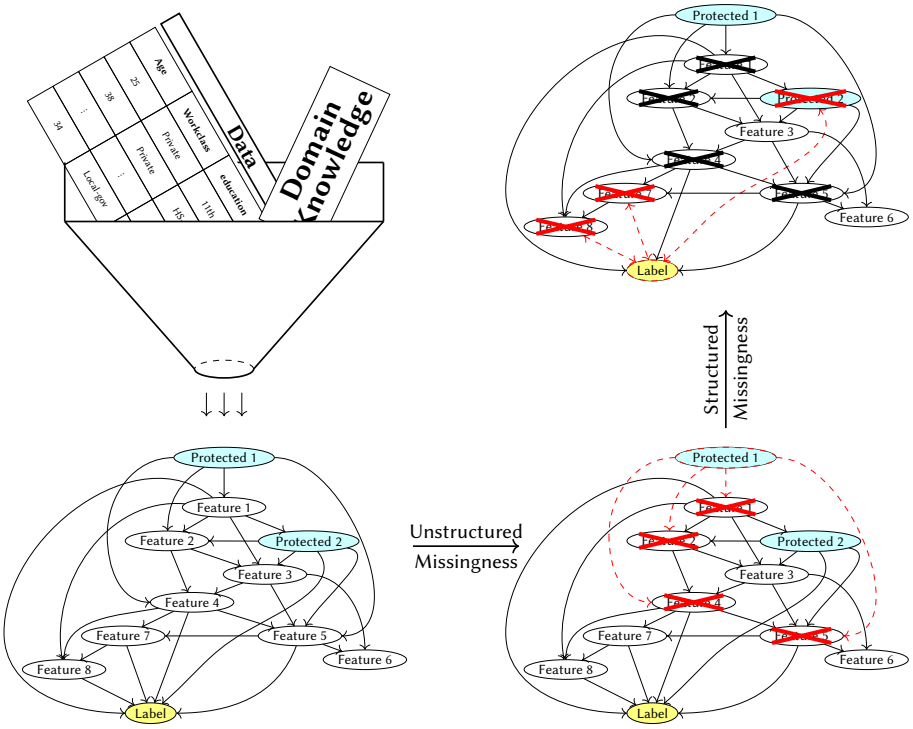


Figure 1: Inducing structures missingness is a two step process first unstructured missingness is applied based on the first probabilistic round a second set of features is dropped.

where the probability of missingness depends on an observed variable, specifically the attribute labeled "Protected 1". Red dashed edges indicate the dependency paths used to determine which features are selected for missingness in a given step, while red crosses denote features in which missing values were introduced at that stage. Red crosses indicate features in which missing values were artificially introduced according to the missingness mechanism. This process results in a missingness mask $\mathbf{M}$, with selected feature values removed. In this simulation, SM was introduced solely based on Protected 1, ensuring that Feature 3 remained fully observed throughout. For clarity and consistency across datasets we selected only a single protected attribute; the approach naturally extends to settings involving multiple protected variables. In a second iteration, missingness is induced in features that are directly associated with the target label, thereby simulating a more structurally dependent form of missingness. Features previously marked with black crosses remain unchanged and do not receive additional missing values.

Algorithm 1: Structured Feature Dropout

Input: Dataset $\mathscr{D} \in \mathbb{K}^{n \times d}$; probability distributions γ, θ
Output: Modified dataset $\mathscr{D}$ with structured missingness (SM)
clean($\mathbf{X}$) ; `// Optional preprocessing or normalization`
$F \leftarrow$ `computeFeatureRelation`($\mathscr{D}$) ; `// Determine feature relationships`
$\mathbf{M} \leftarrow$ `missingMask`($\mathscr{D}, F, \gamma$) ; `// Apply base-level missingness`
$\mathbf{M}' \leftarrow$ `structuredMask`($\mathbf{M}, F, \theta$) ; `// Inject structured missingness`
setMissing($\mathscr{D}, \mathbf{M} + \mathbf{M}'$) ; `// Apply total missingness mask`
return $\mathbf{D}$

3.3. Robust Training for Missing Values

Whenever missing values occur during deployment of a machine learning model, it might be susceptible to reduced performance and fairness. To counteract this effect, we propose Missingness

Robustness Augmentation, a novel method which aims to increase the robustness of an ML model when dealing with structured missingness. The core idea of our proposed method is to enrich the training data with artificially induced missing values. By exposing the model to these conditions during training, we improve its ability to generalize when faced with similar missingness during inference. Importantly, our approach enhances fairness robustness without compromising predictive performance.

The proposed method is listed in algorithm 2. Missingness Robustness Augmentation augments the training data by randomly sampling a proportion α of the dataset and artificially removing a predefined set of features from these samples. This induces additional structured missingness (SM) in the training set, informed by domain knowledge. By doing so, the model is exposed to missingness patterns that resemble those expected at deployment, thereby improving its robustness to fairness and performance degradation under test-time SM.

This algorithm is particularly effective if there is a distributional shift of the missingness pattern from the training data to the testing data. This shift is reduced by adding randomly drawn samples from the dataset and reinserting those into the training set, but with certain features masked/missing.

Training a model on data that includes imputed values can enhance its robustness, provided that the same imputation strategy is consistently applied during both training and testing phases. algorithm 2 is to be applied before the initial imputation step takes place. Exposing the model to a sufficient volume of imputed data during training, allows it to better approximate the potentially skewed distribution it may encounter in real-world scenarios.

The core insight underlying our method is that, in many practical applications, the patterns of structured missingness are at least partially understood or can be inferred from domain expertise [15]. Leveraging this knowledge to simulate realistic missingness during training enables a straightforward yet effective approach to improving model robustness under structured missingness, as supported by empirical results in this and prior work.

In cases where this distribution is unknown, feature selection methods and uniform sampling across the entire dataset may provide reasonable performance. However, care must be taken when determining the number of augmented samples, as excessive additions may introduce noise and increase the computational burden, while insufficient augmentation may fail to capture the underlying data distribution.

Algorithm 2: Missingness Robustness Augmentation

Input: Dataset $\mathscr{D}$ with n rows; missingness ratio $\alpha \in [0, 1]$; feature subset $\mathscr{F} \subseteq$ columns of $\mathscr{D}$
Output: Modified dataset $\tilde{\mathscr{D}}$ with additional missingness
$m \leftarrow \lfloor \alpha \cdot n \rceil$; `// Number of samples to modify`
$\mathscr{D}_{\text{miss}} \leftarrow$ sample m rows uniformly at random from $\mathscr{D}$
foreach $x \in \mathscr{D}_{\text{miss}}$ do
  foreach $f \in \mathscr{F}$ do
    $x[f] \leftarrow$ `missing`
Replace original rows in $\mathscr{D}$ with modified $\mathscr{D}_{\text{miss}}$ to obtain $\tilde{\mathscr{D}}$
return $\tilde{\mathscr{D}}$

4. Experimental Setup

To analyze how SM impacts the fairness of binary classification tasks, we set up a data processing pipeline that was applied to a variety of imputation methods, classifiers, and datasets. The selection of data sets is based on the survey [53], where the authors also provide the bayesian network for each dataset. Both the imputation methods and the classifiers were selected from state-of-the-art methods in the literature. An overview of the datasets used in our experiments is provided in Table 1. The column with the number of entries refers to the cleaned version

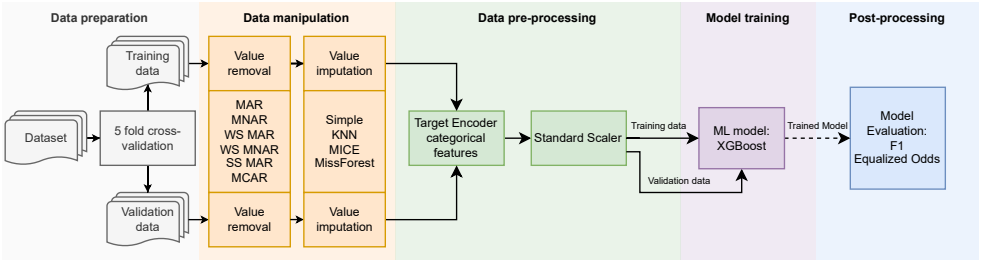


Figure 2: Experimental setup pipeline proposed in this work.

Table 1
Datasets used and some information about them.

Dataset	#Entries	Domain	Protected Attribute	Target Variable
Adult	45,222	Finance	Sex	Income
Bank Marketing	45,211	Finance	Marital	Deposit Subscription
German Credit	1,000	Finance	Sex	Credit Score
Credit card clients	30,000	Finance	Sex	Default payment
Ricci	118	Society	Race	Promotion
Student-Mathematics	649	Education	Sex	Final grade
Student-Portoguese	649	Education	Sex	Final grade

of each dataset, with all samples containing missing values removed prior to experimentation. Although several datasets include multiple protected attributes, we report only the specific attribute used to induce missingness in our experiments, as this was the sole attribute considered when evaluating fairness metrics. The code and additional results of this work are available at https://github.com/fairness_SM.

We use imputation methods from statistics and classical machine learning. Due to the small size of the data, we do not use DL imputation [22]. From the statistical realm we employed simple imputation (using mean for numeric features and mode for categorical) and MICE [32] while using MissForest [33] and KNN [38] from the machine learning domain.

To employ KNN imputation, we slightly modified the approach to work with categorical features. Every categorical feature is one-hot encoded, and during imputation, all encoded values are set to missing and imputed individually. Then we compute their maximum, and the final imputed value is the corresponding categorical feature.

Pipeline. To set up a controlled environment we applied similar data cleaning steps as in [53], removing all missing values which are included in the dataset. To ensure consistency and reduce statistical noise we employed a 5-fold cross-validation-scheme. The original dataset was first partitioned into training and testing subsets, after which missingness was artificially introduced into both. A detailed overview of the entire pipeline can be seen in Figure 2, it consists of 5 steps: (1) Data preparation, to split the dataset in train/validation with the 5-fold cross validation strategy. (2) Data manipulation, to first induce the missing values in train and validation data, with multiple percentage (0%, 10%,30%, 50%, 70%), and then impute the missing values. (3) Data pre-processing, using target encoder [57], to encode categorical features, and then standardize the data to zero mean and unitary varriance. (4) Training of the ML model XGBoost, as it is the state-of-the-art in ML for tabular datasets [58]. (5) Post-processing, to evaluate the trained model with fairness and performance metrics.

Preprocessing. Following the 5-fold cross-validation split, missingness induction, and imputation, we applied a standardized preprocessing pipeline. Given the presence of categorical features in most datasets, we employed target encoding to transform these variables, as it is considered a state-of-the-art approach for supervised learning tasks [59]. Finally, the resulting data were standardized to have zero mean and unit variance to ensure consistent model performance. The full pipeline, outlined in Figure 2, consisted of 5 steps: To ensure controlled experimentation, we eliminate all initial missing values from the dataset before invoking algorithm 1. This step isolates the effects of induced structured missingness (SM), allowing for clearer attribution of observed outcomes to the experimental conditions. Specifically, we avoid mixing different missingness mechanisms to focus exclusively on the impact of SM on fairness.

Missing Mechanisms. Our primary objective is to evaluate the practical implications of SM in real-world applications. Although it is theoretically possible to construct a structured version of MCAR, such a formulation holds limited practical significance. Consequently, we exclude structured MCAR from our analysis. Instead, we focus on MAR and MNAR mechanisms, where missingness is conditioned on or influenced by the protected attribute. For MNAR scenarios, we simulate a latent confounder by removing the protected attribute from the dataset after inducing missingness. This reflects settings in which sensitive variables affect the missingness process but are not explicitly available at training or inference time. We evaluate both weakly and strongly structured variants of MAR and MNAR. Missingness is introduced using a unified procedure (algorithm 1) that generates both structured and unstructured missingness masks in a consistent and reproducible manner.

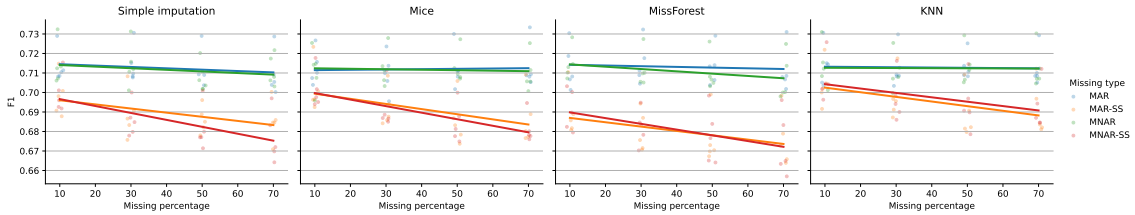
Evaluation Metrics As our primary interest lies in evaluating downstream task performance and fairness, we do not consider reconstruction error, commonly used to assess imputation quality, as it does not necessarily correlate with performance or fairness in predictive modeling tasks. Instead, we focus on well-established metrics from the machine learning literature that directly reflect model behavior: the F1 score for predictive performance and Equalized Odds (EO) for fairness, as we are in a supervised learning setting [60]. EO evaluates fairness based on error rate parity across groups, offering a measure of fairness in classification tasks, particularly in domains where label information is meaningful, and error disparities are critical. For fairness evaluation, we focus on group-level metrics rather than individual fairness. Although individual fairness offers a more fine-grained notion of equitable treatment, it often relies on access to a well-defined causal model, an assumption that is frequently impractical or unverified in real-world applications.

5. Evaluation and Results Discussion

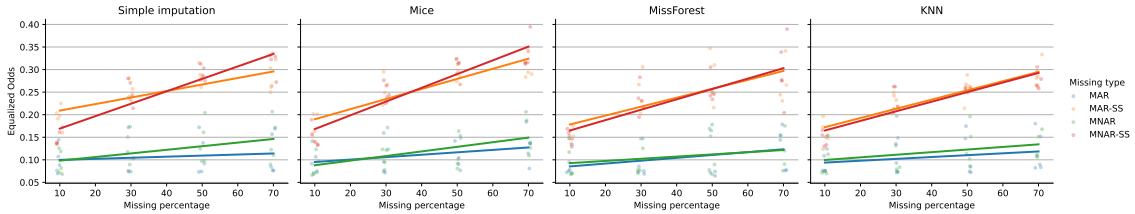
In this section, we present and analyze our findings on the impact of SM. We begin by evaluating the effect of different imputation strategies employed during preprocessing. Subsequently, we investigate the model's behavior under varying degrees of missingness in both the training and test sets, with a particular focus on the asymmetry between the two.

For simplicity, we discuss the results on the Adult dataset [61], as it is a common benchmark in the fairness literature [53], and the trend of the results is qualitatively similar to the other datasets investigated in this work, listed in Table 1.

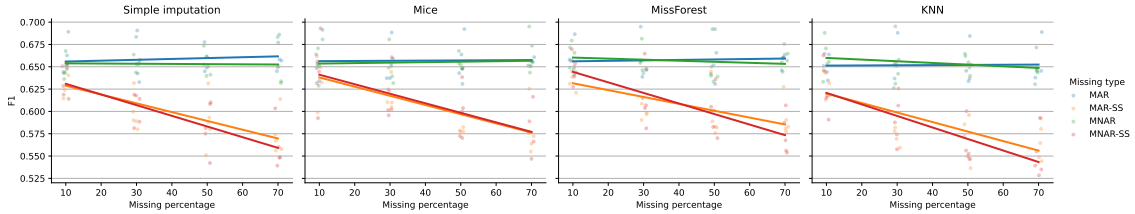
Shortcoming of Imputation Methods. Figure 3 presents both predictive performance (F1-score) and group-level fairness metrics (EO) for the Adult and Bank marketing dataset, under conditions when training and test sets have matching levels of missingness. Across all imputation strategies evaluated, a consistent trend emerges: SM leads to the most pronounced degradation in both performance and fairness metrics. This pattern is clearly reflected in Figure 3, where



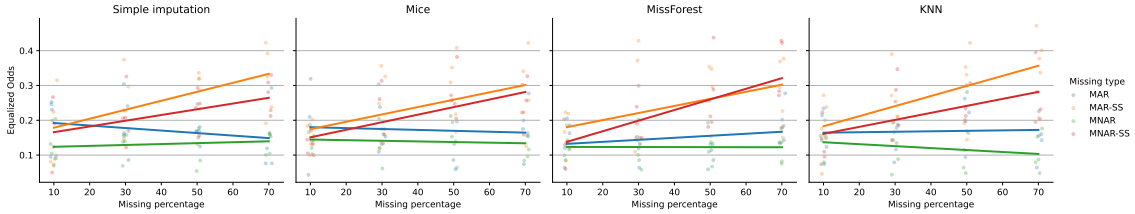
(a) F1-score on Adult dataset.



(b) EO on Adult dataset.



(c) F1-score on Bank marketing dataset.



(d) EO on Bank marketing dataset.

Figure 3: F1 score and Equalized Odds (EO) across different missingness mechanisms, evaluated by imputation method.

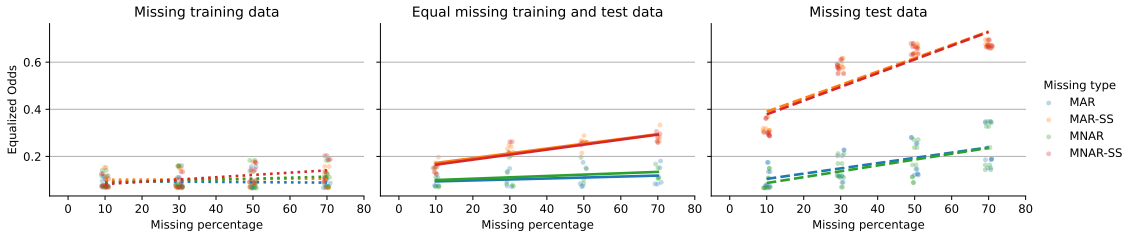
increasing overall missingness correlates with declines in F1 score and EO. The consistency of this trend is evident across multiple datasets, including representative results from the Adult and Bank Marketing datasets. These results align with prior findings [25], which suggest that higher imputation complexity does not necessarily yield improved downstream performance.

Shifts in Missingness. The assumption that training and test sets exhibit identical levels of missingness is rarely met in practice. To evaluate the robustness of the model under more realistic conditions, we performed experiments across a range of missingness levels in both training and test data. A central finding is that the presence and distribution of missing values at inference time have a disproportionately greater impact on both predictive performance and fairness metrics compared to missingness during training. Specifically, models trained on complete data but evaluated on partially observed test data exhibited the most significant degradation in fairness. Figure 4 summarizes results across three experimental conditions: missingness introduced exclusively during training, symmetrically during both training and testing, and exclusively at test time, all of which used simple imputation.

Our results indicate that missingness confined to training had negligible effects on performance and fairness, while test-time missingness introduced considerably more degradation across metrics.



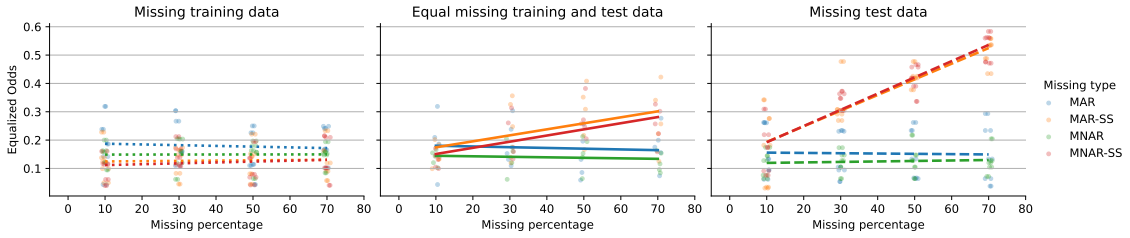
(a) F1 score one Adult dataset.



(b) EO on Adult dataset.



(c) F1 score one Bank marketing dataset.



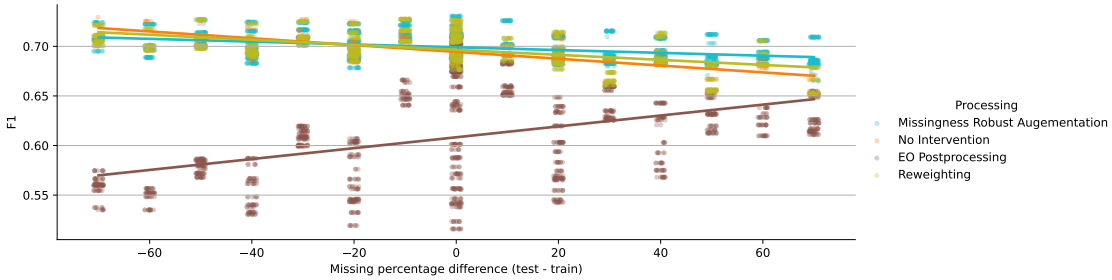
(d) EO on Adult dataset.

Figure 4: Impact of missingness placement (train vs. test) on F1, and EO metrics.

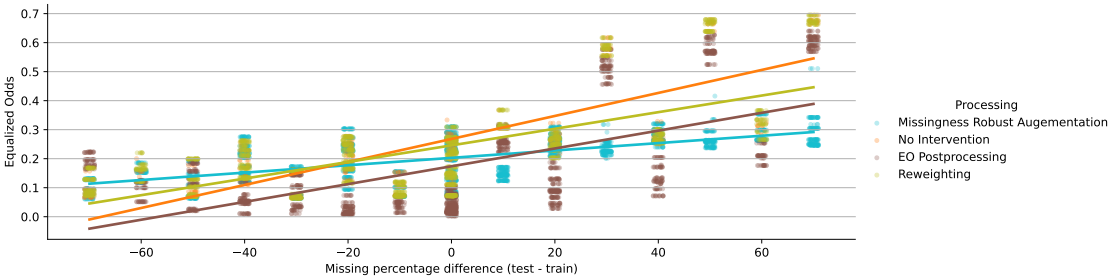
Furthermore, the magnitude of this effect scales with the degree of mismatch between training and test missingness levels. This is particularly problematic in real-world deployments, where ground-truth labels are unavailable at test time, limiting the ability to detect post-hoc fairness degradation. These findings underscore the importance of anticipating structured missingness and implementing mitigation strategies during the model development phase, rather than relying solely on downstream audits.

5.1. Toward Robust Fairness in Incomplete Data Settings

Setup. Motivated by the observed limitations of existing techniques under structured missingness, we developed Missingness Robustness Augmentation, a preprocessing approach aimed at improving robustness and fairness. For clarity, recall that our proposed method is applied in step two of the pipeline proposed in Figure 2. To evaluate its effectiveness, we conducted a comparative analysis against three baseline strategies widely used in fair machine learning: (i) a standard model without any fairness intervention, (ii) a reweighting scheme that adjusts the



(a) f1



(b) Equalized Odds

Figure 5: F1 and EO by missingness difference (test - train). Missingness Robustness Augmentation performs best when there is a substantial mismatch in missingness between training and test sets.

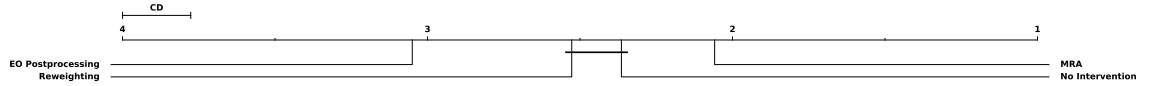
influence of training samples to mitigate biasparticularly by assigning higher weights to instances from the minority group with missing values [62], and (iii) a post-processing method that enforces the EO criterion by modifying classifier outputs [63]. For our evaluation of Missingness Robustness Augmentation we set α to about 10% of total dataset size.

Evaluation. Figure 5 summarizes the overall performance of all data processing strategies evaluated. The x-axis denotes the difference in missingness rates between the training and test sets; higher values indicate a greater proportion of missing data in the test set relative to the training set.

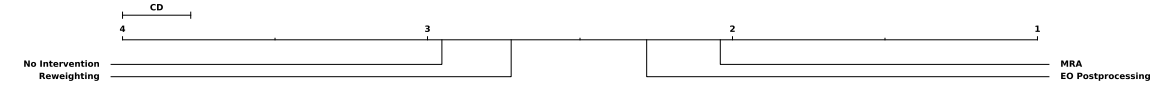
The reweighting strategy utilized in our study is based on the presence of missing values in the training data, applying more weight to data points with missing values. Similarly, when applying Equalized Odds (EO) postprocessing, discrepancies between the distribution of missingness in the training and test sets diminish the fairness improvements while also negatively impacting model performance. In contrast, Missingness Robustness Augmentation remains effective under such distribution shifts, provided the augmented features accurately reflect the missingness patterns expected at test time. However, in scenarios where this alignment cannot be guaranteed, our results (see Figure 4) indicate only marginal effects on both fairness and performance. Consequently, incorporating Missingness Robustness Augmentation presents minimal risk but offers potential benefits, especially in deployment settings where structured missingness may emerge.

Significance. In order to assess the statistical significance of the results across multiple datasets, we use the Friedman non-parameteric test with 0.05 confidence level, followed by Nemenyi post-hoc test [64], and the we plot the results with the Critical Distance (CD) plot in Figure 6. We focus on where the test set exhibits a higher proportion of missing values than the training set, we observe significant differences among data processing methods in both predictive performance and fairness.

Although EO post-processing achieves the best fairness outcomes, it does so at the cost of



(a) Critical distance for F1 score.



(b) Critical distance for EO.

Figure 6: Evaluation of data processing strategies under asymmetric missingness conditions, with test sets exhibiting greater missingness than training sets. Based on all datasets listed in Table 1. Missingness Robustness Augmentation.

reduced accuracy. In contrast, the proposed Missingness Robustness Augmentationoffers the second-best performance in terms of fairness, while also outperforming all other methods in predictive accuracy. This suggests that our method provides a favorable balance between fairness and performance under conditions of mismatched missingness.

6. Conclusion

Structured missingness arises from complex real-world processes and interactions, posing unique challenges for machine learning systems, as highlighted by [15]. This study demonstrates that SM in data can have a substantial impact on both predictive performance and fairness in machine learning systems, particularly when it occurs during inference. Through empirical evaluation, we show that disparities in missingness patterns between training and deployment environments can introduce or exacerbate group-level disparities in model outcomes. To investigate this systematically, we propose a novel methodology for inducing structured missingness, which links the missingness mechanism explicitly to the protected attribute.

To address the effects of SM we introduce a simple and generalizable preprocessing method designed to improve model robustness in the presence of structured missingness. Our approach requires no architectural changes, is easy to implement, and maintains predictive accuracy across tasks. While improvements in fairness are modest, the method yields consistent and stable gains across a range of missingness scenarios, and outperforms standard reweighting and post-processing baselines in terms of reliability and sensitivity to test-time conditions. As our approach does not rely on a complete causal model, we deliberately refrained from evaluating counterfactual fairness, though exploring this perspective remains a promising direction for future work.

Our findings suggest that this method is most beneficial in settings where test-time missingness is prevalent and unevenly distributed across groups, which is common in real-world deployment contexts. Additionally, performance is determined by the alignment between the induced and the test-time missingness. When the induced distribution deviated significantly from the true test-time distribution, the impact on the metrics evaluated was minimal. In contrast, when the distributions aligned closely, we observed a clear improvement in these metrics.

Structured missingness remains an underexplored but pressing challenge in fair machine learning, and our results highlight the need for proactive, pipeline-level strategies that account for missingness during both development and deployment phases.

Our proposed method relies on domain knowledge to implement realistic and effective preventive strategies. Designing approaches that are agnostic to such prior knowledge or that leverage learned feature representations to approximate it remains an important direction for future work. Additionally, addressing the challenges posed by structured missingness when both the training and test sets exhibit comparable levels of missingness remains a complex and unresolved issue.

Future work should explore dynamic or adaptive methods that respond to missingness patterns at inference time, as well as deeper theoretical frameworks to understand the interaction between data missingness and fairness constraints. Another practical suggestion is to explore the impact of structured missingness on deep learning systems, as such systems gain more traction when dealing with tabular data.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] B. Balbierer, L. Heinlein, D. Zipperling, N. Kühl, A Multivocal Literature Review on Privacy and Fairness in Federated Learning (2024).

[2] A. Guerra-Manzanares, L. J. L. Lopez, M. Maniatakos, F. E. Shamout, Privacy-preserving machine learning for healthcare: open challenges and future perspectives, volume 13932, 2023, pp. 25–40. URL: http://arxiv.org/abs/2303.15563. doi:10.1007/978-3-031-39539-0_3, arXiv:2303.15563 [cs].

[3] C. John, E. J. Ekpenyong, C. C. Nworu, Imputation of Missing Values in Economic and Financial Time Series Data Using Five Principal Component Analysis (PCA) Approaches, Central Bank of Nigeria Journal of Applied Statistics (2019) 51–73. URL: https://www.cbn.gov.ng/Out/2019/STD/51%20-%2073_John%20et%20al.pdf. doi:10.33429/Cjas.10119.3/6.

[4] K. Bell, J. Hong, N. McKeown, C. Voss, A New Direction for Machine Learning in Criminal Law (2021).

[5] F. Dakalbab, M. Abu Talib, O. Abu Waraga, A. Bou Nassif, S. Abbas, Q. Nasir, Artificial intelligence & crime prediction: A systematic literature review, Social Sciences & Humanities Open 6 (2022) 100342. URL: https://linkinghub.elsevier.com/retrieve/pii/S2590291122000961. doi:10.1016/j.ssaho.2022.100342.

[6] G. V. Travaini, F. Pacchioni, S. Bellumore, M. Bosia, F. De Micco, Machine Learning and Criminal Justice: A Systematic Review of Advanced Methodology for Recidivism Risk Prediction, International Journal of Environmental Research and Public Health 19 (2022) 10594. URL: https://www.mdpi.com/1660-4601/19/17/10594. doi:10.3390/ijerph191710594.

[7] F. Fioretto, C. Tran, P. V. Hentenryck, K. Zhu, Differential Privacy and Fairness in Decisions and Learning Tasks: A Survey, in: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, 2022, pp. 5470–5477. URL: http://arxiv.org/abs/2202.08187. doi:10.24963/ijcai.2022/766, arXiv:2202.08187 [cs].

[8] F. Martínez-Plumed, C. Ferri, D. Nieves, J. Hernández-Orallo, Fairness and Missing Values, 2019. URL: http://arxiv.org/abs/1905.12728. doi:10.48550/arXiv.1905.12728, arXiv:1905.12728 [cs].

[9] M. B. Zafar, I. Valera, M. G. Rodriguez, K. P. Gummadi, Fairness Constraints: Mechanisms for Fair Classification, 2017. URL: http://arxiv.org/abs/1507.05259, arXiv:1507.05259 [stat].

[10] D. C. Howell, The Treatment of Missing Data, in: The SAGE Handbook of Social Science Methodology, SAGE Publications Ltd, 1 Oliver's Yard,ă55 City Road,ăLondonăEnglandăEC1Y 1SPăUnited Kingdom, 2007, pp. 212–226. URL: https://methods.sagepub.com/book/the-sage-handbook-of-social-science-methodology/n11.xml. doi:10.4135/9781848607958.n11.

[11] D. B. Rubin, Inference and Missing Data (1976).

[12] R. Little, D. Rubin, Statistical analysis with missing data, Wiley series in probability and mathematical statistics. Probability and mathematical statistics, Wiley, 2002. URL: http://books.google.com/books?id=aYPwAAAAMAAJ.

[13] I. F. Ilyas, X. Chu, Data Cleaning, Association for Computing Machinery, New York, NY, USA, 2019.

[14] Y. Xin, R. Song, J. Hao, W. Li, C. Wu, L. Zuo, Y. Cai, X. Zhang, H. Wu, W. Hui, Poor reporting quality and high proportion of missing data in economic evaluations alongside pragmatic trials: a cross-sectional survey, BMC Medical Research Methodology 25 (2025) 61. URL: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-025-02519-z. doi:10.1186/s12874-025-02519-z.

[15] R. Mitra, S. F. McGough, T. Chakraborti, C. Holmes, R. Copping, N. Hagenbuch, S. Biedermann, J. Noonan, B. Lehmann, A. Shenvi, X. V. Doan, D. Leslie, G. Bianconi, R. Sanchez-Garcia, A. Davies, M. Mackintosh, E.-R. Andrinopoulou, A. Basiri, C. Harbron, B. D. MacArthur, Learning from data with structured missingness, Nature Machine Intelligence 5 (2023) 13–23. URL: https://www.nature.com/articles/s42256-022-00596-z. doi:10.1038/s42256-022-00596-z.

[16] L. Radosavljevi, S. M. Smith, T. E. Nichols, A generative model for evaluating missing data methods in large epidemiological cohorts, BMC Medical Research Methodology 25 (2025) 34. URL: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-025-02487-4. doi:10.1186/s12874-025-02487-4.

[17] J. Jackson, R. Mitra, N. Hagenbuch, S. McGough, C. Harbron, A Complete Characterisation of Structured Missingness, 2023. URL: http://arxiv.org/abs/2307.02650, arXiv:2307.02650 [stat].

[18] E. J. Topol, High-performance medicine: the convergence of human and artificial intelligence, Nature Medicine 25 (2019) 44–56. URL: https://www.nature.com/articles/s41591-018-0300-7. doi:10.1038/s41591-018-0300-7.

[19] L. A. Vale Silva, K. Rohr, Pan-Cancer Prognosis Prediction Using Multimodal Deep Learning, in: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), IEEE, Iowa City, IA, USA, 2020, pp. 568–571. URL: https://ieeexplore.ieee.org/document/9098665/. doi:10.1109/ISBI45749.2020.9098665.

[20] F. A. Khan, D. Herasymuk, N. Protsiv, J. Stoyanovich, Still More Shades of Null: An Evaluation Suite for Responsible Missing Value Imputation, 2025. URL: http://arxiv.org/abs/2409.07510. doi:10.48550/arXiv.2409.07510, arXiv:2409.07510 [cs].

[21] W.-C. Lin, C.-F. Tsai, Missing value imputation: a review and analysis of the literature (20062017), Artificial Intelligence Review 53 (2020) 1487–1509. URL: http://link.springer.com/10.1007/s10462-019-09709-4. doi:10.1007/s10462-019-09709-4.

[22] Y. Sun, J. Li, Y. Xu, T. Zhang, X. Wang, Deep learning versus conventional methods for missing data imputation: A review and comparative study, Expert Systems with Applications 227 (2023) 120201. URL: https://linkinghub.elsevier.com/retrieve/pii/S0957417423007030. doi:10.1016/j.eswa.2023.120201.

[23] J. Josse, J. M. Chen, N. Prost, G. Varoquaux, E. Scornet, On the consistency of supervised learning with missing values, Statistical Papers 65 (2024) 5447–5479. URL: https://link.springer.com/10.1007/s00362-024-01550-4. doi:10.1007/s00362-024-01550-4.

[24] P. J. García-Laencina, J. Morales-Sánchez, R. Verdú-Monedero, J. Larrey-Ruiz, J.-L. Sancho-Gómez, A. R. Figueiras-Vidal, Classification with Incomplete Data:, in: E. S. Olivas, J. D. M. Guerrero, M. Martinez-Sober, J. R. Magdalena-Benedito, A. J. Serrano López (Eds.), Handbook of Research on Machine Learning Applications and Trends, IGI Global, 2010, pp. 147–175. URL: http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-60566-766-9.ch007. doi:10.4018/978-1-60566-766-9.ch007.

[25] D. Bertsimas, A. Delarue, J. Pauphilet, Simple Imputation Rules for Prediction with Missing Data: Contrasting Theoretical Guarantees with Empirical Performance, 2024. URL: http://arxiv.org/abs/2104.03158. doi:10.48550/arXiv.2104.03158, arXiv:2104.03158 [stat].

[26] M. L. Morvan, J. Josse, E. Scornet, G. Varoquaux, What's a good imputation to predict with missing values?, 2021. URL: http://arxiv.org/abs/2106.00311. doi:10.48550/arXiv.2106.00311, arXiv:2106.00311 [stat].

[27] S. Caton, S. Caton, U. Ie, S. Malisetty, U. Edu, C. Haas, C. Haas, W. A. At, Impact of

Imputation Strategies on Fairness in Machine Learning (2022).
[28] Y. Zhang, Q. Long, Fairness in Missing Data Imputation, 2021. URL: http://arxiv.org/abs/2110.12002, arXiv:2110.12002 [cs].
[29] M. Fernando, F. Cèsar, N. David, H. José, Missing the missing values: The ugly duckling of fairness in machine learning, International Journal of Intelligent Systems 36 (2021) 3217–3258. URL: https://onlinelibrary.wiley.com/doi/10.1002/int.22415. doi:10.1002/int.22415.
[30] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, O. Tabona, A survey on missing data in machine learning, Journal of Big Data 8 (2021) 140. URL: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00516-9. doi:10.1186/s40537-021-00516-9.
[31] S. Alam, M. S. Ayub, S. Arora, M. A. Khan, An investigation of the imputation techniques for missing values in ordinal data enhancing clustering and classification analysis validity, Decision Analytics Journal 9 (2023) 100341. URL: https://linkinghub.elsevier.com/retrieve/pii/S2772662223001819. doi:10.1016/j.dajour.2023.100341.
[32] S. V. Buuren, K. Groothuis-Oudshoorn, mice : Multivariate Imputation by Chained Equations in R, Journal of Statistical Software 45 (2011). URL: http://www.jstatsoft.org/v45/i03/. doi:10.18637/jss.v045.i03.
[33] D. J. Stekhoven, P. Bühlmann, MissForestnon-parametric missing value imputation for mixed-type data, Bioinformatics 28 (2012) 112–118. URL: https://academic.oup.com/bioinformatics/article/28/1/112/219101. doi:10.1093/bioinformatics/btr597.
[34] E. Hallaji, R. Razavi-Far, M. Saif, DLIN: Deep Ladder Imputation Network, IEEE Transactions on Cybernetics 52 (2022) 8629–8641. URL: https://ieeexplore.ieee.org/document/9370000/. doi:10.1109/TCYB.2021.3054878.
[35] W. Du, D. Cote, Y. Liu, SAITS: Self-Attention-based Imputation for Time Series, Expert Systems with Applications 219 (2023) 119619. URL: http://arxiv.org/abs/2202.08516. doi:10.1016/j.eswa.2023.119619, arXiv:2202.08516 [cs].
[36] V. Fortuin, D. Baranchuk, G. Rätsch, S. Mandt, GP-VAE: Deep Probabilistic Multivariate Time Series Imputation (2019).
[37] S. I. Khan, A. S. M. L. Hoque, SICE: an improved missing data imputation technique, Journal of Big Data 7 (2020) 37. URL: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00313-w. doi:10.1186/s40537-020-00313-w.
[38] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, R. B. Altman, Missing value estimation methods for DNA microarrays, Bioinformatics 17 (2001) 520–525. URL: https://academic.oup.com/bioinformatics/article/17/6/520/272365. doi:10.1093/bioinformatics/17.6.520.
[39] F. Honghai, C. Guoshun, Y. Cheng, Y. Bingru, C. Yumei, A SVM Regression Based Approach to Filling in Missing Values, in: D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, R. Khosla, R. J. Howlett, L. C. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems, volume 3683, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005, pp. 581–587. URL: http://link.springer.com/10.1007/11553939_83. doi:10.1007/11553939_83, series Title: Lecture Notes in Computer Science.
[40] J. Yoon, J. Jordon, GAIN: Missing Data Imputation using Generative Adversarial Nets (2018).
[41] Z. Sun, H. Li, W. Wang, J. Liu, X. Liu, DTIN: Dual Transformer-based Imputation Nets for multivariate time series emitter missing data, Knowledge-Based Systems 284 (2024) 111270. URL: https://linkinghub.elsevier.com/retrieve/pii/S0950705123010195. doi:10.1016/j.knosys.2023.111270.
[42] N. B. Ipsen, P.-A. Mattei, J. Frellsen, HOW TO DEAL WITH MISSING DATA IN SUPERVISED DEEP LEARNING? (2022).

[43] S. C.-X. Li, B. Jiang, B. Marlin, MisGAN: Learning from Incomplete Data with Generative Adversarial Networks, 2019. URL: http://arxiv.org/abs/1902.09599. doi:10.48550/arXiv.1902.09599, arXiv:1902.09599 [cs].

[44] H. Jeong, H. Wang, F. P. Calmon, Fairness without Imputation: A Decision Tree Approach for Fair Prediction with Missing Values, 2022. URL: http://arxiv.org/abs/2109.10431, arXiv:2109.10431 [cs].

[45] Y. Xiao, R. Song, M. Chen, Direct and Unbiased Multiple Imputation Methods for Missing Values of Categorical Variables, Journal of Data Science 10 (2021) 465–481. URL: https://jds-online.org/doi/10.6339/JDS.201207_10%283%29.0007. doi:10.6339/JDS.201207_10(3).0007.

[46] T. Chen, C. Guestrin, XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794. URL: http://arxiv.org/abs/1603.02754. doi:10.1145/2939672.2939785, arXiv:1603.02754 [cs].

[47] C. M. Caruso, P. Soda, V. Guarrasi, Not Another Imputation Method: A Transformer-based Model for Missing Values in Tabular Datasets, 2024. URL: http://arxiv.org/abs/2407.11540, arXiv:2407.11540 [cs].

[48] S. M. Kia, N. M. Rad, D. v. Opstal, B. v. Schie, A. F. Marquand, J. Pluim, W. Cahn, H. G. Schnack, PROMISSING: Pruning Missing Values in Neural Networks, 2022. URL: http://arxiv.org/abs/2206.01640, arXiv:2206.01640 [cs].

[49] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A Survey on Bias and Fairness in Machine Learning, 2022. URL: http://arxiv.org/abs/1908.09635. doi:10.48550/arXiv.1908.09635, arXiv:1908.09635 [cs].

[50] P. Garg, J. Villasenor, V. Foggo, Fairness Metrics: A Comparative Analysis, in: 2020 IEEE International Conference on Big Data (Big Data), IEEE, Atlanta, GA, USA, 2020, pp. 3662–3666. URL: https://ieeexplore.ieee.org/document/9378025/. doi:10.1109/BigData50022.2020.9378025.

[51] M. J. Kusner, J. R. Loftus, C. Russell, R. Silva, Counterfactual Fairness, 2018. URL: http://arxiv.org/abs/1703.06856. doi:10.48550/arXiv.1703.06856, arXiv:1703.06856 [stat].

[52] L. Rosenblatt, R. T. Witter, Counterfactual Fairness Is Basically Demographic Parity, 2023. URL: http://arxiv.org/abs/2208.03843. doi:10.48550/arXiv.2208.03843, arXiv:2208.03843 [cs].

[53] T. L. Quy, A. Roy, V. Iosifidis, W. Zhang, E. Ntoutsi, A survey on datasets for fairness-aware machine learning, WIREs Data Mining and Knowledge Discovery 12 (2022) e1452. URL: http://arxiv.org/abs/2110.00530. doi:10.1002/widm.1452, arXiv:2110.00530 [cs].

[54] Q. Feng, M. Du, N. Zou, X. Hu, Fair Machine Learning in Healthcare: A Review, 2024. URL: http://arxiv.org/abs/2206.14397. doi:10.48550/arXiv.2206.14397, arXiv:2206.14397 [cs].

[55] R. Feng, F. P. Calmon, H. Wang, Adapting Fairness Interventions to Missing Values (2023).

[56] D. H. Lina, A. Silva, Better Fair than Sorry: Adversarial Missing Data Imputation for Fair GNNs, 2025. URL: http://arxiv.org/abs/2311.01591. doi:10.48550/arXiv.2311.01591, arXiv:2311.01591 [cs].

[57] D. Micci-Barreca, A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems, ACM SIGKDD explorations newsletter 3 (2001) 27–32.

[58] R. Shwartz-Ziv, A. Armon, Tabular Data: Deep Learning is Not All You Need, 2021. URL: http://arxiv.org/abs/2106.03253. doi:10.48550/arXiv.2106.03253, arXiv:2106.03253 [cs].

[59] F. Pargent, F. Pfisterer, J. Thomas, B. Bischl, Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features, Computational Statistics 37 (2022) 2671–2692. URL: https://link.springer.com/10.1007/s00180-022-01207-6. doi:10.1007/s00180-022-01207-6.

[60] M. Hardt, E. Price, N. Srebro, Equality of Opportunity in Supervised Learning, 2016. URL: http://arxiv.org/abs/1610.02413. doi:10.48550/arXiv.1610.02413, arXiv:1610.02413 [cs].

[61] B. Becker, R. Kohavi, Adult, UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

[62] J. Chai, X. Wang, Fairness with Adaptive Weights (2022).
[63] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, K. Q. Weinberger, On Fairness and Calibration (2017).
[64] J. Demšar, Statistical comparisons of classifiers over multiple data sets, Journal of Machine learning research 7 (2006) 1–30.