FAIM: Fair Imputation with Adversarial Training for Mitigating Bias in Missing Data

Rasta Tadayontahmasebi [1] **Haewon Jeong** [2] **Ramtin Pedarsani** [2]

Abstract

Imputation is a critical preprocessing step for handling missing data, yet conventional imputation methods can introduce or amplify unfairness across protected groups. We theoretically show that imputation fairness directly impacts downstream fairness in terms of accuracy parity at inference time, when the predictive model is trained solely on fully observed data. Motivated by this insight, we introduce a novel adversarial framework called **F**air **A**dversarial **IM**putation (**FAIM**) that can be integrated with any gradient-based imputation model. Our method incorporates a tunable fairness-accuracy trade-off parameter, allowing practitioners to balance imputation performance and imputation fairness. We empirically validate FAIM on two real-world datasets, showing significant improvements in group-wise imputation fairness. Furthermore, we assess the downstream fairness impact on a synthetic dataset derived from a real-world dataset, confirming that fairer imputations lead to fairer predictive outcomes at inference time.

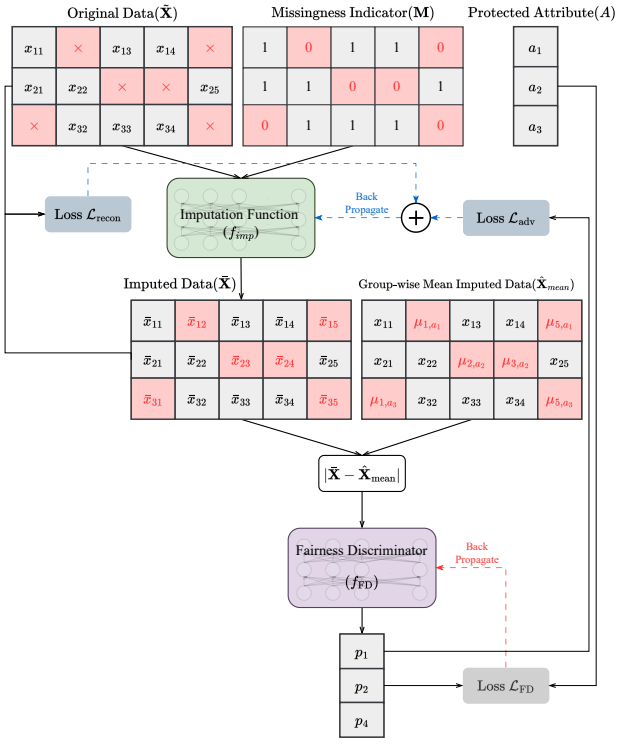

Figure 1: **FAIM architecture.** The fairness discriminator f_{FD} learns to predict the protected attribute from imputation errors. The imputation function f_{imp} is optimized with a combined reconstruction and adversarial loss to minimize group-identifiable signals and promote fairer imputations.

1 Introduction

Missing values are common in real-world datasets and can occur for many reasons. Some arise from random events such as sensor errors or data corruption (Song & Szafir, 2018; Wei & Link, 2019), while others are systematically linked to sociodemographic factors like age, race, sex, education, or income (Little & Rubin, 2019; Cheema, 2014; Jeanselme et al., 2022; Fernando et al., 2021). Additional causes include privacy concerns (Tourangeau & Yan, 2007), accessibility issues (Zhou et al., 2017), language barriers (Karras & Kornfeld, 2020), and question sensitivity or social desirability bias (Tourangeau & Yan, 2007).

Most machine learning (ML) pipelines are not inherently equipped to handle missing data, so missing values are typically addressed during preprocessing, either by dropping incomplete records or imputing missing entries with plausible values before training a model. However, when missingness patterns differ across demographic groups, these preprocessing decisions can significantly influence both the accuracy and fairness of downstream models (Newman, 2014; Martínez-Plumed et al., 2019; Feng et al., 2023). Despite the prevalence of missing data in real-world settings, the fair

[1]Department of Computer Science, University of California, Santa Barbara [2]Department of Electrical and Computer Engineering, University of California, Santa Barbara. Correspondence to: Rasta Tadayontahmasebi <rasta@ucsb.edu>, Haewon Jeong <haewon@ece.ucsb.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

ML literature often assumes complete datasets (Dwork et al., 2012; Calmon et al., 2017; Zafar et al., 2019), neglecting how disparities in imputation performance can propagate group-level bias. As a result, fair treatment of missing data remains a critical but underexplored issue in responsible ML.

In this paper, we shift the focus to a fundamental and overlooked question: *How can we impute missing values fairly, ensuring comparable performance across protected groups?* We address this question by introducing **FAIM**, a novel adversarial training framework that encourages group-invariant imputation quality. FAIM draws inspiration from the adversarial debiasing framework of Zhang et al. (2018), which promotes fairness by training a predictor alongside an adversary that attempts to recover the protected attribute. In our setup, the imputation function plays the role of the predictor, while the adversary learns to infer the protected attribute based on the imputation errors. This minimax interaction drives the imputation model to equalize its performance across groups. Our framework is broadly applicable: it supports any gradient-based imputation model that reconstructs the full data vector and handles both categorical and numerical features. Notably, it does not require access to the protected attribute at test time, making it practical for privacy-sensitive applications. A tunable hyperparameter controls the trade-off between imputation fairness and accuracy. We implement FAIM using a simple multi-layer perceptron (MLP) imputer and demonstrate its effectiveness in improving imputation fairness on two real-world datasets.

Our work is most closely related to FIGAN, introduced by Zhang & Long (2022), whose definition of imputation fairness aligns with ours and focuses on minimizing performance disparities across protected groups. FIGAN is a Generative Adversarial Network (GAN)-based (Goodfellow et al., 2014) imputation method augmented with a regularization term to encourage group-level parity. However, it assumes all features are numerical, which limits its applicability to datasets with categorical variables. It also requires access to the protected attribute at inference time, which may not be feasible. Most importantly, their work does not explore the relationship between imputation fairness and downstream model fairness. This is a key gap that our work addresses.

We consider a realistic deployment scenario in which a predictive model is trained on fully observed data, but inference time inputs may contain missing values that require imputation. We provide both theoretical and empirical evidence, using a synthetic dataset derived from the real-world Law School dataset (Wightman, 1998), that improving imputation fairness, that is, reducing group-level disparities in imputation quality, can directly improve downstream fairness, as measured by accuracy parity. In a simplified linear setting, we formally prove that imputation unfairness leads to accuracy disparities in the downstream model at inference time.

To summarize, our contributions in this work are fourfold: (1) we propose FAIM, the first adversarially trained framework for fair imputation; (2) we demonstrate a smooth and controllable trade-off between imputation fairness and imputation accuracy through a tunable parameter; (3) we theoretically establish, under a linear model, and empirically validate using both linear and non-linear classifiers, the connection between imputation fairness and downstream fairness at inference time under a complete-case training setup; and (4) we design FAIM to handle both numerical and categorical features without requiring the protected attribute at inference time.

2 Related Works

Based on Rubin's missingness framework (Rubin, 1976), missing data can be categorized into three cases: (1) *Missing Completely at Random* (MCAR) occurs when the missingness is entirely independent of both observed and unobserved variables. For example, data can be lost due to random sensor errors. (2) *Missing at Random* (MAR) refers to cases where the missingness depends only on observed variables. An example of MAR is when younger individuals are more likely to withhold their salary, given that age is an observed variable. (3) *Missing Not at Random* (MNAR) occurs when the missingness depends on unobserved variables, including the missing values themselves. For instance, individuals in the highest or lowest income brackets may choose not to disclose their income, and other observed data cannot fully explain this behavior.[1] In this work, we provide empirical results on all three missing mechanisms with different missing rates.

The simplest and commonly used method for handling missing values in ML pipelines is deletion, where rows containing any missing values are discarded. However, it is strongly advised in the missing value literature to use all the available data, as even discarding as little as 2-3% of data with missing values can significantly decrease the model performance (Newman, 2014). Furthermore, Martínez-Plumed et al. (2019) empirically show that keeping the rows with missing values often improves the fairness of the predictive model, as missingness is often not random and is systematically related to the protected attribute. While certain ML models, such as decision trees with surrogate splits (Breiman et al., 2017), or models that treat missingness as an informative attribute (Twala et al., 2008; Jeong et al., 2022), can accommodate missing values directly, the majority of models require complete data. As a result, im-

[1]The formal definition of MCAR, MAR, and MNAR can be found in the Supplementary Materials.

puting missing values by replacing them with reasonable approximations has become a critical preprocessing step in most ML workflows. Notably, the UCI Machine Learning Repository (Asuncion et al., 2007), one of the most widely used sources of benchmark datasets, includes numerous datasets in which missing values have already been imputed prior to public release.

Although retaining incomplete rows through imputation can help improve fairness (Martínez-Plumed et al., 2019), recent theoretical work suggests that the widely used "impute-then-classify" paradigm can, in fact, degrade group fairness (Feng et al., 2023). Supporting this, Nezami et al. (2024) empirically demonstrate that using imputed data to train predictive models for college student success increases accuracy but often exacerbates unfairness, particularly for Black and Hispanic students, compared to models trained on data where missing values have been removed. Motivated by these findings and by the fact that, in many real-world scenarios, classifiers are trained on fully observed data, this work focuses on the downstream effects of imputation fairness. Specifically, we study how the fairness of imputations made at inference time impacts the fairness of models trained on complete cases.

Recent studies have highlighted the social harms that can result from biased imputation methods (Caton et al., 2022; Fernando et al., 2021; Wang & Singh, 2021; Zhang & Long, 2021; Khan et al., 2024). For example, Zhang & Long (2021) define imputation fairness as the disparity in imputation accuracy between privileged and unprivileged groups and find that unfairness increases with greater missingness disparity, overall missingness, and class imbalance. We adopt a similar fairness definition in our work. The work most closely related to ours is that of Zhang & Long (2022), who introduce the concept of *imputation fairness risk* and provide theoretical bounds under the assumption of correctly specified imputation models. They further present a GAN-based (Goodfellow et al., 2014) fairness-aware imputation approach, incorporating a regularization term to improve group-level fairness. While this work offers a promising direction, it leaves open the question of how such fairness improvements in imputation translate to fairness in downstream predictive models. This concern is echoed by Shadbahr et al. (2023) high-performing classifiers can still arise from poorly imputed data, hypothesizing that the imputed datapoints can function as data augmentation or regularization. This raises a key question: *Does improving imputation fairness necessarily lead to fairer downstream predictions*? Addressing this from an empirical standpoint, Khan et al. (2024) introduce a comprehensive evaluation suite for responsible imputation. Their framework assesses methods based on imputation quality, fairness, and the predictive performance, fairness, and stability of models trained and tested on the imputed data.Their results suggest no consistent correlation between imputation fairness and downstream model fairness.

In contrast to these prior works, our study focuses specifically on inference-time fairness for models trained on complete-case data, and we provide a theoretical proof of a direct link between imputation fairness and downstream fairness under this setting. Moreover, unlike the approach of Zhang & Long (2022), our framework does not require access to the protected attribute at inference time, making it more practical in privacy-sensitive applications.

Our work is inspired by the *adversarial debiasing* framework proposed by Zhang et al. (2018), which mitigates bias by jointly training a predictor and an adversary in a minimax setup. In their method, the predictor is trained to make accurate predictions, while the adversary attempts to infer the protected attribute from the predictor's output layer. The predictor is penalized when the adversary succeeds, thus encouraging representations that are predictive of the target label but invariant to the sensitive attribute. This adversarial objective can be adapted to promote various group fairness metrics, including *demographic parity*, *equalized odds*, and *equal opportunity*[2], depending on how the adversary is defined. Our method builds on this idea in the context of imputation: we treat the imputer as the predictor and use an adversary that predicts the protected attribute given the error made by the imputer on the observed datapoints, thereby encouraging equal performance over different protected groups.

3 Problem Definition and Theoretical Results

In this section, we formally define the imputation process and present a theoretical analysis of how imputation fairness influences downstream fairness in a linear classifier trained on complete-case data.

3.1 Preliminaries

Missing Value Imputation. Let $\mathbf{X} = (X_1, \ldots, X_d)$ be a fully observed d-dimensional random variable taking place in the space of $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_d$. $\mathbf{X}$ is associated with a sensitive attribute A and a target label Y. Without loss of generality, we assume that the sensitive attribute and the target label are binary, i.e., $A, Y \in \{0, 1\}$.

We define a new space as $\tilde{\mathcal{X}} = \tilde{\mathcal{X}}_1 \times \cdots \times \tilde{\mathcal{X}}_d$, where each component space is given by $\tilde{\mathcal{X}}_j = \mathcal{X}_j \cup \{\texttt{na}\}$ for $j \in \{1, \ldots, d\}$. Here, na represents a missing entry. Let $\tilde{\mathbf{X}} = (\tilde{X}_1, \ldots, \tilde{X}_d)$ denote the partly observed random variable taking values in $\tilde{\mathcal{X}}$. Suppose that the random variable $\mathbf{M} = (M_1, \ldots, M_d) \in \{0, 1\}^d$ is a binary indicator vector, where

[2]Although demographic parity, equalized odds, and equal opportunity fairness measures are not directly used in this paper, their formal definitions can be found in the Supplementary Materials.

$M_i = 1$ indicates that X_i is observed and $M_i = 0$ indicates that X_i is missing (i.e., na). Finally, we define the missing dataset $\mathcal{D}$ as $\mathcal{D} = \{(\tilde{\mathbf{X}}^{(i)}, \mathbf{M}^{(i)}, A^{(i)}, Y^{(i)})\}_{i=1}^n$.

The imputation function $f_{\text{imp}} : \tilde{\mathcal{X}} \times \{0,1\}^d \to \mathcal{X}$ takes $\tilde{\mathbf{X}}$ and a corresponding binary mask $\mathbf{M}$ as inputs and outputs the fully imputed vector $\bar{\mathbf{X}}$. Now, $\bar{\mathbf{X}}$, $\hat{\mathbf{X}} \in \mathcal{X}$, are defined as follows:

$$\bar{\mathbf{X}} = f_{\text{imp}}(\tilde{\mathbf{X}}, \mathbf{M}) \tag{1}$$

$$\hat{\mathbf{X}} = \mathbf{M} \odot \tilde{\mathbf{X}} + (1 - \mathbf{M}) \odot \bar{\mathbf{X}}, \tag{2}$$

where $\odot$ denotes element-wise multiplication. The random variable $\bar{\mathbf{X}}$ denotes the fully imputed data vector, where all entries, including the observed ones, are reconstructed by the imputation function. This assumption is justified, as many gradient-based imputers reconstruct the full data vector(Yoon et al., 2018; Gondara & Wang, 2018; Mattei & Frellsen, 2019; Kotelnikov et al., 2023). $\hat{\mathbf{X}}$ represents the completed data vector where only the missing entries are replaced with their corresponding values in $\bar{\mathbf{X}}$, and the observed entries remain unchanged. In the remainder of this paper, lowercase letters are used to denote realizations of random variables.

Imputation Quality To evaluate the performance of the imputation method, we use the well-established metrics *Root Mean Squared Error* (RMSE) for numerical features and *Accuracy Error* (AR) for categorical features. Additionally, following the approach of (Miao et al., 2022), we use *Average Root Mean Squared Error* (ARMSE), which combines these two metrics. Lower ARMSE values indicate better imputation quality. We define these metrics formally as follows:

$$RMSE(\mathbf{x}_{\cdot j}, \hat{\mathbf{x}}_{\cdot j}) = \sqrt{\frac{\sum_{i=1}^n (1 - m_j^{(i)}) \cdot \left(x_j^{(i)} - \hat{x}_j^{(i)}\right)^2}{\sum_{i=1}^n (1 - m_j^{(i)})}} \tag{3}$$

$$AR(\mathbf{x}_{\cdot j}, \hat{\mathbf{x}}_{\cdot j}) = \frac{\sum_{i=1}^n (1 - m_j^{(i)}) \cdot \mathbb{I}\left[x_j^{(i)} \neq \hat{x}_j^{(i)}\right]}{\sum_{i=1}^n (1 - m_j^{(i)})} \tag{4}$$

$$ARMSE(\mathbf{X}, \hat{\mathbf{X}}) = \frac{1}{d}\left(\sum_{j \in \mathcal{F}_c} AR(\mathbf{x}_{\cdot j}, \hat{\mathbf{x}}_{\cdot j}) + \sum_{j \in \mathcal{F}_n} RMSE(\mathbf{x}_{\cdot j}, \hat{\mathbf{x}}_{\cdot j})\right) \tag{5}$$

where $\mathcal{F}_c$ and $\mathcal{F}_n$ denote the categorical and numerical feature sets, $\mathbf{x}_{\cdot j} = (x_j^{(1)}, \ldots, x_j^{(n)})$ is the vector of values for feature j, and $\mathbb{I}[\cdot]$ is the indicator function. Smaller values of all the metrics introduced suggest better imputation performance.

Imputation Fairness To assess the fairness of an imputation method, we introduce *ARMSE parity* metric formally defined as follows:

$$\textit{ARMSE parity} = \Big| ARMSE_{A=0}(\mathbf{X}, \hat{\mathbf{X}}) - ARMSE_{A=1}(\mathbf{X}, \hat{\mathbf{X}}) \Big| \tag{6}$$

This metric intuitively captures the performance disparity of an imputation model across protected groups, where a perfectly fair imputation method would yield an ARMSE parity of 0.

3.2 Theoretical Results

In the theorem below, we prove that when the downstream classifier relies on an approximately linear scoring function (e.g., logistic regression), any accuracy disparity in the imputation step propagates directly into an accuracy disparity in the downstream task.

Theorem 3.1. *Let $h : \mathbb{R}^d \to \mathbb{R}$ be a downstream classifier's score function. Assume that h satisfies the following:*

$$\mathbb{E}[\|h(X) - Y\|] \sim o(1),$$

$$L \cdot \|\hat{\mathbf{x}} - \mathbf{x}\| - \epsilon \leq \|h(\hat{\mathbf{x}}) - h(\mathbf{x})\| \leq L \cdot \|\hat{\mathbf{x}} - \mathbf{x}\| + \epsilon.$$

Let two imputation algorithms f_1 and f_2 have different ARMSE parities, i.e., without loss of generality,

$$\mathbb{E}[||X - f_1(\tilde{X})||] = \alpha \cdot \mathbb{E}[||X - f_2(\tilde{X})||], \quad (\alpha > 1).$$

Then, the downstream accuracy parity follows:

$$\begin{aligned} &|\mathbb{E}[\|h \circ f_1(\tilde{X}) - Y\||S = 0] - \mathbb{E}[\|h \circ f_1(\tilde{X}) - Y\||S = 1]| \\ &\quad \geq \alpha \cdot |\mathbb{E}[\|h \circ f_2(\tilde{X}) - Y\||S = 0] \\ &\qquad - \mathbb{E}[\|h \circ f_2(\tilde{X}) - Y\||S = 1]| - o(1). \end{aligned}$$

The proof of this theorem is provided in Appendix D

4 FAIM Framework

We propose **FAIM** (Fair Adversarial IMputation), a novel framework that leverages adversarial training to reduce disparities in imputation performance across protected groups. FAIM incorporates a fairness discriminator trained to predict the protected attribute from imputation errors, encouraging group-invariant imputation quality through an adversarial objective. This approach is inspired by the *Adversarial Debiasing* technique introduced by Zhang et al. (2018), in which a classifier is trained alongside an adversary that attempts to infer the sensitive attribute from the model's predictions. By jointly optimizing the predictor and adversary in a minimax fashion, their method encourages fairness under metrics such as *statistical parity* and *equalized odds*. FAIM extends this idea to the imputation setting, promoting fair treatment during data preprocessing.

Figure 1, illustrates the FAIM framework, which comprises of two neural networks: the imputation function f_{imp} and the fairness discriminator f_{FD}. f_{imp}, which can be any gradient-based imputation model, reconstructs the missing data matrix $\tilde{\mathbf{X}}$ as described in Section 3.1.

Algorithm 1 `Pseudo-code of FAIM`

Input: $\mathcal{D} = \{(\tilde{\mathbf{x}}^{(i)}, \mathbf{m}^{(i)}, a^{(i)})\}$ consisting of incomplete data vectors $\tilde{\mathbf{x}}^{(i)}$, missingness masks $\mathbf{m}^{(i)}$, and protected attributes $a^{(i)}$
Initialize: Parameters of f_{imp} and f_{FD}
for each training step **do**
(1) Fairness discriminator optimization
Draw k_{FD} samples from the dataset $\{(\tilde{\mathbf{x}}^{(j)}, \mathbf{m}^{(j)}, a^{(j)})\}_{j=1}^{k_{\text{FD}}}$
for $j = 1, \ldots, k_{\text{FD}}$ **do**
$\bar{\mathbf{x}}^{(j)} \leftarrow f_{\text{imp}}(\tilde{\mathbf{x}}^{(j)}, \mathbf{m}^{(j)})$
$\hat{\mathbf{x}}^{(j)}_{\text{mean}} \leftarrow \texttt{GroupwiseMeanImpute}(\tilde{\mathbf{x}}^{(j)}, \mathbf{m}^{(j)})$
$\mathbf{x}^{(j)}_{\text{err}} \leftarrow |\bar{\mathbf{x}}^{(j)} - \hat{\mathbf{x}}^{(j)}_{\text{mean}}|$
end for
Update f_{FD} using Stochastic Gradient Descent (SGD)

$$\nabla_{\text{FD}} - \sum_{j=1}^{k_{\text{FD}}} \mathcal{L}_{\text{FD}}(a^{(j)}, f_{\text{FD}}(\mathbf{x}^{(j)}_{\text{err}}))$$

(2) Imputation function optimization
Draw k_{imp} samples from the dataset $\{(\tilde{\mathbf{x}}^{(j)}, \mathbf{m}^{(j)}, \mathbf{a}^{(j)})\}_{j=1}^{k_{\text{imp}}}$
for $j = 1, \ldots, k_{\text{imp}}$ **do**
$\bar{\mathbf{x}}^{(j)} \leftarrow f_{\text{imp}}(\tilde{\mathbf{x}}^{(j)}, \mathbf{m}^{(j)})$
end for
Update f_{imp} using SGD

$$\nabla_{\text{imp}} \sum_{j=1}^{k_{\text{imp}}} \mathcal{L}_{\text{recon}}(\tilde{\mathbf{x}}^{(j)}, \bar{\mathbf{x}}^{(j)}) + \gamma \mathcal{L}_{\text{adv}}(f_{\text{FD}}(\mathbf{x}^{(j)}_{\text{err}}))$$

end for

4.1 Fairness Discriminator

Similar to the *Adversarial Debiasing* framework, we employ a fairness discriminator f_{FD} as an *adversary*. However, unlike typical classifiers where the predictor's output is a logits vector, our predictor outputs a fully imputed data matrix. Formally, the fairness discriminator is a function $f_{\text{FD}} : \mathbb{R}^d \to [0, 1]$. The input to the f_{FD} network is then calculated as the absolute error of the imputation function, formally denoted as

$$\mathbf{X}_{\text{err}} = |\hat{\mathbf{X}}_{\text{mean}} - \bar{\mathbf{X}}|, \tag{7}$$

where $\hat{\mathbf{X}}_{\text{mean}}$ represents the group-wise mean imputed data.

Why do we use $\hat{\mathbf{X}}_{mean}$ in (7)? When constructing the input to the fairness discriminator f_{FD}, a key challenge arises: missing values are not present in the input, and their patterns are often correlated with the protected attribute. If we replace these missing positions in the error vector with constant values (e.g., zeros), it can lead the fairness discriminator to exploit these disparities in the missingness patterns rather than the actual imputation errors, undermining its purpose. To mitigate this issue, we use group-wise mean imputation as a simple yet effective strategy to fill missing entries. This approach reduces the risk of leaking protected attribute information through missingness structure while providing a stable reference for computing meaningful imputation errors. Furthermore, using absolute error as input guides the model to equalize reconstruction quality across groups. Since absolute error correlates with RMSE and accuracy, reducing its disparity promotes imputation fairness measured by ARMSE parity.

The fairness discriminator is then trained on these error vectors to predict the protected attribute:

$$\hat{A} = f_{\text{FD}}(\mathbf{X}_{\text{err}}) \tag{8}$$

$$\mathcal{L}_{\text{FD}}(A, \hat{A}) = \mathbb{E}_{\mathbf{X}_{\text{err}}|A=1}[\log(\hat{A})] + \mathbb{E}_{\mathbf{X}_{\text{err}}|A=0}[\log(1 - \hat{A})] \tag{9}$$

Intuitively, the fairness discriminator is trained to maximize the probability of correctly predicting A based on the absolute error made by the imputation function.

4.2 Fair Imputer Objective

The imputation function f_{imp} is designed to predict missing values both accurately and fairly. To achieve this, it is trained using a combined loss consisting of a standard reconstruction loss, denoted by $\mathcal{L}_{\text{recon}}$, and an adversarial loss, denoted by $\mathcal{L}_{\text{adv}}$. The reconstruction loss $\mathcal{L}_{\text{recon}}$ ensures the accuracy of predicting missing values, typically using cross-entropy for categorical features and mean squared error ($l2$-loss) for numerical features. Details on the specific reconstruction loss as well as the imputation function used in our experiments are provided in Section 5. The adversarial loss $\mathcal{L}_{\text{adv}}$ is defined as:

$$\mathcal{L}_{\text{adv}}(\hat{A}) = (\hat{A} - 0.5)^2, \tag{10}$$

with $\hat{A}$ denoting the output of the fairness discriminator and taking values in $[0, 1]$. This $l2$-loss penalizes confident predictions by the fairness discriminator. It encourages $\hat{A}$ to stay close to 0.5, making the discriminator's prediction ambiguous for all groups. As a result, the imputation function learns to produce group-invariant errors, leading to improved fairness. While alternative loss functions can achieve a similar effect, we find the formulation in (10) to be the most stable and effective in practice.

Finally, the overall loss used to train the imputer is given by

$$\mathcal{L}_{\text{imp}} = \mathcal{L}_{\text{recon}}(\tilde{\mathbf{X}}, \bar{\mathbf{X}}) + \gamma \mathcal{L}_{\text{adv}}(\hat{A}), \tag{11}$$

where γ is a tunable hyperparameter that controls the trade-off between imputation accuracy and fairness.

The training procedure for the FAIM framework, outlined in Algorithm 1, follows the adversarial training paradigm introduced by (Goodfellow et al., 2014), alternating between two

fully connected neural networks: the imputation model f_{imp} and the fairness discriminator f_{FD}. At each training step, f_{FD} is first updated to predict the protected attribute from the imputation error vector, computed as the absolute difference between the output of f_{imp} and a group-wise mean-imputed baseline. Then, keeping f_{FD} fixed, we update f_{imp} using a combined loss that balances reconstruction accuracy and adversarial fairness, with the trade-off governed by the tunable hyperparameter γ.

5 Experiments and Evaluations

In this section, we evaluate the imputation and downstream fairness performance of our proposed method and its fairness-enhanced variant in comparison to several widely used baselines across multiple datasets.

MLP Imputer. In our experiments, both the f_{imp} and f_{FD} networks are implemented as MLPs with three hidden layers, each of dimension 16, using ReLU as the activation function. The weights are initialized using Xavier initialization (Glorot & Bengio, 2010). Since neural networks cannot inherently process missing values, we replace the missing entries in $\tilde{\mathbf{X}}$ with randomly generated noise values that are independent of all other features before feeding the data into the network. Formally, the input to the imputation function is defined as:

$$\mathbf{Z} \in \mathbb{R}^d, Z_i \sim \mathcal{U}(0, 0.01) \text{ independently for } i = 1, \ldots, d$$
$$\tilde{\mathbf{X}}_{\text{noise}} = \mathbf{M} \odot \tilde{\mathbf{X}} + (1 - \mathbf{M}) \odot \mathbf{Z} \tag{12}$$

where $\mathbf{M}$ is a binary mask indicating observed entries, and the noise vector $\mathbf{Z}$ is sampled independently for each dimension. The noise injection strategy follows the implementation of GAIN (Yoon et al., 2018). The imputation function then receives both the noise-injected input and the mask: $f_{\text{imp}}(\tilde{\mathbf{x}}_{\text{noise}}, \mathbf{m})$.

To enable the imputation function to handle both numerical and categorical features, we define the reconstruction loss $\mathcal{L}_{\text{recon}}$ as follows:

$$\mathcal{L}_{\text{recon}}(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{d} m_i L_{\text{recon}}(x_i, x_i'), \tag{13}$$

where

$$L_{\text{recon}}(x_i, x_i') = \begin{cases} \alpha \times (x_i' - x_i)^2, & \text{if } x_i \text{ is continuous,} \\ \beta \times (-x_i \log(x_i')), & \text{if } x_i \text{ is binary.} \end{cases}$$

Here, α and β are tunable hyperparameters that control the relative importance of imputing continuous versus binary (one-hot encoded categorical) features. The reconstruction loss is only computed on the observed points.

Finally, both networks are alternatively updated using the Adam optimizer and a fixed learning rate.

Baselines. To ensure a comprehensive evaluation, we compare FAIM against six widely used imputation baselines spanning diverse methodological categories. These baselines include: (1) statistical methods: `mean` and `group-wise mean`; (2) machine learning-based methods: `MICE`(Van Buuren & Groothuis-Oudshoorn, 2011), `MissForest`(Stekhoven & Bühlmann, 2012), and `KNN Imputer`(Troyanskaya et al., 2001); (3) matrix completion: `SoftImpute`(Mazumder et al., 2010); and (4) generative deep learning: `GAIN` (Yoon et al., 2018).[3] Implementation details of our baselines as well as hyperparameter tuning for our method are provided in Appendix C.

Datasets. We assess imputation fairness on two real-world datasets: the UCI Adult dataset (Dua & Graff, 2019) and the ACSIncome (Folk Income) dataset (Ding et al., 2021). To evaluate the effect of imputation fairness on downstream inference, we further experiment with synthetic data based on the Law School dataset (Wightman, 1998). Additional dataset details are provided in Appendix B, and the details of the synthetic data generation are provided in Section 5.2.

Data Missingness Patterns. In our experiments, we start with complete datasets and inject missingness to evaluate imputation accuracy. Following Khan et al. (2024), we introduce socially meaningful missing patterns by selecting the 3–4 features with the highest Kullback–Leibler Divergence (KLD) between protected groups. This allows us to evaluate FAIM under conditions where group-wise feature distributions differ most.

We evaluate three overall missingness rates: 20%, 30%, and 40%, applied consistently across all missingness mechanisms. For MCAR, missing values are injected uniformly at random into *the selected features*. For MAR, we design missing patterns such that the missing probability depends *only on the sensitive attribute*, simulating systemic bias by assigning higher missingness rates to the disadvantaged group (i.e., females in Adult, non-White individuals in Folk Income, and our Law School-based synthetic dataset). For MNAR, we condition the missing probability *only on the feature values themselves*; for example, in the Adult dataset, individuals with a non-married marital status are more likely to withhold that information, resulting in a higher missingness probability for such values. Table 2 details the selected features and injection conditions for the Adult dataset. Further details for other datasets are provided in Appendix B.1.

5.1 Improved Imputation Fairness

Figure 2 shows imputation fairness improvements on the Adult dataset under MCAR with 30% missingness. The `MLPImputer` serves as our baseline, using a simple MLP

[3] We exclude FIGAN as a baseline due to the unavailability of code and the lack of overlap in datasets.

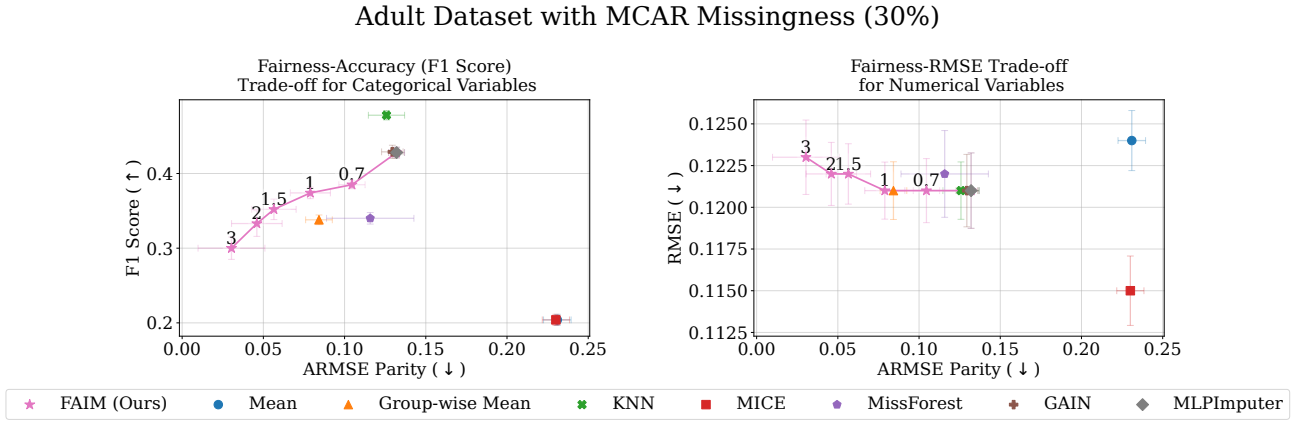

Figure 2: FAIM substantially reduces imputation unfairness, lowering ARMSE parity from approximately 0.13 to 0.03 with a modest impact on RMSE and F1 score. Results are shown on the **Adult** dataset under **MCAR** with **30%** missingness. FAIM points are annotated with γ values, which control the fairness-performance trade-off.

Table 1: Missingness configurations for the Adult dataset at 30% missing rate across different mechanisms. Rare occupations are defined as those with a population frequency of less than 10%.

Mechanism	**Missing Column**	**Conditional Column** (I)	$\mathbf{Pr(m = 0 \mid I}$ **is underprivileged)**	$\mathbf{Pr(m = 0 \mid I}$ **is privileged)**
MCAR	Relationship, Marital Status, Occupation, Hours per Week	N/A	0.3	0.3
MAR	Relationship, Marital Status, Occupation, Hours per Week	sex	0.2 (female)	0.1 (male)
MNAR	Relationship	Relationship	0.25 (not married)	0.05 (married)
	Marital Status	Marital Status	0.25 (not married)	0.05 (married)
	Occupation	Occupation	0.25 (rare occupations)	0.25 (common occupations)
	Hours per Week	Hours per Week	0.18 (< 40)	0.12 (≥ 40)

imputation function without any fairness regularization. FAIM starts with an ARMSE parity around 0.13, similar to that of `MissForest` and `KNN`, and reduces it to 0.03, significantly outperforming these baselines. This fairness gain comes with a modest increase in RMSE and a more noticeable drop in F1 score. For γ values up to 2, FAIM maintains imputation performance comparable to `MissForest` and `group-wise mean` imputation, while offering substantially better fairness. Furthermore, even at the highest γ value, FAIM's imputation performance remains within the range of the other baselines.

In Figure 3, we show results on the Folk Income dataset under MAR with 30% missingness. Compared to the Adult dataset, the baseline ARMSE parity values are generally lower. Our `MLPImputer` starts with an ARMSE parity of about 0.038, close to values observed for `KNN` and `GAIN`. FAIM improves this to 0.012 at the highest γ value, outperforming all baselines in terms of fairness. It is worth noting that this improvement comes at minimal cost to the F1 score for the categorical variables and an improvement in the RMSE for the numerical variables. At the highest γ level, our method outperforms all other baselines in terms of imputation fairness while having comparable imputation performance to the other baselines. `SoftImpute` results are omitted due to high variance that obscure trends.

5.2 Improved Downstream Fairness

To study the impact of imputation fairness on the downstream classification fairness when the model is trained on fully observed data, we construct a synthetic dataset derived from the Law School dataset. This dataset contains both numerical and ordinal categorical variables. We select two numerical features (LSAT and Decile1b) and one categorical feature (Family Income) that exhibit high group-wise distributional divergence (measured by KLD). The protected attribute, race, is retained as in the original dataset. To amplify group disparities, we modify the LSAT distribution to increase its divergence between racial groups. Specifically, we center the LSAT scores for the advantaged group by subtracting their group mean, thereby shifting their distribution and increasing the KLD from the disadvantaged group. We then generate a target score as a linear combination of the three selected features, assigning LSAT twice the weight of the others. Finally, we binarize the target to simulate group imbalance by setting separate thresholds for each group such that 80% of White individuals and 60% of non-White individuals receive a positive label.

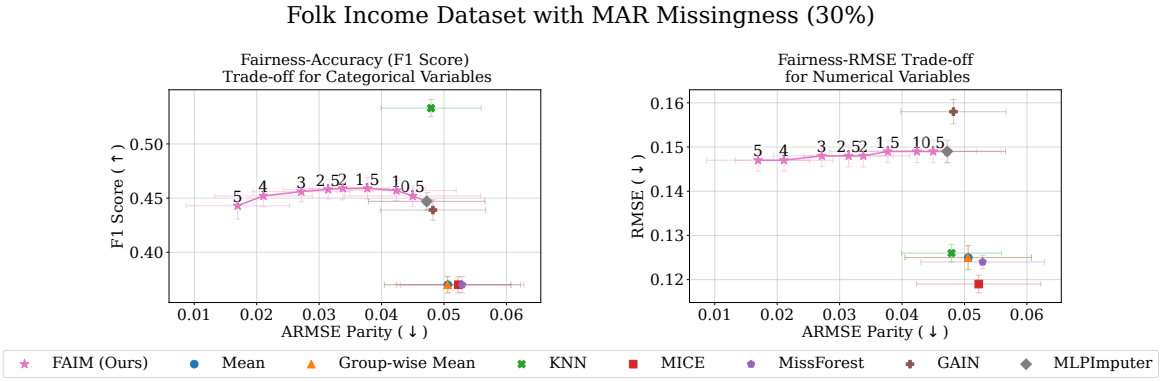

Figure 3: FAIM reduces ARMSE parity on the **Folk Income** dataset from approximately 0.038 to 0.017, with minimal impact on imputation F1 score and a slight improvement in RMSE. Results are shown under the **MAR** setting with **30%** missingness. FAIM points are annotated with γ values.

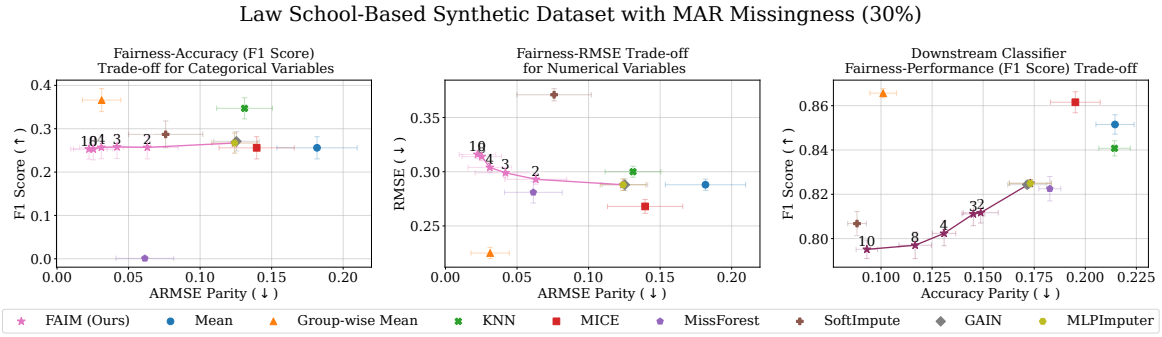

Figure 4: FAIM reduces ARMSE parity from around 0.12 to 0.02 with minimal impact on imputation RMSE and F1 score, and improves downstream accuracy parity from 0.17 to 0.09, with a slight drop in downstream F1 from 0.82 to 0.79. Results are shown for the **Law School-based synthetic** dataset under **MAR** with **30%** missingness. FAIM points are labeled with γ values.

We split the Law School-based synthetic dataset in half for training and testing. We train logistic regression and random forest models on the complete training set. During the test time, missing values are then introduced (details of missing patterns can be found in Table 4). While our theoretical analysis focuses on linear models, our empirical results show that downstream fairness also improves when using a non-linear model such as random forest, demonstrating the broader applicability of our approach. The corresponding results using a random forest classifier are included in Appendix F.

Figure 4 shows imputation and downstream fairness results on the Law School-based synthetic dataset under 30% MAR missingness. We vary γ and evaluate downstream fairness using accuracy parity from a logistic regression model. As γ increases, FAIM consistently reduces ARMSE parity, with minimal change in imputation F1 and a modest increase in RMSE. At higher γ values, FAIM achieves the best fairness while maintaining competitive imputation performance. The `group-wise mean` imputer also performs well across both metrics. In terms of downstream performance, the F1 score for our baseline, `MLPImputer`, is initially close to `GAIN` and `MissForest`, starting around 0.83. It gradually decreases to 0.79 as fairness improves, with accuracy parity improving from about 0.175 to 0.09.

It is important to note that selecting an appropriate γ value is crucial for balancing imputation fairness and downstream fairness with overall performance. Practitioners should carefully tune this parameter to ensure desirable outcomes across both fairness and accuracy.

6 Conclusion and Future Work

We present FAIM, a novel adversarial framework for fair imputation that promotes group-invariant performance by training a fairness discriminator on imputation errors. FAIM improves imputation fairness on real-world datasets, and we show both theoretically and empirically that fairer imputations lead to improved downstream fairness, specifically in terms of accuracy parity, when models are trained on fully observed data. The framework does not require access to the

protected attribute at inference and supports both numerical and categorical features.

In future work, we plan to replace the simple `MLPImputer` with more advanced models, such as diffusion-based imputers (Zheng & Charoenphakdee, 2022; Ouyang et al., 2023; Wen et al., 2024; Zhang et al., 2024), and extend our evaluation to a broader range of real-world datasets.

Impact Statement

This work addresses fairness in data imputation, a critical yet often overlooked aspect of machine learning pipelines. Since missing data is prevalent in real-world applications such as healthcare, education, and criminal justice, disparities in how imputation methods perform across demographic groups can propagate or amplify social biases in downstream decisions. By proposing a fairness-aware imputation framework that does not require access to protected attributes at inference time, our method aims to mitigate such harms while respecting privacy constraints. While our approach is a step toward more equitable machine learning systems, its deployment should still be accompanied by domain-specific evaluations and stakeholder input to ensure responsible use.

References

Asuncion, A., Newman, D., et al. Uci machine learning repository, 2007.

Breiman, L., Friedman, J., Olshen, R. A., and Stone, C. J. *Classification and regression trees*. Routledge, 2017.

Calmon, F., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., and Varshney, K. R. Optimized pre-processing for discrimination prevention. *Advances in neural information processing systems*, 30, 2017.

Caton, S., Malisetty, S., and Haas, C. Impact of imputation strategies on fairness in machine learning. *Journal of Artificial Intelligence Research*, 74:1011–1035, 2022.

Cheema, J. R. A review of missing data handling methods in education research. *Review of Educational Research*, 84(4):487–508, 2014.

Ding, F., Vaughan, J. W., Wallach, H., and Vaughan, J. W. Retiring adult: New datasets for fair machine learning. *arXiv preprint arXiv:2108.04884*, 2021. URL https://arxiv.org/abs/2108.04884.

Dua, D. and Graff, C. UCI machine learning repository, 2019. URL https://archive.ics.uci.edu/ml/datasets/adult. Accessed: 2025-05-10.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012.

Feng, R., Calmon, F., and Wang, H. Adapting fairness interventions to missing values. *Advances in Neural Information Processing Systems*, 36:59388–59409, 2023.

Fernando, M.-P., Cèsar, F., David, N., and José, H.-O. Missing the missing values: The ugly duckling of fairness in machine learning. *International Journal of Intelligent Systems*, 36(7):3217–3258, 2021.

Glorot, X. and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256. JMLR Workshop and Conference Proceedings, 2010.

Gondara, L. and Wang, K. Mida: Multiple imputation using denoising autoencoders. In *Advances in Knowledge Discovery and Data Mining: 22nd Pacific-Asia Conference, PAKDD 2018, Melbourne, VIC, Australia, June 3-6, 2018, Proceedings, Part III 22*, pp. 260–272. Springer, 2018.

Goodfellow, I. et al. Generative adversarial nets. *Advances in neural information processing systems*, 2014.

Jeanselme, V., De-Arteaga, M., Zhang, Z., Barrett, J., and Tom, B. Imputation strategies under clinical presence: Impact on algorithmic fairness. In *Machine Learning for Health*, pp. 12–34. PMLR, 2022.

Jeong, H., Wang, H., and Calmon, F. P. Fairness without imputation: A decision tree approach for fair prediction with missing values. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 9558–9566, 2022.

Karras, J. E. and Kornfeld, M. Survey design and linguistic equity. *Journal of Survey Statistics and Methodology*, 8 (1):73–91, 2020. doi: 10.1093/jssam/smz045.

Khan, F. A., Herasymuk, D., Protsiv, N., and Stoyanovich, J. Still more shades of null: An evaluation suite for responsible missing value imputation. *arXiv preprint arXiv:2409.07510*, 2024.

Kotelnikov, A., Baranchuk, D., Rubachev, I., and Babenko, A. Tabddpm: Modelling tabular data with diffusion models. In *International Conference on Machine Learning*, pp. 17564–17579. PMLR, 2023.

Little, R. J. and Rubin, D. B. *Statistical analysis with missing data*. John Wiley & Sons, 2019.

Martínez-Plumed, F., Ferri, C., Nieves, D., and Hernández-Orallo, J. Fairness and missing values. *arXiv preprint arXiv:1905.12728*, 2019.

Mattei, P.-A. and Frellsen, J. Miwae: Deep generative modelling and imputation of incomplete data sets. In *International conference on machine learning*, pp. 4413–4423. PMLR, 2019.

Mazumder, R., Hastie, T., and Tibshirani, R. Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research*, 11: 2287–2322, 2010.

Miao, X., Wu, Y., Chen, L., Gao, Y., and Yin, J. An experimental survey of missing data imputation algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 35(7):6630–6650, 2022.

Newman, D. A. Missing data: Five practical guidelines. *Organizational research methods*, 17(4):372–411, 2014.

Nezami, N., Haghighat, P., Gándara, D., and Anahideh, H. Assessing disparities in predictive modeling outcomes for college student success: The impact of imputation techniques on model performance and fairness. *Education Sciences*, 14(2):136, 2024.

Ouyang, Y., Xie, L., Li, C., and Cheng, G. Missdiff: Training diffusion models on tabular data with missing values. *arXiv preprint arXiv:2307.00467*, 2023.

Rubin, D. B. Inference and missing data. *Biometrika*, 63(3): 581–592, 1976.

Shadbahr, T., Roberts, M., Stanczuk, J., Gilbey, J., Teare, P., Dittmer, S., Thorpe, M., Torné, R. V., Sala, E., Lió, P., et al. The impact of imputation quality on machine learning classifiers for datasets with missing values. *Communications Medicine*, 3(1):139, 2023.

Song, H. and Szafir, D. A. Where's my data? evaluating visualizations with missing data. *IEEE transactions on visualization and computer graphics*, 25(1):914–924, 2018.

Stekhoven, D. J. and Bühlmann, P. Missforest—nonparametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

Tourangeau, R. and Yan, T. Sensitive questions in surveys. *Psychological Bulletin*, 133(5):859–883, 2007. doi: 10.1037/0033-2909.133.5.859.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001.

Twala, B. E., Jones, M., and Hand, D. J. Good methods for coping with missing data in decision trees. *Pattern Recognition Letters*, 29(7):950–956, 2008.

Van Buuren, S. and Groothuis-Oudshoorn, K. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45:1–67, 2011.

Wang, Y. and Singh, L. Analyzing the impact of missing values and selection bias on fairness. *International Journal of Data Science and Analytics*, 12(2):101–119, 2021.

Wei, Z. and Link, S. Embedded functional dependencies and data-completeness tailored database design. *Proceedings of the VLDB Endowment*, 12(11):1458–1470, 2019.

Wen, Y., Wang, Y., Yi, K., Ke, J., and Shen, Y. Diffimpute: Tabular data imputation with denoising diffusion probabilistic model. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2024.

Wightman, L. F. Lsac national longitudinal bar passage study. Technical report, Law School Admission Council, 1998.

Yoon, J., Jordon, J., and Schaar, M. Gain: Missing data imputation using generative adversarial nets. In *International conference on machine learning*, pp. 5689–5698. PMLR, 2018.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., and Gummadi, K. P. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20(75):1–42, 2019.

Zhang, B. H., Lemoine, B., and Mitchell, M. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340, 2018.

Zhang, H., Fang, L., and Yu, P. S. Unleashing the potential of diffusion models for incomplete data imputation. *arXiv preprint arXiv:2405.20690*, 2024.

Zhang, Y. and Long, Q. Assessing fairness in the presence of missing data. *Advances in neural information processing systems*, 34:16007–16019, 2021.

Zhang, Y. and Long, Q. Fairness-aware missing data imputation. In *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*, 2022.

Zheng, S. and Charoenphakdee, N. Diffusion models for missing value imputation in tabular data. In *NeurIPS 2022 Workshop on Table Representation Learning*, 2022. URL https://arxiv.org/abs/2210.17128.

Zhou, J., Rau, P.-L. P., and Salvendy, G. Older adults' responses to font size, color contrast, and letter spacing on mobile devices. *Behaviour & Information Technology*, 36(2):119–129, 2017. doi: 10.1080/0144929X.2016.1212093.

A Formal Definitions of Missing Data Mechanisms and Group Fairness Metrics

Let $\mathbf{X} = (X_1, \ldots, X_d) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_d$ be a fully observed d-dimensional random variable representing the input features. Each data point is associated with a binary sensitive attribute $A \in \{0, 1\}$ and a binary target label $Y \in \{0, 1\}$.

Let $\tilde{\mathbf{X}} = (\tilde{X}_1, \ldots, \tilde{X}_d) \in \tilde{\mathcal{X}} = (\mathcal{X}_1 \cup \{\text{na}\}) \times \cdots \times (\mathcal{X}_d \cup \{\text{na}\})$ denote the partially observed version of $\mathbf{X}$, where `na` indicates a missing value. We define the missingness indicator vector $\mathbf{M} = (M_1, \ldots, M_d) \in \{0, 1\}^d$, where $M_j = 1$ if X_j is observed and $M_j = 0$ otherwise.

A.1 Missing Data Mechanisms

The missingness mechanism is characterized by the conditional distribution $\mathbb{P}(\mathbf{M} \mid \mathbf{X}, A, Y)$, and falls into one of the following categories:

MCAR: Data is said to be missing completely at random (MACR) if the missingness is independent of both the observed and unobserved data. Formally:

$$\mathbb{P}(\mathbf{M} \mid \mathbf{X}, A, Y) = \mathbb{P}(\mathbf{M}).$$

MAR: Data is said to be missing at random (MAR) if the missingness is only dependent on the observed data i.e., where $M_i = 1$. Let $\mathbf{X}_{\text{obs}}$ be the observed components of $\mathbf{X}$. MAR missingness is formally defined as follows:

$$\mathbb{P}(\mathbf{M} \mid \mathbf{X}, A, Y) = \mathbb{P}(\mathbf{M} \mid \mathbf{X}_{\text{obs}}, A, Y),$$

MNAR: Data is said to be missing not at random (MNAR) if the missingness depends on the unobserved components $\mathbf{X}_{\text{mis}}$ and cannot be explained by the observed data alone.

$$\mathbb{P}(\mathbf{M} \mid \mathbf{X}, A, Y) \text{ depends on } \mathbf{X}_{\text{mis}},$$

A.2 Group Fairness Metrics

We also define the following group fairness metrics, where $\hat{Y} \in \{0, 1\}$ denotes the predicted label derived from the (possibly imputed) features:

Demographic Parity (Statistical Parity). This criterion requires the model to produce positive predictions at equal rates across groups, regardless of the true label. It is formally satisfied when

$$\mathbb{P}(\hat{Y} = 1 \mid A = 0) = \mathbb{P}(\hat{Y} = 1 \mid A = 1).$$

Equalized Odds. This metric requires the model to have both equal true positive rates and equal false positive rates across groups. It is satisfied if

$$\mathbb{P}(\hat{Y} = 1 \mid A = 0, Y = y) = \mathbb{P}(\hat{Y} = 1 \mid A = 1, Y = y), \quad \text{for } y \in \{0, 1\}.$$

Equal Opportunity. A relaxation of equalized odds, this metric only requires equal true positive rates across groups. It holds when

$$\mathbb{P}(\hat{Y} = 1 \mid A = 0, Y = 1) = \mathbb{P}(\hat{Y} = 1 \mid A = 1, Y = 1).$$

Accuracy Parity. This metric requires that the overall prediction accuracy is equal across groups defined by the sensitive attribute. Formally, it is satisfied when

$$\mathbb{P}(\hat{Y} = Y \mid A = 0) = \mathbb{P}(\hat{Y} = Y \mid A = 1).$$

B Dataset Information

Adult. The Adult dataset (Dua & Graff, 2019), also known as the Census Income dataset, is a widely used benchmark for fairness research. It originates from the 1994 U.S. Census and contains demographic and employment-related information for 48,842 individuals. The binary classification task is to predict whether an individual's annual income exceeds \$50,000 based on 14 attributes, including age, education, occupation, and hours worked per week. In our experiments, we use the standard processed version of the dataset. The sensitive attribute is gender, with "female" designated as the disadvantaged group.

ACSIncome. The ACSIncome dataset is a subset of the Folktables benchmark (Ding et al., 2021), which is derived from U.S. Census data collected between 2014–2018 across all 50 states. Specifically, the ACSIncome task (referred to as "folk-income") is a binary classification problem aimed at predicting whether an individual's annual income exceeds $50,000, based on 10 features, 8 categorical and 2 numerical, including educational attainment, work hours per week, marital status, and occupation. For our experiments, we use data from the state of Georgia from the year 2018, subsampled to 40,000 instances. The sensitive attribute used for fairness evaluation is race, with "non-White" treated as the disadvantaged group.

Law School. The Law School dataset (Wightman, 1998) was collected by the Law School Admission Council (LSAC) through a survey conducted in 1991, covering 20,798 applicants from 163 U.S. law schools. Each record includes 11 features (5 categorical and 6 numerical), such as LSAT scores and undergraduate GPA. The prediction task involves determining whether a student will pass the bar exam. In our setup, we treat race as the sensitive attribute, with "non-White" designated as the disadvantaged group for fairness analysis. For generating our synthetic dataset we use LSAT, Decile1b, and family income as the features. The binary target variable is created using a linear weighting of these features and a threshold.

B.1 Missingness Configurations

Table 2: Missingness configuration for the Adult dataset at 20% and 40% missing rates across different mechanisms. Rare occupations are defined as those with a population frequency of less than 10%.

Mechanism	**Missing Column**	**Conditional Column (I)**	$\mathbf{Pr}(\mathbf{m}=\mathbf{0} \mid \mathbf{I}$ is underprivileged)	$\mathbf{Pr}(\mathbf{m}=\mathbf{0} \mid \mathbf{I}$ is privileged)
MCAR	Relationship, Marital status, Occupation, Hours per week	N/A	0.2 / 0.4	0.2 / 0.4
MAR	Relationship, Marital status, Occupation, Hours per week	sex	0.133 / 0.266 (female)	0.067 / 0.134 (male)
MNAR	Relationship	Relationship	0.168 / 0.336 (not married)	0.032 / 0.067 (married)
	Marital status	Marital status	0.168 / 0.336 (not married)	0.032 / 0.067 (married)
	Occupation	Occupation	0.133 / 0.266 (rare)	0.067 / 0.134 (common)
	Hours per week	Hours per week	0.12 / 0.24 (< 40)	0.08 / 0.16 (≥ 40)

Table 3: Missingness configuration for the Folk Income dataset at 20%, 30%, and 40% missing rates across different mechanisms.

Mechanism	**Missing Column**	**Conditional Column (I)**	$\mathbf{Pr}(\mathbf{m}=\mathbf{0} \mid \mathbf{I}$ is underprivileged)	$\mathbf{Pr}(\mathbf{m}=\mathbf{0} \mid \mathbf{I}$ is privileged)
MCAR	Hours worked per week, Relationship, Marital status	N/A	0.2 / 0.3 / 0.4	0.2 / 0.3 / 0.4
MAR	Hours worked per week, Relationship, Marital status	race	0.132 / 0.2 / 0.267 (non-white)	0.067 / 0.1 / 0.134 (white)
MNAR	Hours worked per week	Hours worked per week	0.168 / 0.25 / 0.334 (< 40)	0.032 / 0.05 / 0.066 (≥ 40)
	Relationship	Relationship	0.168 / 0.25 / 0.334 (not married)	0.032 / 0.05 / 0.066 (married)
	Marital status	Marital status	0.168 / 0.25 / 0.334 (not married)	0.032 / 0.05 / 0.066 (married)

Table 4: Missingness configuration for the synthetic Law School dataset at 20%, 30%, and 40% missing rates across different missing mechanisms. The LSAT and Decile1B values are normalized.

Mechanism	**Missing Column ($\mathcal{F}^m$)**	**Conditional Column (I)**	$\mathbf{Pr}(\mathcal{F}^{\mathbf{m}} \mid \mathbf{I}$ is dis)	$\mathbf{Pr}(\mathcal{F}^{\mathbf{m}} \mid \mathbf{I}$ is priv)
MCAR	LSAT, Decile1B, Family income	N/A	0.2 / 0.3 / 0.4	0.2 / 0.3 / 0.4
MAR	LSAT, Decile1B, Family income	race	0.132 / 0.2 / 0.267 (non-white)	0.067 / 0.1 / 0.134 (white)
MNAR	LSAT	LSAT	0.132 / 0.2 / 0.267 (≤ 0.8)	0.067 / 0.1 / 0.134 (> 0.8)
	Decile1B	Decile1B	0.132 / 0.2 / 0.267 (< 0.5)	0.067 / 0.1 / 0.134 (≥ 0.5)
	Family income	Family income	0.132 / 0.2 / 0.267 (< 4)	0.067 / 0.1 / 0.134 (≥ 4)

C Experimental Setup

Statistical and machine learning-based baselines are implemented using the `scikit-learn` library. For `MissForest`, we use `IterativeImputer` with a `Random Forest` regressor. These baselines use default hyperparameters in our experiments. For `MLPImputer` and `GAIN`, we fix the batch size to 128 and set the hyperparameters α and β to 5, as this consistently yielded the best results. We tune the learning rate for the imputation function and the fairness discriminator in the range $[5e^{-5}, 5e^{-4}]$ with a step size of $5e^{-5}$, and the number of training iterations in the range $[1000, 2000]$ with a step size of 500. All experiments are conducted on CPU, categorical features are one-hot encoded, and results are averaged over 10 random train-test splits.

D Theoretical Analysis of Imputation and Downstream Fairness

In this section we provide the proof of Theroem 3.1.

Proof. Let $E_{h\circ f} = \|h \circ f(\tilde{X}) - Y\|$. Then,

$$\begin{aligned}
\mathbb{E}[E_{h\circ f_1}|S=1] &= \mathbb{E}[\|h \circ f_1(\tilde{X}) - Y\||S=1] && (14)\\
&= \mathbb{E}[\|h \circ f_1(\tilde{X}) + h(X) - h(X) - Y\||S=1] && (15)\\
&\leq \mathbb{E}[\|h \circ f_1(\tilde{X}) - h(X)\||S=1] + \mathbb{E}[\|h(X) - Y\||S=1] && (16)\\
&\leq L \cdot \mathbb{E}[\|X - f_1(X)\||S=1] + \epsilon + \delta_1, && (17)
\end{aligned}$$

where $\delta_1 = \mathbb{E}[\|h(X) - Y\||S=1] \sim o(1)$. By using the reverse triangular inequality, we obtain:

$$\begin{aligned}
\mathbb{E}[E_{h\circ f_1}|S=1] &= \mathbb{E}[\|h \circ f_1(\tilde{X}) - Y\||S=1] && (18)\\
&= \mathbb{E}[\|h \circ f_1(\tilde{X}) + h(X) - h(X) - Y\||S=1] && (19)\\
&\geq \mathbb{E}[\|h \circ f_1(\tilde{X}) - h(X)\||S=1] - \mathbb{E}[\|h(X) - Y\||S=1] && (20)\\
&\geq L \cdot \mathbb{E}[\|X - f_1(X)\||S=1] - \epsilon - \delta_1. && (21)
\end{aligned}$$

By combining (17) and (21), we get:

$$L \cdot \mathbb{E}[\|X - f_1(X)\||S=1] - \epsilon - \delta_1 \leq \mathbb{E}[E_{h\circ f_1}|S=1] \leq L \cdot \mathbb{E}[\|X - f_1(X)\||S=1] + \epsilon + \delta_1.$$

Letting $\delta_2 = \mathbb{E}[E_{h\circ f_1}|S=0] \sim o(1)$, we can show similar inequalities:

$$L \cdot \mathbb{E}[\|X - f_1(X)\||S=0] - \epsilon - \delta_2 \leq \mathbb{E}[E_{h\circ f_1}|S=0] \leq L \cdot \mathbb{E}[\|X - f_1(X)\||S=0] + \epsilon + \delta_2.$$

Hence,

$$\begin{aligned}
|\mathbb{E}[E_{h\circ f_1}|S=0] - \mathbb{E}[E_{h\circ f_1}|S=1]| \geq & L \cdot |\mathbb{E}[\|X - f_1(X)\||S=0] - \mathbb{E}[\|X - f_1(X)\||S=1]| \\
& - 2\epsilon - \delta_1 - \delta_2. && (22)
\end{aligned}$$

By letting $\delta_3 = \mathbb{E}[E_{h\circ f_2}|S=1]$ and $\delta_4 = \mathbb{E}[E_{h\circ f_2}|S=0]$, we can show:

$$\begin{aligned}
|\mathbb{E}[E_{h\circ f_2}|S=0] - \mathbb{E}[E_{h\circ f_2}|S=1]| \leq & L \cdot |\mathbb{E}[\|X - f_2(X)\||S=0] - \mathbb{E}[\|X - f_2(X)\||S=1]| \\
& + 2\epsilon + \delta_3 + \delta_4 \\
= & L \cdot \frac{1}{\alpha} |\mathbb{E}[\|X - f_1(X)\||S=0] - \mathbb{E}[\|X - f_1(X)\||S=1]| \\
& + 2\epsilon + \delta_3 + \delta_4. && (23)
\end{aligned}$$

Using (22) and (23), we can derive:

$$\begin{aligned}
|\mathbb{E}[E_{h\circ f_1}|S=0] - \mathbb{E}[E_{h\circ f_1}|S=1]| &\geq \alpha \cdot |\mathbb{E}[E_{h\circ f_2}|S=0] - \mathbb{E}[E_{h\circ f_2}|S=1]| \\
&\quad - 2\epsilon - \delta_1 - \delta_2 - \alpha(2\epsilon + \delta_3 + \delta_4) \\
&= \alpha \cdot |\mathbb{E}[E_{h\circ f_2}|S=0] - \mathbb{E}[E_{h\circ f_2}|S=1]| - o(1). && (24)
\end{aligned}$$

□

E Imputation Fairness Experiment Results

The marker labels represent the γ hyperparameter values used to control the trade-off between imputation fairness and performance, with higher γ values leading to improved fairness (lower ARMSE parity), usually with some loss to the imputation performance.

E.1 MCAR Experiments

E.1.1 Adult Dataset Results

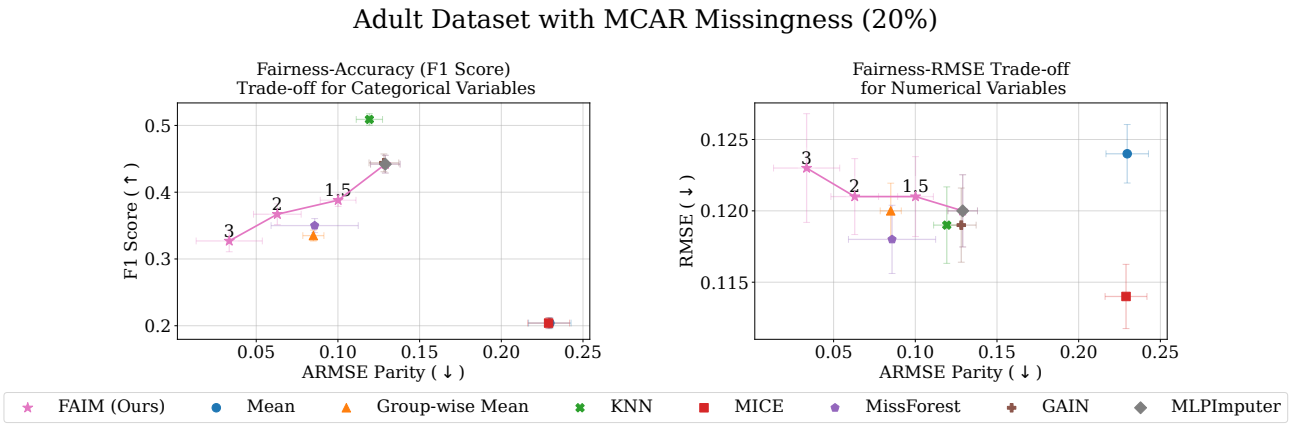

Figure 5: Imputation fairness improvement for the **Adult** dataset under **MCAR** at **20%** missingness.

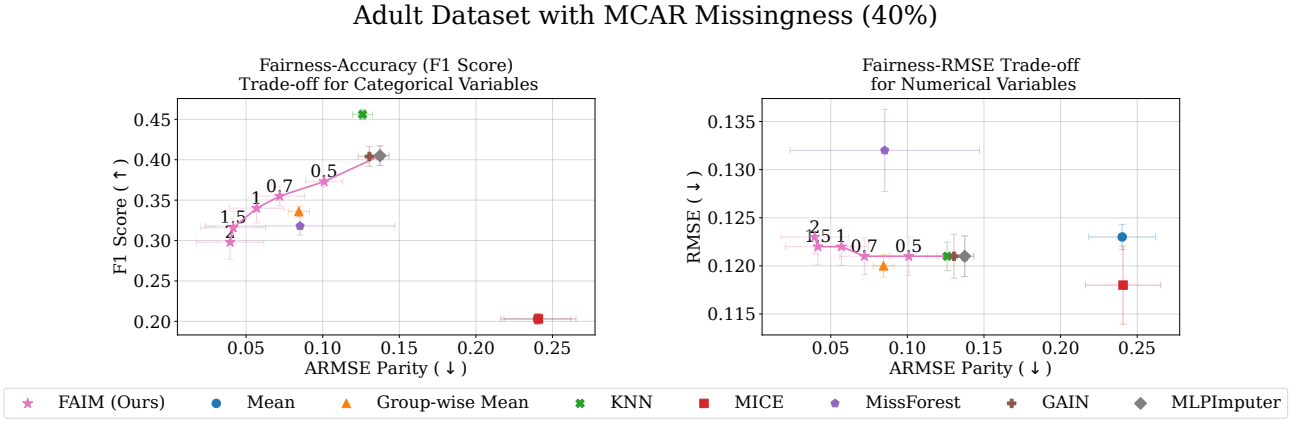

Figure 6: Imputation fairness improvement for the **Adult** dataset under **MCAR** at **40%** missingness.

E.1.2 Folk Income Dataset Results

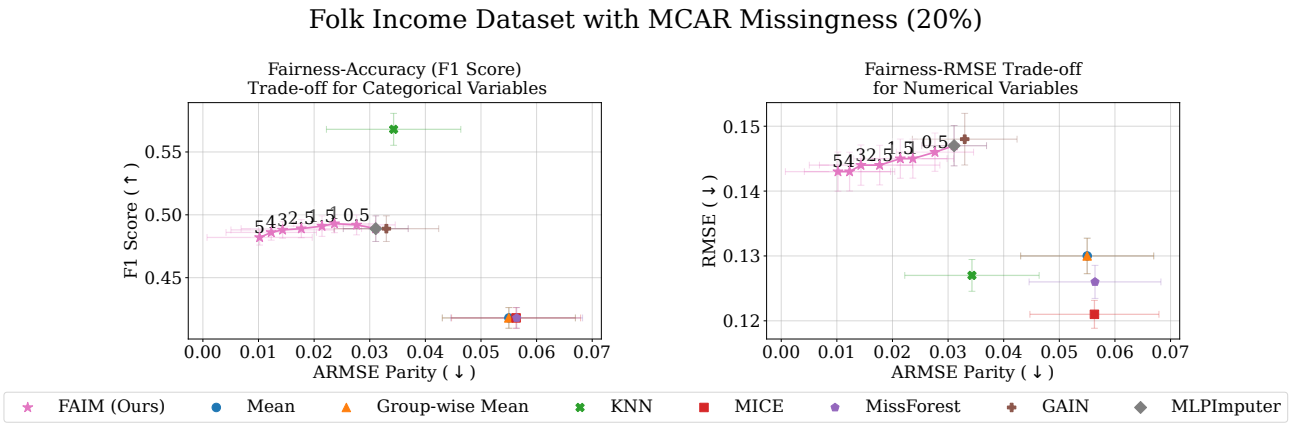

Figure 7: Imputation fairness improvement for the **Folk Income** dataset under **MCAR** at **20%** missingness.

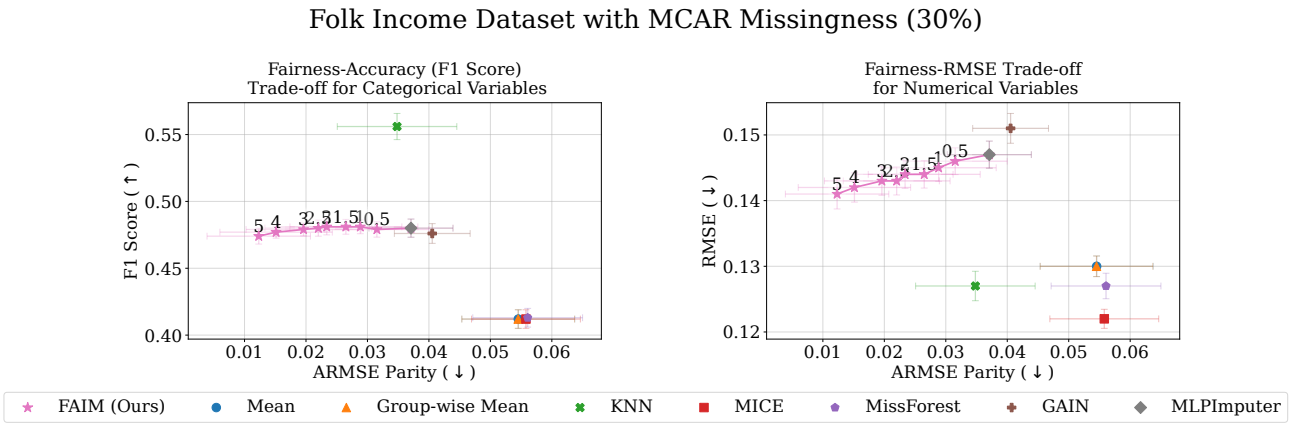

Figure 8: Imputation fairness improvement for the **Folk Income** dataset under **MCAR** at **30%** missingness.

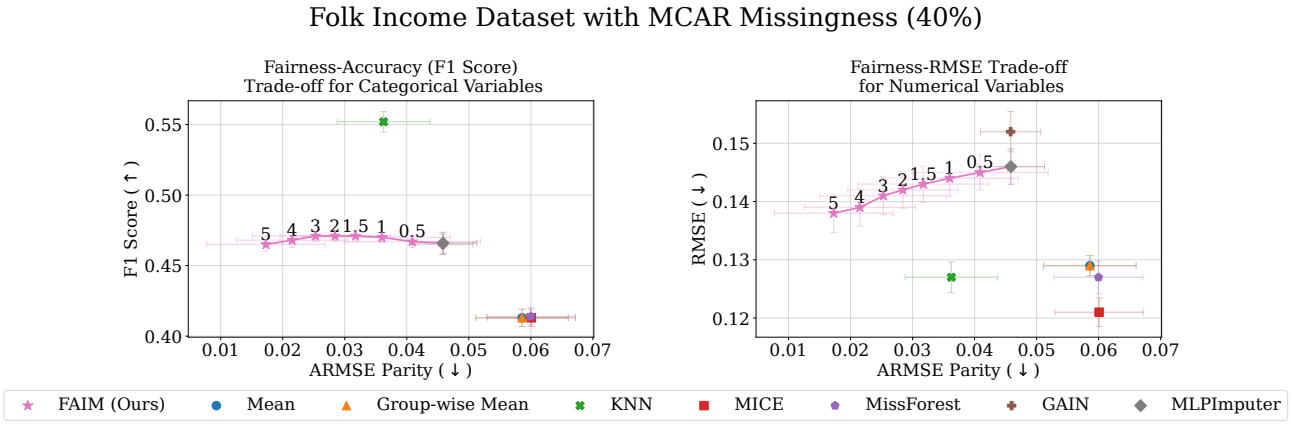

Figure 9: Imputation fairness improvement for the **Folk Income** dataset under **MCAR** at **30%** missingness.

E.2 MAR Experiments

E.2.1 Adult Dataset Results

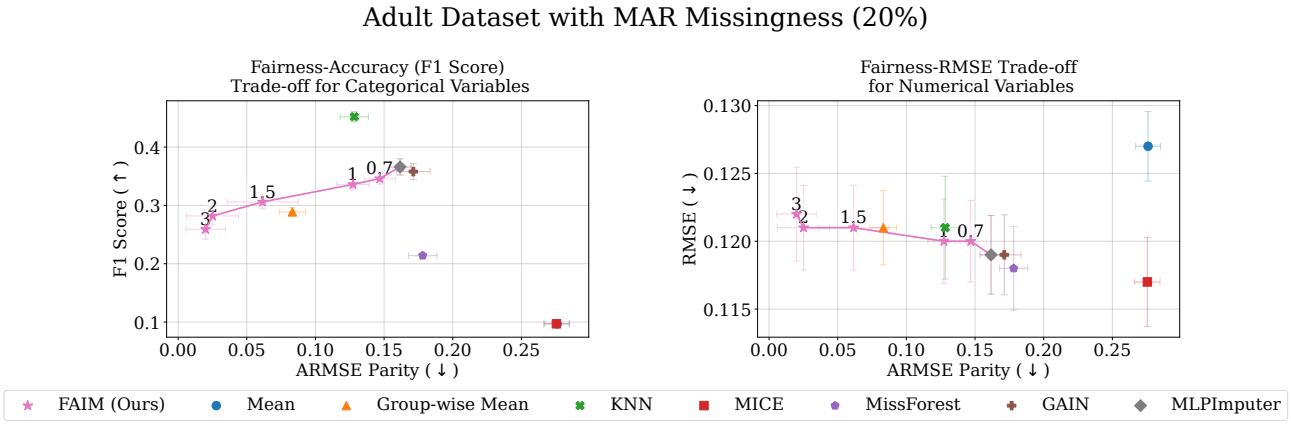

Figure 10: Imputation fairness improvement for the **Adult** dataset under **MAR** at **20%** missingness.

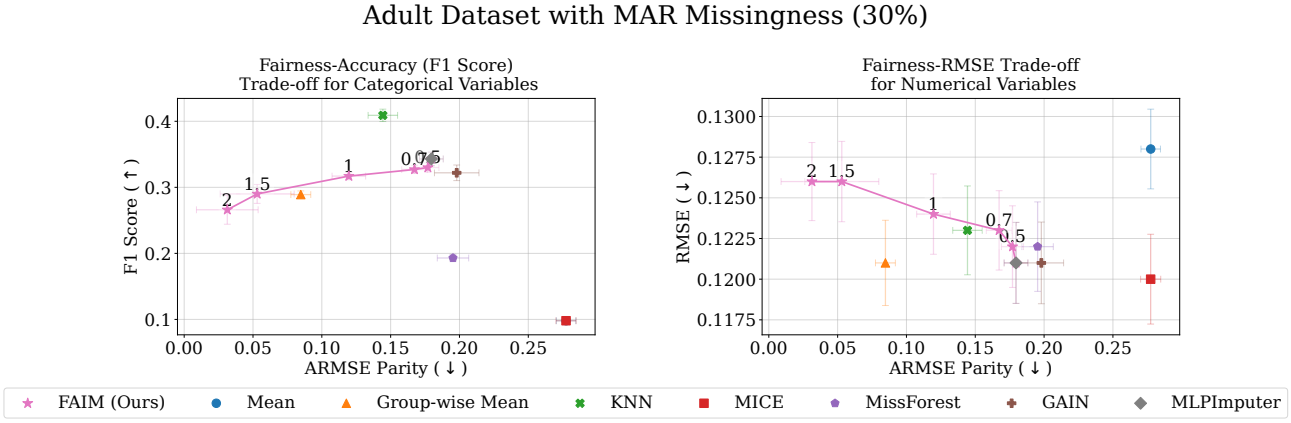

Figure 11: Imputation fairness improvement for the **Adult** dataset under **MAR** at **30%** missingness.

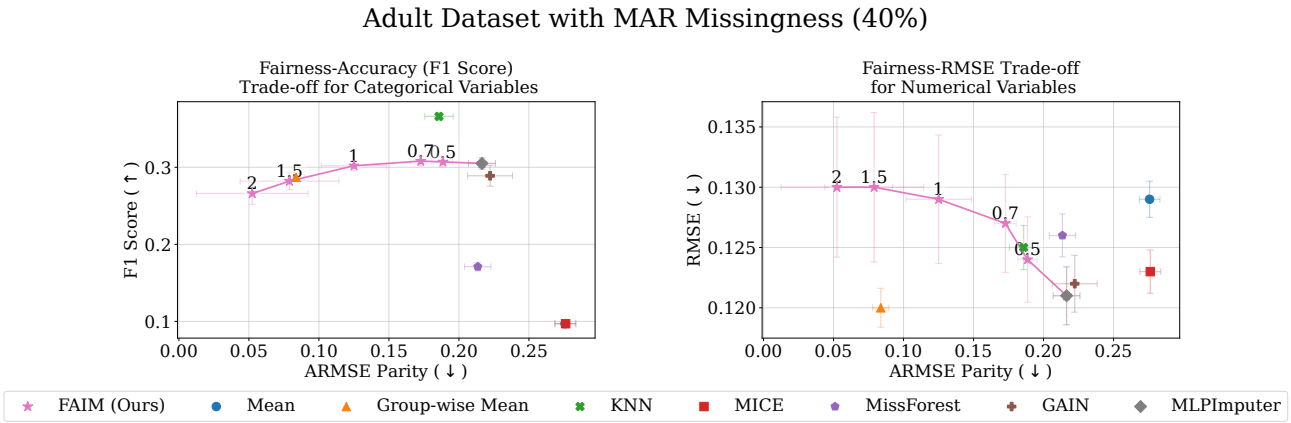

Figure 12: Imputation fairness improvement for the **Adult** dataset under **MAR** at **40%** missingness.

E.2.2 FOLK INCOME DATASET RESULTS

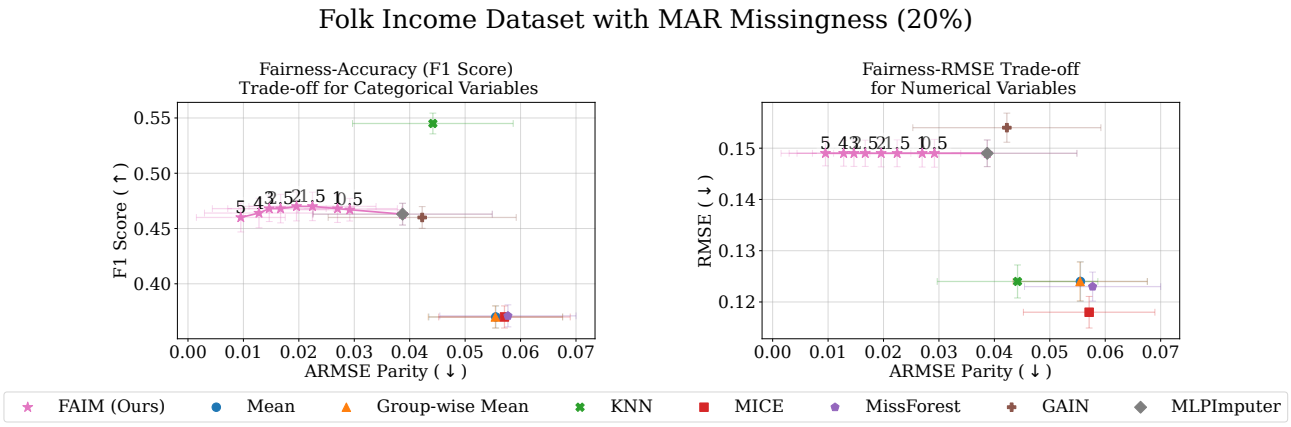

Figure 13: Imputation fairness improvement for the **Folk Income** dataset under **MAR** at **20%** missingness.

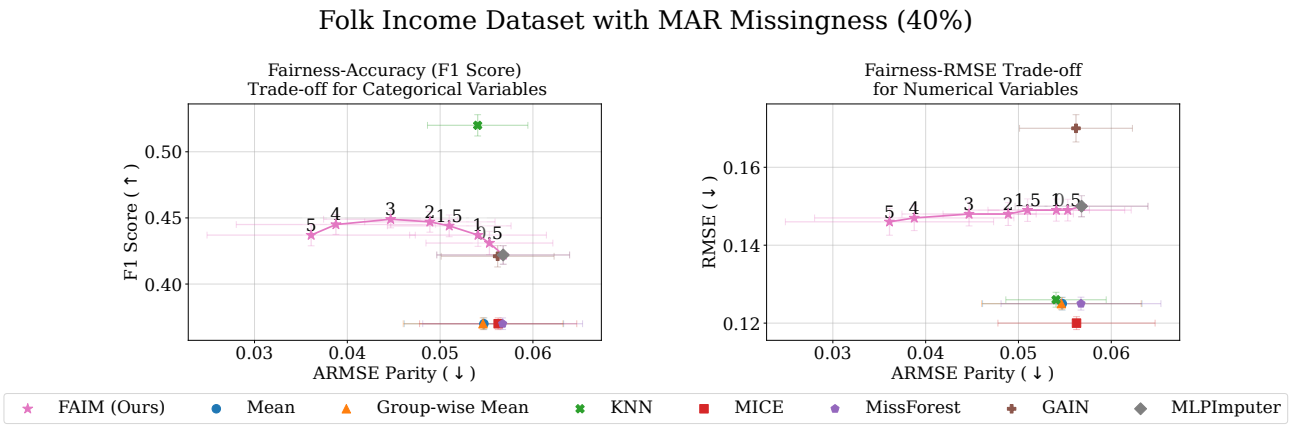

Figure 14: Imputation fairness improvement for the **Folk Income** dataset under **MAR** at **40%** missingness.

E.3 MNAR Experiments

E.3.1 ADULT DATASET RESULTS

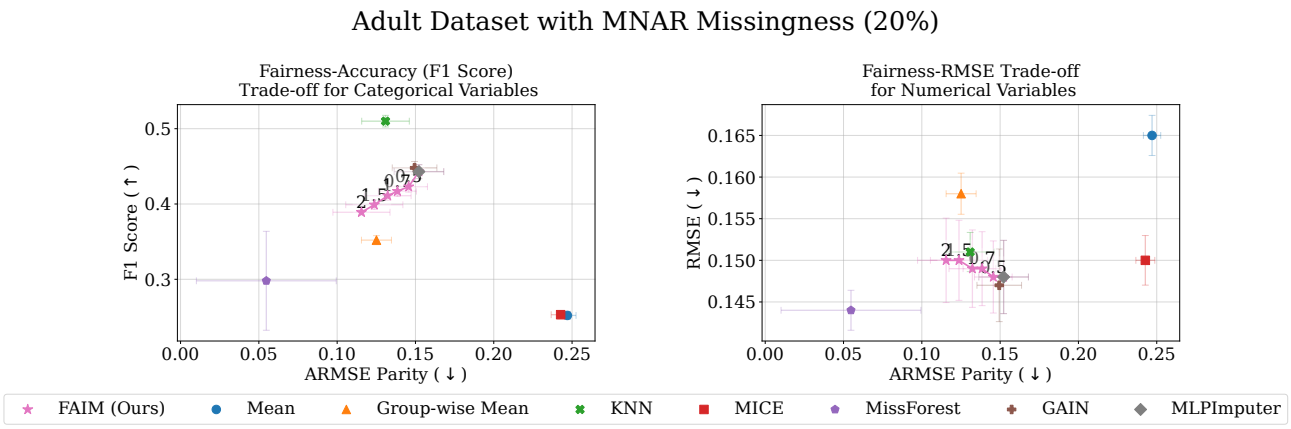

Figure 15: Imputation fairness improvement for the **Adult** dataset under **MNAR** at **20%** missingness.

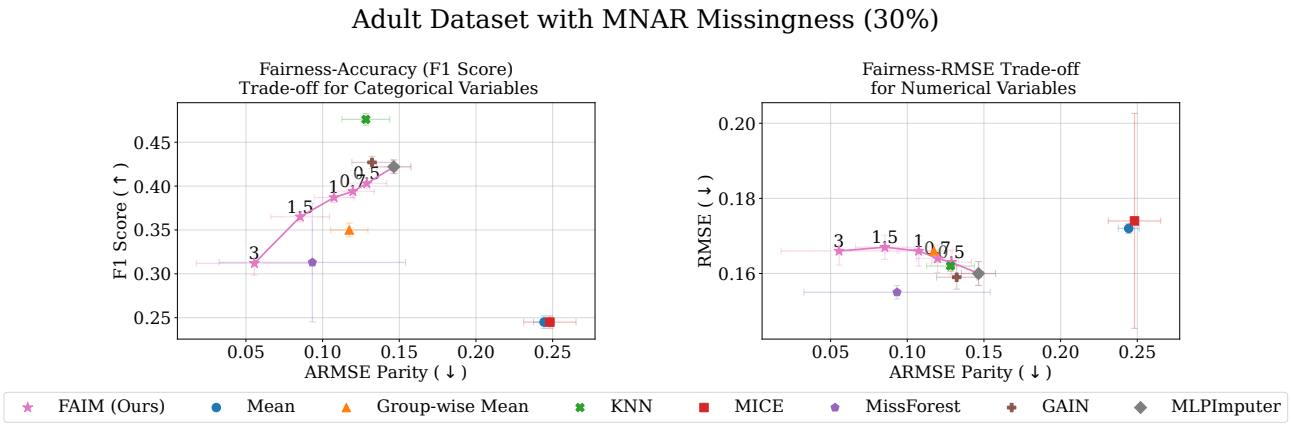

Figure 16: Imputation fairness improvement for the **Adult** dataset under **MNAR** at **30%** missingness.

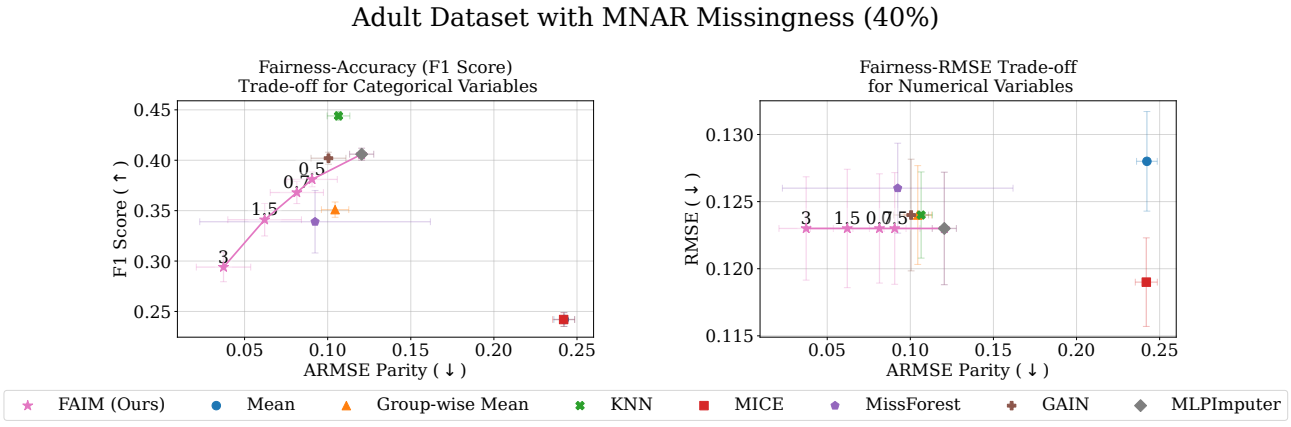

Figure 17: Imputation fairness improvement for the **Adult** dataset under **MNAR** at **40%** missingness.

E.3.2 Folk Income Dataset Results

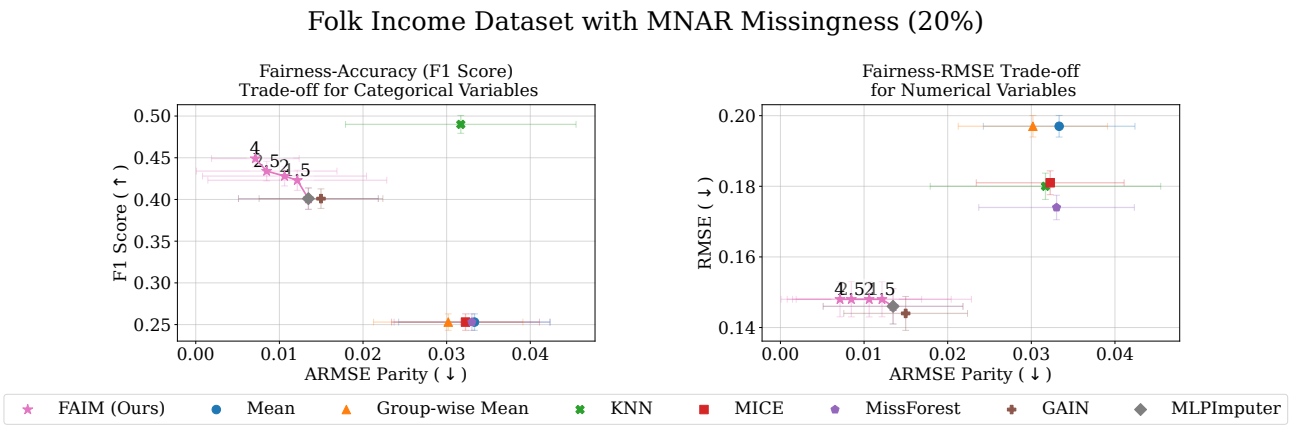

Figure 18: Imputation fairness improvement for the **Folk Income** dataset under **MNAR** at **20%** missingness.

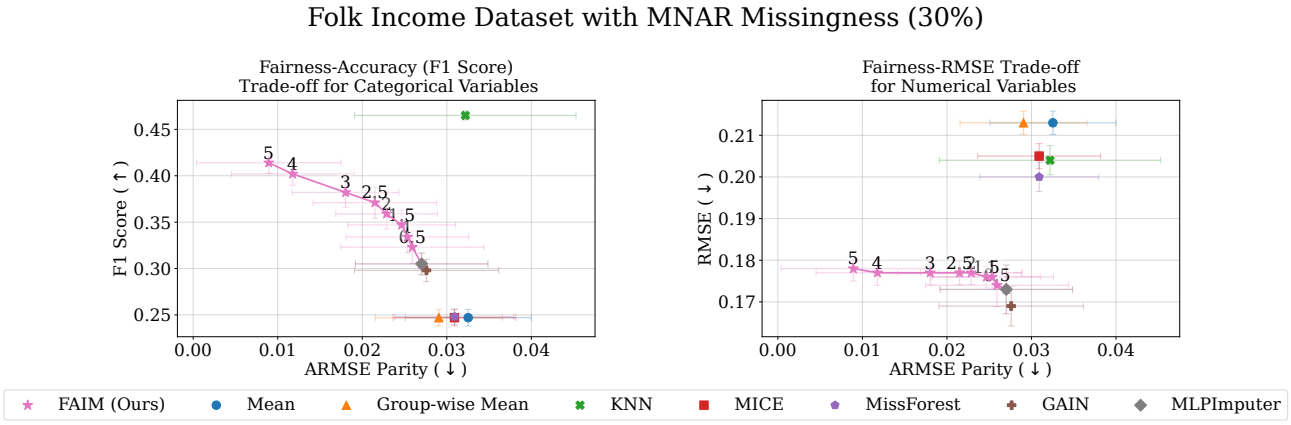

Figure 19: Imputation fairness improvement for the **Folk Income** dataset under **MNAR** at **30%** missingness.

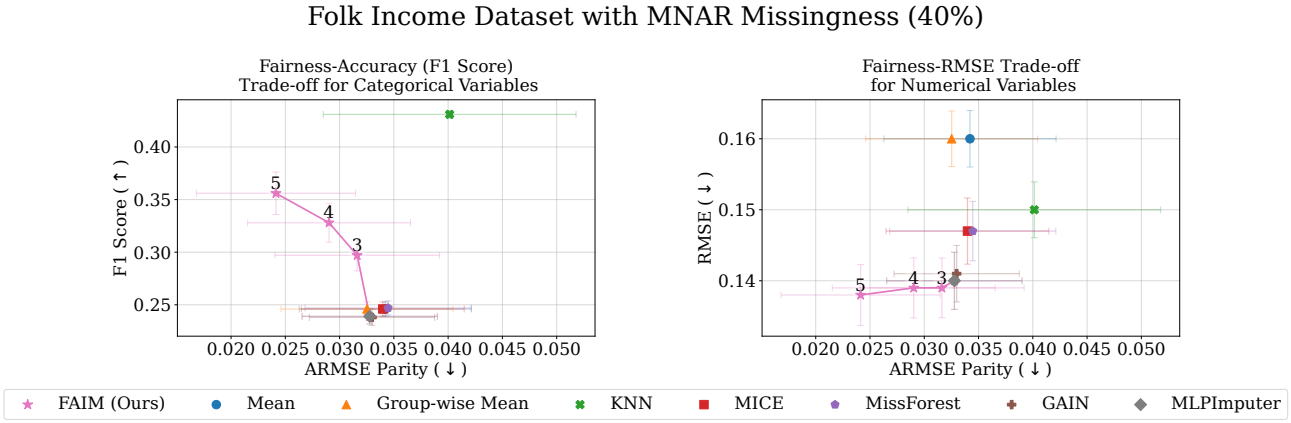

Figure 20: Imputation fairness improvement for the **Folk Income** dataset under **MNAR** at **40%** missingness.

F Downstream Fairness Experiment Results

In this section, we present empirical results on our synthetic dataset derived from the Law School dataset, evaluating all three missing data mechanisms at 20%, 30%, and 40% missingness levels. As in Section E, the markers are annotated with the γ hyperparameter, which controls the trade-off between imputation performance and fairness. Higher values of γ generally lead to improved imputation fairness and better downstream accuracy parity. We report results using both linear and non-linear classifiers, specifically logistic regression and random forest.

F.1 MCAR Experiments

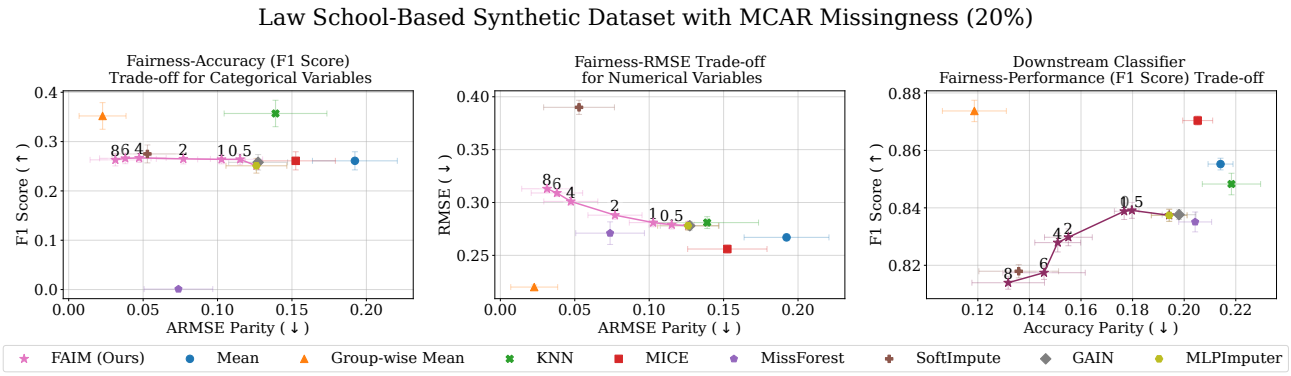

Figure 21: Imputation and downstream fairness improvement on the **synthetic** dataset under **MCAR** at **20%** missingness. Left: F1 score vs. ARMSE parity. Middle: RMSE vs. ARMSE parity. Right: downstream F1 score vs. accuracy parity for the **logistic regression** classifier.

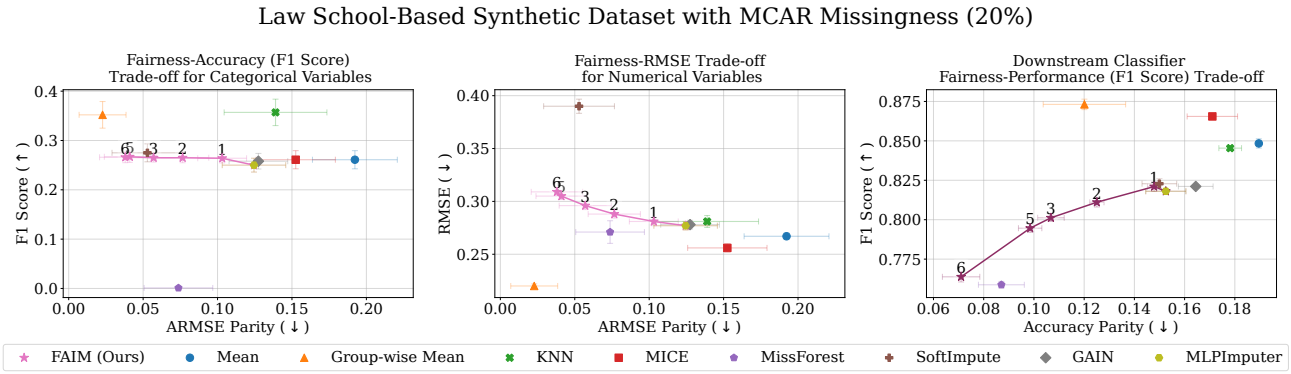

Figure 22: Imputation and downstream fairness improvement on the **synthetic** dataset under **MCAR** at **20%** missingness. Left: F1 score vs. ARMSE parity. Middle: RMSE vs. ARMSE parity. Right: downstream F1 score vs. accuracy parity for the **random forest** classifier.

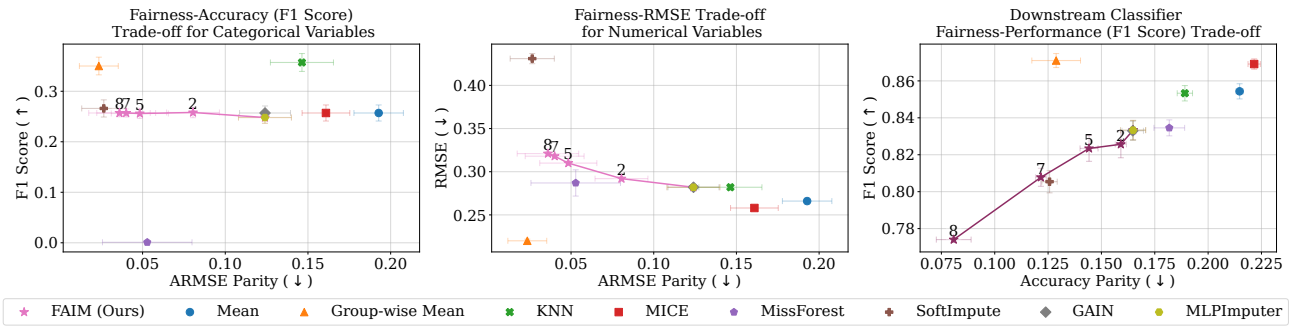

Figure 23: Imputation and downstream fairness improvement on the **synthetic** dataset under **MCAR** at **30%** missingness. Left: F1 score vs. ARMSE parity. Middle: RMSE vs. ARMSE parity. Right: downstream F1 score vs. accuracy parity for the **logistic regression** classifier.

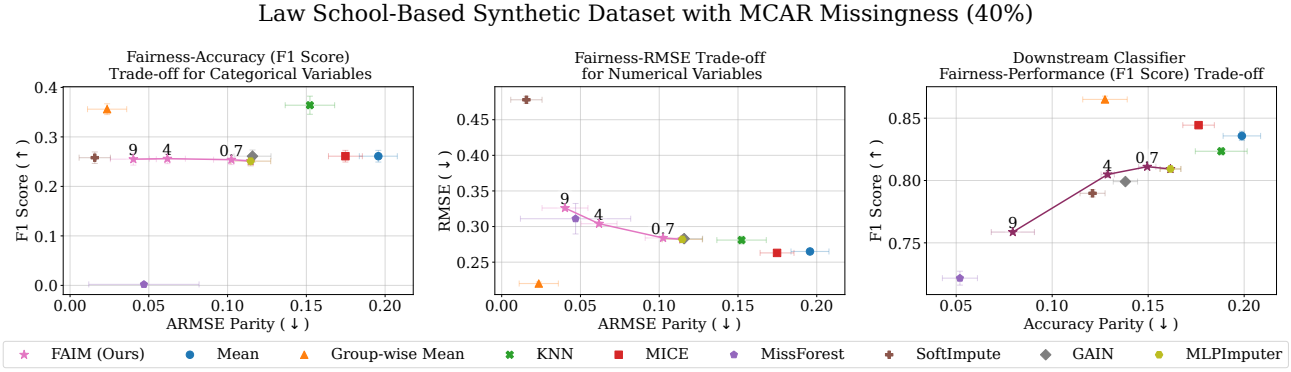

Figure 26: Imputation and downstream fairness improvement on the **synthetic** dataset under **MCAR** at **40%** missingness. Left: F1 score vs. ARMSE parity. Middle: RMSE vs. ARMSE parity. Right: downstream F1 score vs. accuracy parity for the **random forest** classifier.

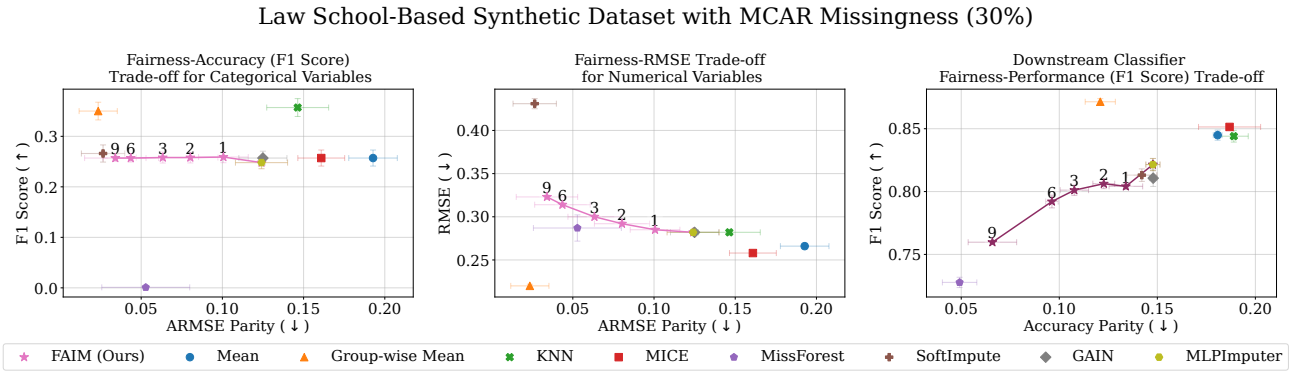

Figure 24: Imputation and downstream fairness improvement on the **synthetic** dataset under **MCAR** at **30%** missingness. Left: F1 score vs. ARMSE parity. Middle: RMSE vs. ARMSE parity. Right: downstream F1 score vs. accuracy parity for the **random forest** classifier.

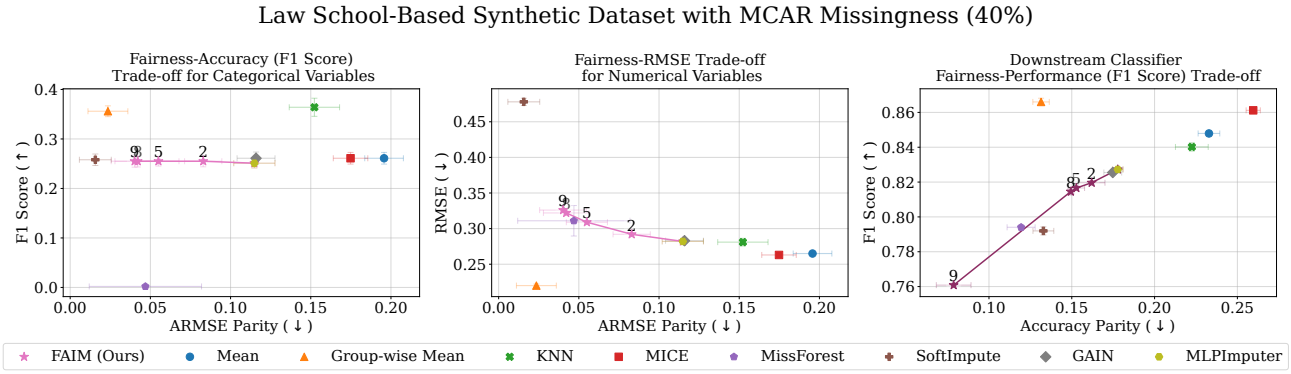

Figure 25: Imputation and downstream fairness improvement on the **synthetic** dataset under **MCAR** at **40%** missingness. Left: F1 score vs. ARMSE parity. Middle: RMSE vs. ARMSE parity. Right: downstream F1 score vs. accuracy parity for the **logistic regression** classifier.

F.2 MAR Experiments

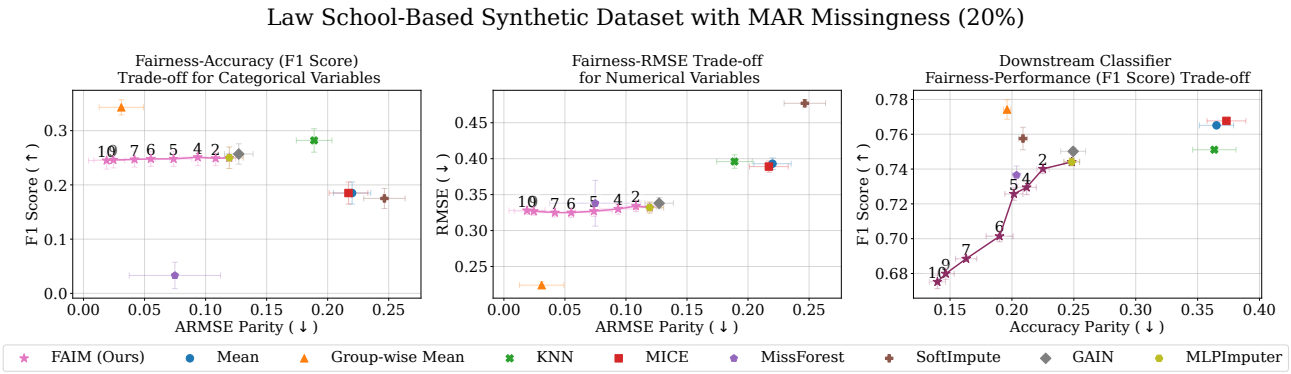

Figure 27: Imputation and downstream fairness improvement on the **synthetic** dataset under **MAR** at **20%** missingness. Left: F1 score vs. ARMSE parity. Middle: RMSE vs. ARMSE parity. Right: downstream F1 score vs. accuracy parity for the **logistic regression** classifier.

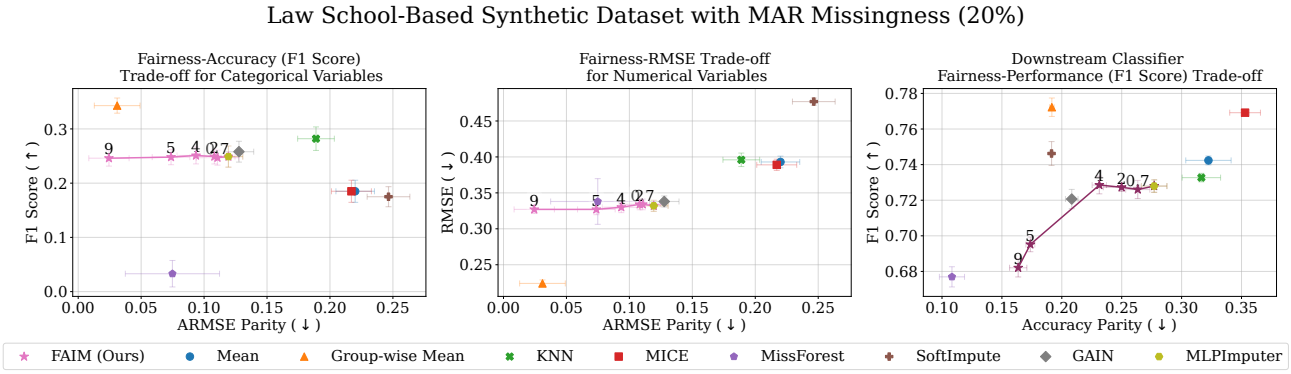

Figure 28: Imputation and downstream fairness improvement on the **synthetic** dataset under **MAR** at **20%** missingness. Left: F1 score vs. ARMSE parity. Middle: RMSE vs. ARMSE parity. Right: downstream F1 score vs. accuracy parity for the **random forest** classifier.

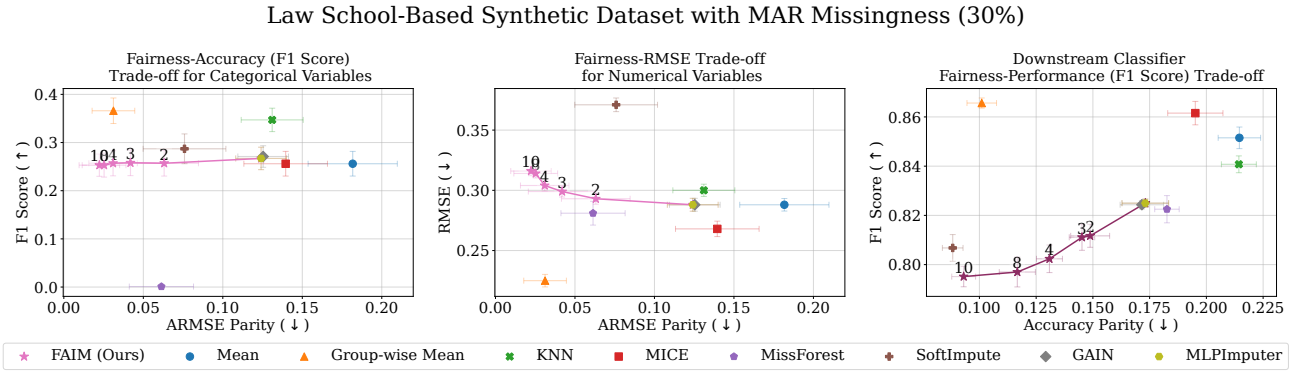

Figure 29: Imputation and downstream fairness improvement on the **synthetic** dataset under **MAR** at **30%** missingness. Left: F1 score vs. ARMSE parity. Middle: RMSE vs. ARMSE parity. Right: downstream F1 score vs. accuracy parity for the **logistic regression** classifier.

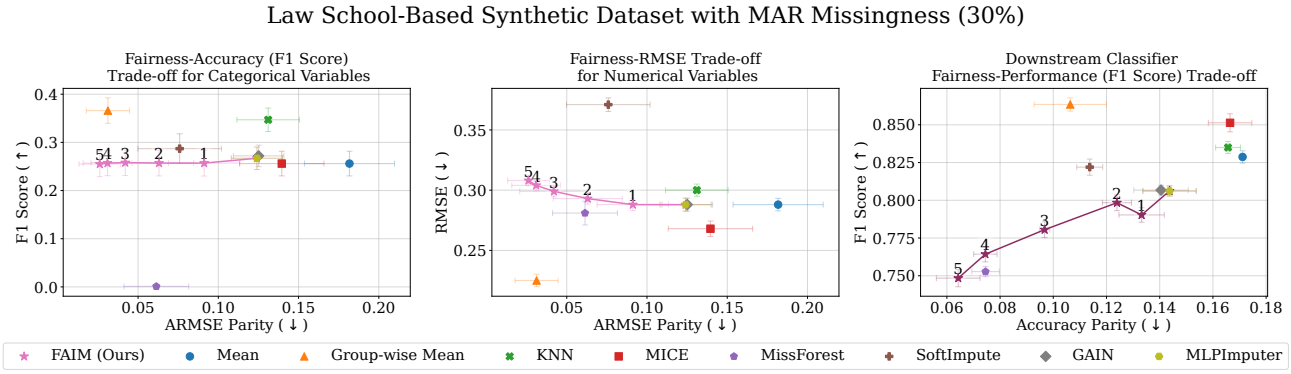

Figure 30: Imputation and downstream fairness improvement on the **synthetic** dataset under **MAR** at **30%** missingness. Left: F1 score vs. ARMSE parity. Middle: RMSE vs. ARMSE parity. Right: downstream F1 score vs. accuracy parity for the **random forest** classifier.

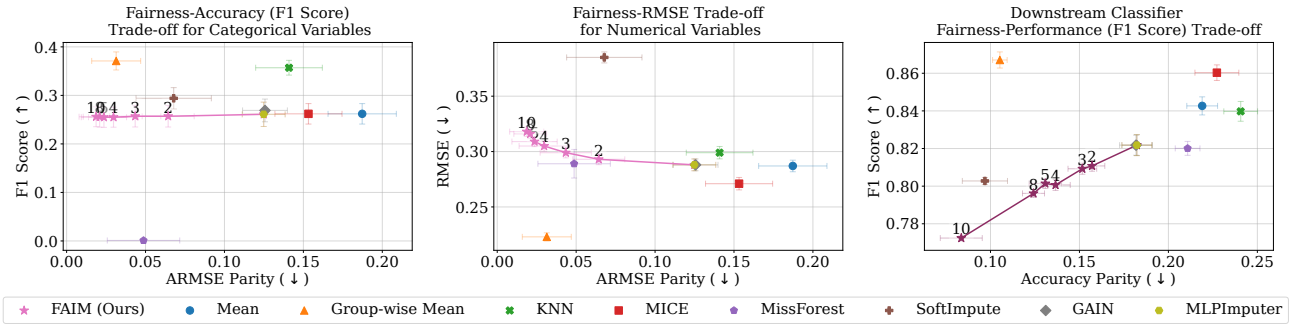

Figure 31: Imputation and downstream fairness improvement on the **synthetic** dataset under **MAR** at **40%** missingness. Left: F1 score vs. ARMSE parity. Middle: RMSE vs. ARMSE parity. Right: downstream F1 score vs. accuracy parity for the **logistic regression** classifier.

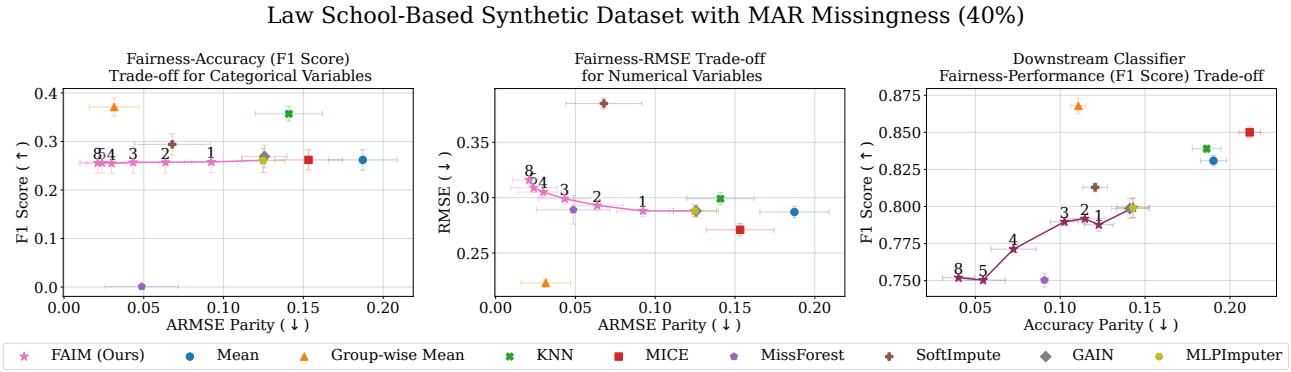

Figure 32: Imputation and downstream fairness improvement on the **synthetic** dataset under **MAR** at **40%** missingness. Left: F1 score vs. ARMSE parity. Middle: RMSE vs. ARMSE parity. Right: downstream F1 score vs. accuracy parity for the **random forest** classifier.

F.3 MNAR Experiments

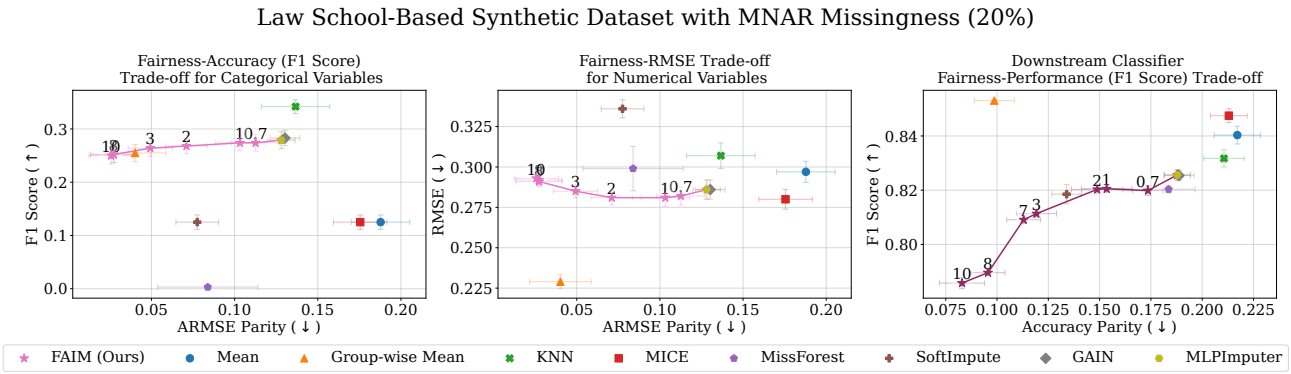

Figure 33: Imputation and downstream fairness improvement on the **synthetic** dataset under **MNAR** at **20%** missingness. Left: F1 score vs. ARMSE parity. Middle: RMSE vs. ARMSE parity. Right: downstream F1 score vs. accuracy parity for the **logistic regression** classifier.

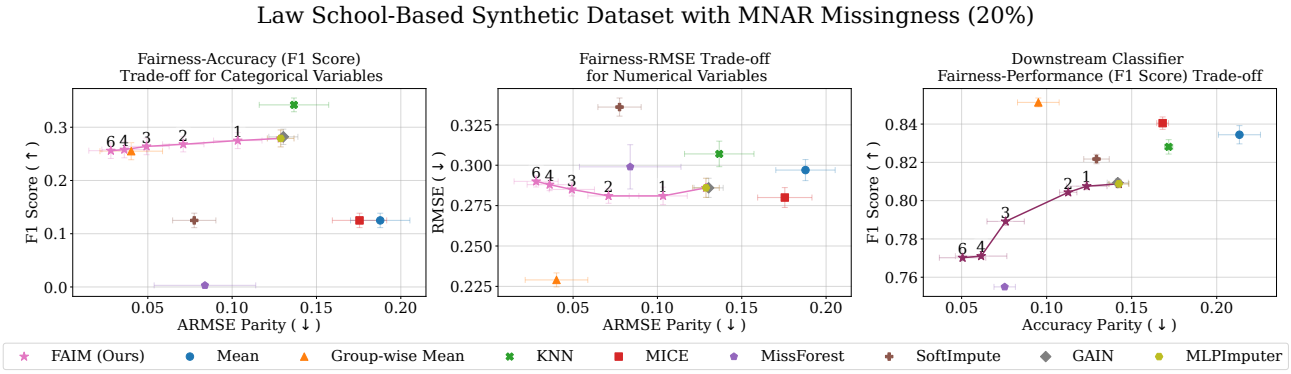

Figure 34: Imputation and downstream fairness improvement on the **synthetic** dataset under **MNAR** at **20%** missingness. Left: F1 score vs. ARMSE parity. Middle: RMSE vs. ARMSE parity. Right: downstream F1 score vs. accuracy parity for the **random forest** classifier.

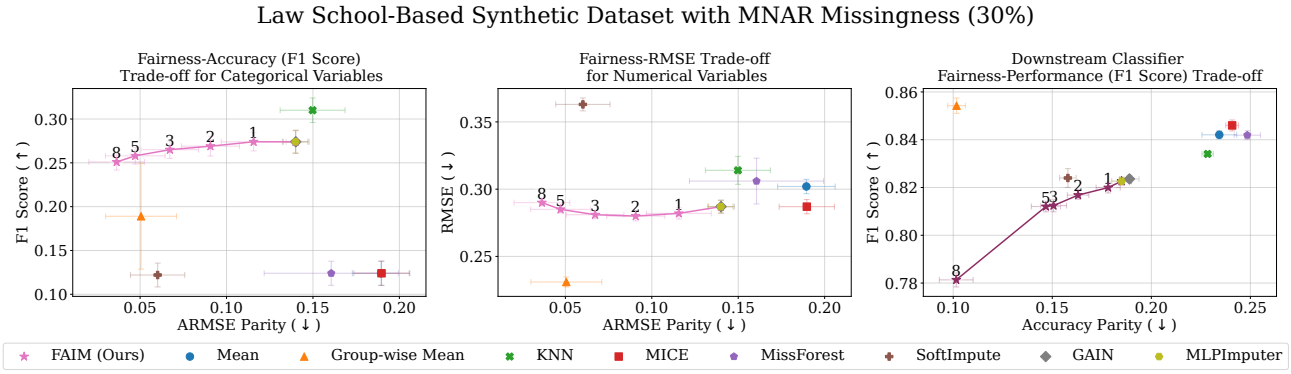

Figure 35: Imputation and downstream fairness improvement on the **synthetic** dataset under **MNAR** at **30%** missingness. Left: F1 score vs. ARMSE parity. Middle: RMSE vs. ARMSE parity. Right: downstream F1 score vs. accuracy parity for the **logistic regression** classifier.

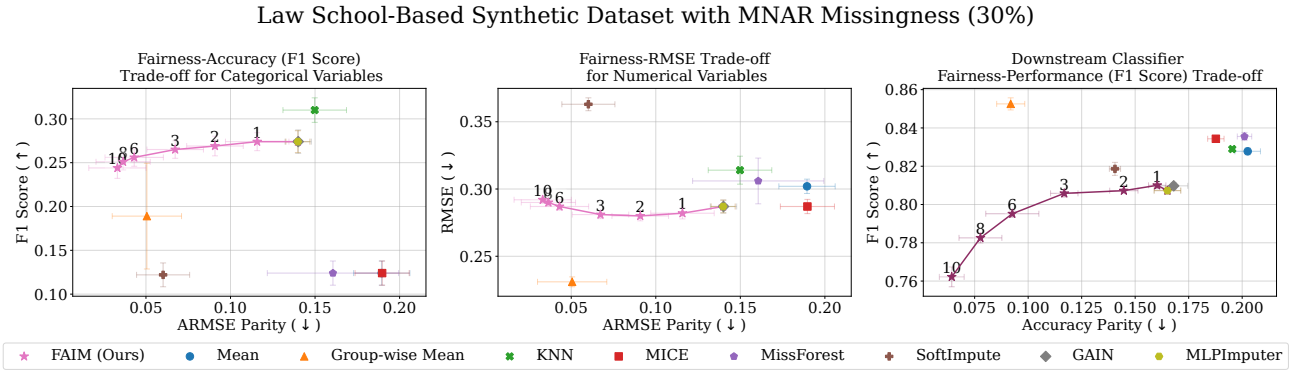

Figure 36: Imputation and downstream fairness improvement on the **synthetic** dataset under **MNAR** at **30%** missingness. Left: F1 score vs. ARMSE parity. Middle: RMSE vs. ARMSE parity. Right: downstream F1 score vs. accuracy parity for the **random forest** classifier.

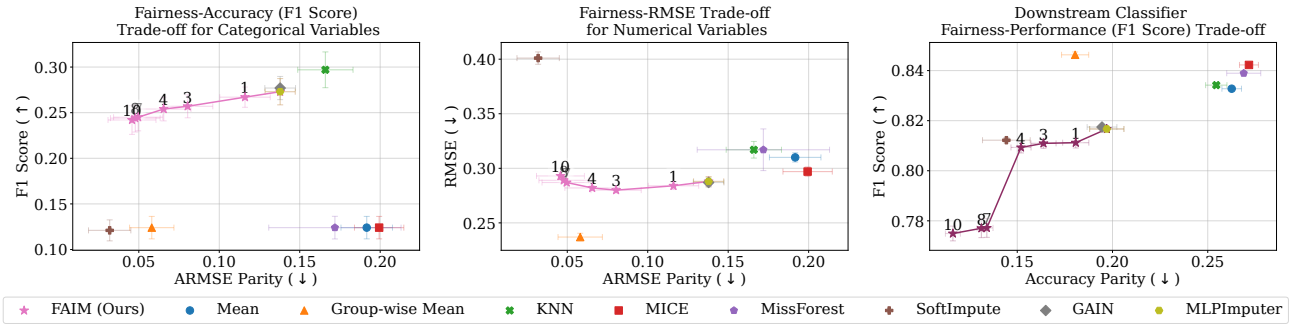

Figure 37: Imputation and downstream fairness improvement on the **synthetic** dataset under **MNAR** at **40%** missingness. Left: F1 score vs. ARMSE parity. Middle: RMSE vs. ARMSE parity. Right: downstream F1 score vs. accuracy parity for the **logistic regression** classifier.

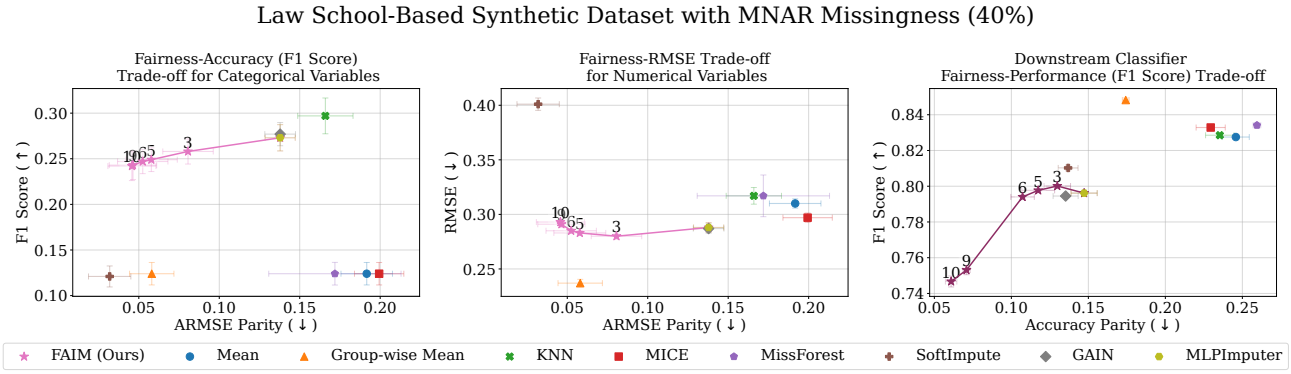

Figure 38: Imputation and downstream fairness improvement on the **synthetic** dataset under **MNAR** at **40%** missingness. Left: F1 score vs. ARMSE parity. Middle: RMSE vs. ARMSE parity. Right: downstream F1 score vs. accuracy parity for the **random forest** classifier.