Uncovering Algorithmic Fairness in Deep Learning–Based Imputation of Multivariate Clinical Time Series in Heart Failure Patients

**Rayssa Muniz [1], Victor Lima[1],
Paloma Saldanha[1], Andrea Ribeiro[3], Rodrigo de Paula[3]
Martin Cadeiras[4], Carlo da Silva[2], Hítalo Silva [2], Paulo Rocha[2] Diego Pinheiro[1]**

[1] Universidade Católica de Pernambuco (UNICAP)

[2]Universidade de Pernambuco (UPE)

[3]Universidade Federal de Pernambuco (UFPE)

[4]University of California, Davis (UCDAVIS)

{rayssa.00000829609,victor.00000829618,
diego.silva,paloma.saldanha}@unicap.br, {ph.alcantara.rocha,
hitaloosilva}@gmail.com, andrea.marianogueira@ufpe.br,
rpmonteiro90@yahoo.com.br, mcadeiras@ucdavis.edu,
cmrs@poli.br

Abstract. *Deep learning for missing data imputation (MDI) in healthcare time series has advanced, but fairness concerns remain underexplored. Traditional evaluations focus on global error metrics (e.g., MAE), overlooking disparities across variables and protected subgroups. We analyze five state-of-the-art MDI models (BRITS, SAITS, USGAN, GPVAE, MRNN) on a heart failure dataset (PhysioNet), assessing fairness via Lorenz curves and the Gini coefficient. Results reveal that low MAE does not imply fairness. SAITS was most efficient (MAE = 0.241) but least fair (Gini = 0.615), while MRNN was less efficient (MAE = 0.672) yet fairer (Gini = 0.439). We highlight the need to balance accuracy and fairness in MDI for clinical applications.*

1. Introduction

Multivariate time series (MTS) arising from healthcare data are often incomplete and irregular, making reliable clinical inference challenging [Tipirneni and Reddy 2022, Liu et al. 2023, Wang et al. 2024]. Spanning domains from intensive care to chronic disease management, MTS can represent a wide range of data types, including wearable data (e.g., heart rate), clinical measurements (e.g., lab results), and omics profiles (e.g., gene expression). Within this healthcare context, MTS are shaped by patients' clinical conditions and the natural course of care delivery, leading to sparsity due to missing data and irregularity due to asynchronous sampling. For instance, heart rate data from a wearable sensor may be recorded every few seconds, while cholesterol levels from lab tests are measured only every few weeks—resulting in misaligned and incomplete time series. Given their sparsity and irregularity, a fundamental step in handling healthcare MTS is missing data imputation (MDI).

State-of-the-art MDI methods increasingly rely on deep learning models. Initially, MDI was dominated by statistical approaches that filled in missing values using simple heuristics such as the mean, the last observed value, or even zero [Wang et al. 2024]. These replacements, while easy to implement, often misrepresent temporal dependencies and correlations between variables within the observed time interval. Even statistical models such as Autoregressive Integrated Moving Average (ARIMA) and its seasonal extension (SARIMA), which capture only linear dependencies, fail to model the complex temporal relationships and variation patterns present in MTS [Mesquita et al. 2024]. These limitations have been overcome by deep learning methods, such as Self-attention-Based Imputation for Time Series (SAITS) [Du et al. 2023], Bidirectional Recurrent Imputation for Time Series (BRITS) [Cao et al. 2018], Gaussian Process Variational Autoencoder (GPVAE) [Fortuin et al. 2020], Unsupervised Generative Adversarial Network (USGAN) [Miao et al. 2021] and Multidirectional Recurrent Neural Network (M-RNN) [Yoon et al. 2019]. These methods can learn the approximate distribution of the underlying data from the observed data [Wang et al. 2024]. Nevertheless, MDI in healthcare MTS is typically assessed only by aggregate efficiency metrics—such as mean absolute error (MAE)—while the algorithmic fairness of the underlying deep-learning models remains unexamined.

Algorithmic fairness—particularly in healthcare—cannot be overlooked, as algorithms used for clinical inference and decision-making have the potential to cause harm [Pfohl et al. 2024, Obermeyer et al. 2019]. For instance, lack of algorithmic fairness can lead to biased treatment assignments or diagnostic testing for patients from different sociodemographic groups—often defined by protected variables such as race, gender, or socioeconomic status—even when they present similar clinical profiles [Omar et al. 2024]. Missing data is a major source of algorithmic unfairness. Although multiple notions of fairness have been proposed, they primarily focus on downstream classification tasks, often overlooking fairness concerns introduced during deep learning–based missing data imputation (MDI) [Verma and Rubin 2018, Min et al. 2025].

In this paper, we characterize not only the algorithmic efficiency but also the fairness of state-of-the-art deep learning models for MDI. Efficiency is measured in terms of imputation error using the MAE, while fairness is assessed through the distribution of imputation error using Lorenz Curves and the Gini coefficient, which quantifies inequality in the distribution of imputation errors. We conduct our analysis using a comprehensive real-world clinical dataset of heart failure patients from PhysioNet 2012 [Silva et al. 2012b]. Our results reveal a trade-off between efficiency and fairness, where more efficient models tend to be less fair, and vice versa. This work sheds light on the algorithmic fairness of deep learning models for MDI and underscores the need for imputation approaches in healthcare MTS that consider not only efficiency but also fairness.

2. Background

Multivariate Time Series. Let $(S, X, Y) \sim \mathcal{F}$ be a triplet drawn from an (unknown) distribution, where:

- $S = [s_1, \ldots, s_T]$ is a sequence of time stamps;
- $X \in \mathbb{R}^{T \times D}$ is a MTS of length T with D variables, and $\mathbf{x}_t \in \mathbb{R}^D$ denotes the t-th row (the measurements collected at time s_t);
- Y contains optional labels or outcomes associated with the series (e.g., mortality or readmission).

Missingness mask. Given that in MTS, we observe only a subset of X, a missing mask $M \in \{0,1\}^{T\times D}$ is defined by

$$m_{t,d} = \begin{cases} 1, & \text{if } x_{t,d} \text{ is observed (i.e., variable } d \text{ was measured at } s_t), \\ 0, & \text{otherwise.} \end{cases}$$

Imputation model class. Consider a family of imputation functions $\{f_\theta : \theta \in \Theta\}$. Each f_θ takes the partially observed pair (X, M) and outputs an imputed series:

$$\hat{X}_\theta = f_\theta(X, M) \in \mathbb{R}^{T\times D} \quad ,$$

such that an estimate for entry (t, d) is defined as $\hat{x}^{(\theta)}_{t,d} = \hat{X}_\theta$.

Parameter fitting. The optimal parameters $\hat{\theta} = \mathcal{A}(X, M)$ are obtained with some algorithm $\mathcal{A}$ by minimizing the expected reconstruction loss over the missing entries:

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \mathbb{E}_{(X,M)\sim\mathcal{F}} \left[\sum_{t=1}^{T} \sum_{d=1}^{D} (1 - m_{t,d})\, \mathcal{L}\big(x_{t,d},\, \hat{x}^{(\theta)}_{t,d}\big) \right],$$

where $\mathcal{L}$ is a pointwise loss such as the squared error $\mathcal{L}(x_{t,d}, \hat{x}^{(\theta)}_{t,d}) = (x_{t,d} - \hat{x}^{(\theta)}_{t,d})^2$.

Missing Data Mechanisms. Missing data can arise from different underlying mechanisms, typically classified as Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR) [Liu et al. 2023]. Under MCAR, the probability of missingness is independent of both observed and unobserved data; that is, data are missing purely by chance. In the MAR case, missingness may depend on the observed data but not on the missing values themselves. Finally, under MNAR, the probability of missingness depends on the missing values and possibly also on the observed data, implying a systematic, non-random pattern of missingness.

3. Related Work

3.1. Deep Learning Models for Missing Data Imputation in Multivariate Time Series

Currently, state-of-the-art approaches for MDI in MTS are predominantly based on deep learning models. Among these is SAITS [Du et al. 2023], a model that applies the self-attention technique to focus on different parts of a time series, assigning different levels of importance to different parts of the time series. SAITS captures the dependencies between different moments of the series using Diagonal Self-Attention Masks (DSMA) that have multiple attention heads.

BRITS [Cao et al. 2018], another prominent model, is based on a bidirectional Recurrent Neural Network (RNN). The model imputes missing values by treating them as variables within a recurrent dynamical system, learning their values from the observed data in the RNN graph.

The USGAN model [Miao et al. 2021] employs a generative adversarial framework, featuring a generator to fill in missing values and a discriminator to verify the authenticity of the imputed data. This model integrates the discriminator with a temporal reminder matrix, a method that introduces additional complexity to the discriminator's training to yield performance improvements.

GPVAE [Fortuin et al. 2020] offers a hybrid approach by combining a Variational Autoencoder (VAE) with a Gaussian Process (GP). The VAE component maps the incomplete time series data into a complete latent space representation. Within this latent space, the GP component then models the temporal dynamics, capturing correlations that are used to impute the missing values.

Finally, the MRNN model [Yoon et al. 2019] operates both within and across data streams, performing interpolation within each stream and imputation across different streams. It employs a bidirectional RNN for interpolation, followed by a multilayer perceptron (MLP) for imputation. MRNN accounts for time lags between measurements and uniquely treats missing values as fixed constants, such as zero or the variable mean, which remain unchanged during the training process.

3.2. Algorithmic Bias and Fairness

Deep learning methods, despite achieving efficient results, may have some limitations such as interpretability, fairness and equity, which can be critical especially in healthcare scenarios [Meng et al. 2022]. Several criteria have been proposed to define fairness in algorithmic decision-making [Verma and Rubin 2018]. One approach focuses on statistical parity, whereby individuals or groups should receive similar treatment according to measurable indicators. Another emphasizes equality of outcomes, according to which fairness is achieved when all demographic groups, such as those defined by gender, age, or race (protected groups or variables), receive positive results at comparable rates. Deviations from these patterns may signal algorithmic unfairness. However, beyond statistical metrics, fairness in artificial intelligence must also be understood as a normative value. Algorithms trained on historical data are prone to reproducing or even exacerbating preexisting social inequalities, thus raising concerns about algorithmic injustice. Ensuring fairness, therefore, entails more than satisfying mathematical definitions: it requires aligning automated decisions with human values such as equity and dignity, adopting mechanisms for transparency and accountability, and safeguarding the rights of vulnerable or historically marginalized groups [Russell 2020].

4. Methods

4.1. Multivariate Time Series Dataset

This study used the PhysioNet Challenge 2012 dataset [Silva et al. 2012b], which contains data from $12,000$ patients inserted into an Intensive Care Unit (ICU) context. Among the variables, there is a subset collected upon the patient's admission to the ICU, such as age, gender, height, type of ICU, and admission weight, which are treated as static variables because those values do not change throughout the observed time series. There are also variables that correspond to time series clinical measurements, such as heart rate, temperature, and platelet count. The data present in the set correspond to the first 48 hours after the patient's admission to the ICU, and the time markings, the moments when the observed variables are measured, occur every 1 hour.

The dataset comprises a comprehensive panel of routine ICU measurements, including laboratory results, vital signs, and respiratory parameters. While all available variables were considered during training and evaluation, we report results for a selected subset—Cholesterol, Troponin I, Troponin T, AST, pH, and Urine output. These variables were chosen due to their diverse missingness patterns and the pronounced variability in mean absolute error (MAE) observed across models and subgroups. The full list of variables is available in the dataset documentation [Silva et al. 2012a].

4.2. Stratification by Variable and Protected Subgroups

The dataset was divided into demographic subgroups to better analyze and characterize bias in the imputation generated by a model. Data were stratified by two protected variables: gender and age. For gender, the levels are female and male. For age, patients are grouped as 65 years or older and under 65 years.

4.3. Pre-processing of Absence Levels in Multivariate Time Series

To preprocess the data, we used PyPOTS [Wenjie 2023], an open-source library ecosystem for machine learning research on partially observed time series that provides components for each stage of the data pipeline. The Times Series Data Beans (TSDB) library was used to load the PhysioNet Challenge 2012 database. BenchPOTS, another PyPOTS component, provided a standard and unified preprocessing pipeline for a variety of datasets, including the PhysioNet Challenge 2012. The component was used to remove patients who have data measured only at the time of admission to the Intensive Care Unit (ICU) and do not have time series data, after which 11,988 patients remained. In addition, BenchPOTS was also used to divide the test dataset into protected subgroups. Finally, the PyGrinder toolkit was used to introduce synthetic missing values into the dataset. An artificial missingness was applied to the data on a scale of 10% under a MCAR mechanism.

4.4. Measuring Algorithmic Efficiency as Imputation Error

Algorithmic efficiency is defined as the mean absolute imputation error. Define the random variable

$$\varepsilon_{t,d}^{(\theta)} = (1 - m_{t,d})\,|x_{t,d} - \hat{x}_{t,d}^{(\theta)}|.$$

Then the mean absolute error of model f_θ is

$$\mathrm{MAE}(\theta) = \frac{\mathbb{E}_{(X,M)\sim\mathcal{F}}\Big[\sum_{t,d}\varepsilon_{t,d}^{(\theta)}\Big]}{\mathbb{E}_{(X,M)\sim\mathcal{F}}\Big[\sum_{t,d}(1 - m_{t,d})\Big]},$$

i.e. the expected absolute error *per missing entry*. Given a test set with missing-index set $\mathcal{I}_0 = \{(t,d) \mid m_{t,d} = 0\}$ and $n = |\mathcal{I}_0|$, the empirical MAE for f_θ is

$$\widehat{\mathrm{MAE}}(\theta) = \frac{1}{n}\sum_{(t,d)\in\mathcal{I}_0} |x_{t,d} - \hat{x}_{t,d}^{(\theta)}|.$$

Lower values of $\mathrm{MAE}(\theta)$ or $\widehat{\mathrm{MAE}}(\theta)$ indicate higher imputation accuracy.

4.5. Characterization of Algorithmic Fairness

Algorithmic fairness is characterized based on the distribution of imputation error. Given a model $f_{\hat{\theta}}$, the imputed value at position (t, d) is $\hat{x}^{\hat{\theta}}_{t,d} = f_{\hat{\theta}}(X, M)_{t,d}$, and the corresponding imputation error is $\varepsilon^{\hat{\theta}}_{t,d} = \mathcal{L}(x_{t,d}, \hat{x}^{\hat{\theta}}_{t,d}), \quad$ for $m_{t,d} = 0$. Let $\varepsilon \mid \hat{\theta} \sim \mathcal{D}(\hat{\theta})$ denote the conditional distribution of imputation errors under model $f_{\hat{\theta}}$, defined over the set of missing entries. The Lorenz curve $L_{\hat{\theta}} : [0, 1] \to [0, 1]$ is

$$L_{\hat{\theta}}(p) = \frac{1}{\mathbb{E}[\varepsilon \mid \hat{\theta}]} \int_0^p F^{-1}_{\varepsilon|\hat{\theta}}(q)\, dq,$$

where $F^{-1}_{\varepsilon|\hat{\theta}}$ is the quantile function of $\varepsilon \mid \hat{\theta}$, the conditional distribution of imputation errors. The associated (population) Gini coefficient is

$$\text{Gini}(\hat{\theta}) = 1 - 2 \int_0^1 L_{\hat{\theta}}(p)\, dp.$$

Given the n imputation errors $\varepsilon_1, \ldots, \varepsilon_n$ for the missing entries in a test set, sort them as $\varepsilon_{(1)} \leq \cdots \leq \varepsilon_{(n)}$. The empirical Lorenz curve at $p_k = k/n$ $(k = 1, \ldots, n)$ is

$$\hat{L}_{\hat{\theta}}\left(\frac{k}{n}\right) = \frac{\sum_{i=1}^{k} \varepsilon_{(i)}}{\sum_{i=1}^{n} \varepsilon_{(i)}}.$$

The corresponding empirical Gini coefficient is

$$\widehat{\text{Gini}}(\hat{\theta}) = 1 - \frac{2}{n} \sum_{k=1}^{n} \hat{L}_{\hat{\theta}}\left(\frac{k}{n}\right) = \frac{2 \sum_{i=1}^{n} i\, \varepsilon_{(i)}}{n \sum_{i=1}^{n} \varepsilon_{(i)}} - \frac{n+1}{n}.$$

Together, $L_{\hat{\theta}}(\cdot)$ and $\text{Gini}(\hat{\theta})$ (or their empirical counterparts) quantify how unequally imputation errors are distributed—lower Gini values indicate fairer error allocation across missing entries.

4.6. Experimental Setup

The dataset was partitioned into training (64%), validation (16%), and testing (20%) subsets. The 64% training portion does not contain synthetic missing values and is used by the models to learn how to infer missing values in MTS. The 16% validation subset contains synthetic missing values and is used to evaluate the performance of the models during training, avoiding overfitting. The remaining 20% testing subset also contains synthetic missing values and is used in statistical analyses to measure the imputation error of the models.

We evaluated five models: SAITS, BRITS, USGAN, GPVAE, and MRNN. All models were trained using the Adam optimizer with a learning rate of 0.01, a batch size of 32, and a maximum of 10 epochs, with early stopping patience set to 3. Each model had a distinct architecture and set of hyperparameters:

Table 1. Model architectures and hyperparameters

Model	# Parameters	Key Architectural/Hyperparameter Details
SAITS	7,201,082	4 attention heads, embedding size = 256, feed-forward size = 128, key/value size = 64
BRITS	2,393,044	Hidden state size = 128
USGAN	12,585,017	Hidden size (G/D) = 256, 1 training step per generator/discriminator per iteration
GPVAE	2,296,052	Latent space size = 37, Cauchy kernel, encoder: 2×128 neurons, decoder: 2×256 neurons
MRNN	1,079,051	Hidden state size = 128

4.7. Statistical Analysis

To describe the missing rate of the dataset, an analysis was performed to quantify the number of measurements in the test set. The total number of measurements for each protected subgroup was calculated by multiplying the number of patients in the subgroup by the number of observation hours (48) and the number of time series variables (37). For a more granular analysis at the variable level, the total measurements for each variable within a subgroup were calculated by multiplying the number of patients in that subgroup by the observation hours.

The missing data rate for both the overall subgroup and for each specific variable was then determined by dividing the count of missing entries by the corresponding total number of measurements. For this analysis the ‘undefined’ gender subgroup was excluded as it constitutes only 0.13% of the test data.

To assess the statistical significance of the estimates, confidence intervals were constructed using the bootstrap resampling method. Ninety-five percent confidence intervals were calculated by identifying the percentiles 2.5^{th} and 97.5^{th} of $9,000$ estimates generated by resampling with repetition.

5. Results

5.1. Missing Rate per Variable and Subgroup

The descriptive analysis of the test set reveals different levels of missing rate between protected subgroups (Table 2). Within the gender subgroups, the Male subgroup has a missing data rate of 45.10%, which is substantially higher than the 34.51% rate observed in the Female subgroup. For the age subgroups, individuals older than 65 show a higher missing data rate (41.81%) than those younger than 65 (37.92%).

Table 2. Descriptive Statistics Containing Missing Rate by Protected Subgroup

Variable	Subgroup	Measurements	Missing Rate (%)
Gender	Female	1,843,488	34.51
	Male	2,411,808	45.10
Age	Age ≥ 65	2,237,760	41.81
	Age < 65	2,022,864	37.92

The variation in missing data extends to the variable level, where the missing rates differ within the same subgroup (Table 3). In the subgroup of patients aged 65 or over, the missing rate for Cholesterol was 52.42%, in contrast to the 15.22% rate for Urine. A consistent pattern was also observed in the gender group, where the Male subgroup showed a higher missing rate than the Female subgroup across every variable.

Table 3. Missing Rate by Variable and Protected Subgroup

Variable	**Subgroup**	**Measurements**	**Cholesterol** (%)	**TroponinI** (%)	**TroponinT** (%)	**AST** (%)	**Ph** (%)	**Urine** (%)
Gender	Female	49,824	43.19	43.18	42.79	42.52	38.38	13.35
	Male	65,184	56.51	56.49	55.99	55.67	49.18	17.36
Age	Age ≥ 65	60,480	52.42	52.38	51.80	51.80	46.10	15.22
	Age < 65	54,672	47.41	47.41	47.11	46.51	41.58	15.58

5.2. Imputation Disparities of MAE by Variables and Protected Subgroups

The estimated MAE from the models results shows significant variation both between clinical variables and across protected subgroups (Table 4). Evaluating the USGAN model's performance for the Female subgroup between variables, a MAE of 2.35 was observed for Cholesterol in contrast with a MAE of 0.06 for pH. Across different subgroups, within the GPVAE model, the MAE for Cholesterol was 1.47 for the Female subgroup compared to 0.79 recorded for the Male subgroup. Across all models, Cholesterol consistently registered the highest imputation errors, while pH values were imputed with the greatest accuracy, indicated by the lowest MAE values.

Table 4. 95% Confidence Intervals for the Imputation Errors by Variables for the Gender Group

Variable	**SAITS**		**BRITS**		**USGAN**		**GPVAE**		**MRNN**	
	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
Cholesterol	1.35 (0.79, 1.92)	0.83 (0.55, 1.11)	1.16 (0.49, 1.83)	0.86 (0.53, 1.19)	**2.35** **(0.55, 4.15)**	2.10 (1.22, 2.98)	**1.47** **(0.97, 1.96)**	**0.79** **(0.42, 1.15)**	1.31 (0.61, 2.01)	0.85 (0.49, 1.20)
Troponin I	0.37 (0.23, 0.51)	0.58 (0.24, 0.92)	0.52 (0.38, 0.66)	0.73 (0.31, 1.15)	1.47 (1.10, 1.84)	1.48 (0.62, 2.30)	0.64 (0.51, 0.77)	0.62 (0.48, 0.76)	0.61 (0.47, 0.74)	0.66 (0.46, 0.86)
Troponin T	0.34 (0.18, 0.50)	0.35 (0.21, 0.50)	0.35 (0.20, 0.50)	0.43 (0.28, 0.57)	0.46 (0.32, 0.60)	0.46 (0.35, 0.58)	0.45 (0.31, 0.60)	0.53 (0.38, 0.67)	0.56 (0.44, 0.69)	0.58 (0.47, 0.69)
AST	0.15 (0.06, 0.24)	0.09 (0.05, 0.13)	0.32 (0.16, 0.48)	0.12 (0.07, 0.17)	0.31 (0.21, 0.40)	0.23 (0.19, 0.27)	0.50 (0.31, 0.68)	0.27 (0.21, 0.33)	0.60 (0.44, 0.77)	0.41 (0.37, 0.44)
pH	0.04 (0.03, 0.04)	0.04 (0.04, 0.04)	0.03 (0.03, 0.03)	0.03 (0.03, 0.03)	**0.06** **(0.06, 0.07)**	0.06 (0.06, 0.07)	0.06 (0.05, 0.06)	0.05 (0.05, 0.05)	0.12 (0.12, 0.13)	0.12 (0.12, 0.13)
Urine	0.31 (0.29, 0.33)	0.37 (0.35, 0.39)	0.29 (0.27, 0.31)	0.35 (0.33, 0.37)	0.31 (0.29, 0.33)	0.36 (0.34, 0.38)	0.41 (0.39, 0.43)	0.47 (0.44, 0.49)	0.59 (0.57, 0.61)	0.63 (0.61, 0.66)

5.3. Efficiency and Fairness Trade-off in Deep Learning Models

A comparison of the imputation models highlights an inverse relationship between efficiency and fairness (Table 5). It was observed that, in terms of efficiency, the SAITS and BRITS models stand out for presenting the lowest MAE values compared to that of the other models. However, in terms of fairness, they are the most unfair, having the highest Gini coefficients. On the other hand, the MRNN model presents the lowest efficiency, evidenced by a higher MAE, but demonstrates greater fairness in relation to the other models, with the lowest Gini coefficient value.

Table 5. Trade-off between algorithmic efficiency—measured by 95% confidence intervals of Mean Absolute Error (MAE)—and algorithmic fairness—measured by 95% confidence intervals of the Gini coefficient.

Model	MAE	Gini
SAITS	**0.241 (0.237, 0.244)**	0.615 (0.611, 0.620)
BRITS	0.270 (0.262, 0.276)	0.639 (0.628, 0.649)
USGAN	0.276 (0.272, 0.280)	0.576 (0.570, 0.582)
GPVAE	0.452 (0.448, 0.456)	0.507 (0.503, 0.510)
MRNN	0.672 (0.667, 0.676)	**0.439 (0.436, 0.442)**

5.4. Characterization of Imputation Error Disparities

The distribution of imputation errors across different models and variables was visualized using Lorenz curves (Figure 1). The figure presents these curves for six variables: Cholesterol, Troponin I, Troponin T, AST, pH, and Urine, evaluated on the SAITS, BRITS, USGAN, GPVAE, and MRNN models. An aggregate Lorenz curve summarizing the overall error distribution for each of the five models is also included (Figure 1, rightmost lower plot).

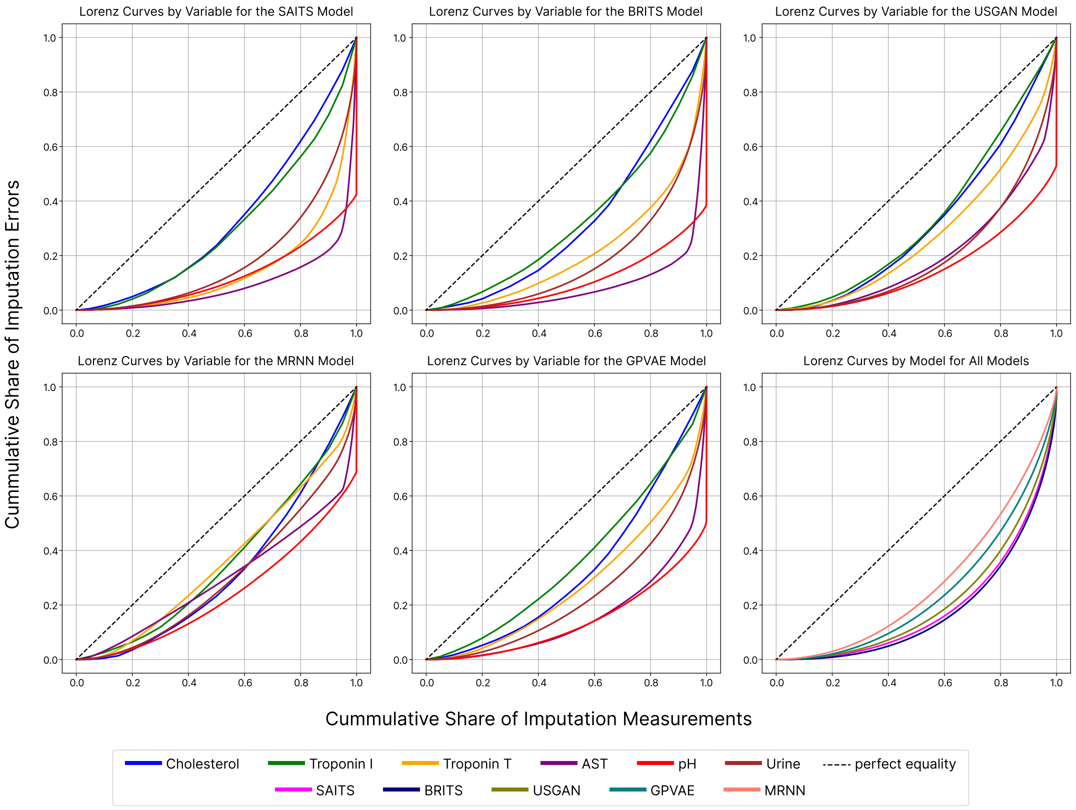

Figure 1. Lorenz curves of imputation errors for each deep learning model, stratified by variable. The bottom-right plot shows Lorenz curves for all models, stratified by model, considering all time series variables in the dataset.

6. Discussion

Our findings highlight that is essential to rigorously scrutinize AI models before deployment, particularly those intended for high-stakes applications where erroneous outputs may cause harm. In our analysis, we show that relying solely on the MAE metric can mask disparities in model performance between variables and across protected subgroups. While the global MAE metric provides a summary measure of accuracy, it fails to capture algorithmic fairness at finer levels of granularity. The SAITS, BRITS, USGAN, GPVAE, and MRNN models exhibit varying degrees of disparity when evaluated by specific variables and subgroups, as summarized by the Gini coefficient in Table 5.

Our analysis of algorithmic fairness at finer levels of granularity also reveals that imputation biases vary in direction depending on the specific variable and subgroup (Table 4). All models demonstrated lower error when imputing the Cholesterol and AST variables for the Male subgroup, suggesting a performance disparity detrimental to the Female subgroup. Conversely, for the Urine variable, all models performed better for the Female subgroup. This demonstrates that performance biases are not uniform and depend on the intersection of protected attributes and clinical variables.

Regarding the rate of missingness of a variable and its corresponding imputation error, Cholesterol, which had the highest rate of missing data, also consistently presented the highest MAE values across all models and subgroups (Table 3). While the Troponin I and Troponin T variables exhibited similarly high missingness rates, their imputation errors differed significantly in some cases. For the Female subgroup under the USGAN model, Troponin I had an MAE of 1.47, whereas Troponin T had an MAE of 0.46. This divergence suggests that imputation disparities are not driven solely by missingness rates but may also result from complex underlying data correlations. These findings may serve as a practical guide in contexts where specific variables are of primary interest and model selection must account for both accuracy and fairness in imputation performance.

The aggregate Lorenz curves for each model shows the overall algorithmic fairness of their imputation error distributions (Figure 1). The curves for the SAITS and BRITS models deviate most significantly from the line of perfect equality, indicating the greatest disparities in their imputation of missing data. In contrast, the MRNN model's curve lies closest to the line of equality, demonstrating superior algorithmic fairness compared to the other models.

A variable-level analysis of the Lorenz curves further highlights the trade-off between model accuracy and fairness (Figure 1). The pH variable, which consistently achieved the lowest MAE, also exhibited one of the most inequitable error distributions, with its Lorenz curve deviating significantly from the line of equality across all models. In contrast, the Cholesterol and Troponin I variables were imputed more equitably in the SAITS, BRITS, USGAN, and GPVAE models. Across these analyses, the MRNN model distinguished itself as the most consistent, showing less variation in fairness across different variables and generally maintaining Lorenz curves closest to the line of equality as the fairest model overall.

7. Conclusion

Missing data imputation has become a fundamental step in learning from multivariate time series, particularly in healthcare, where it can significantly impact downstream tasks related to clinical decision-making. Despite the growing adoption of deep learning models for this purpose, concerns about their algorithmic fairness in missing data imputation have largely been overlooked. Our findings, however, reveal varying degrees of fairness across models, variables, and subgroups.

While algorithmic efficiency is essential, it must be recognized that algorithmic fairness is equally important to ensure that fundamental ethical and legal principles are not compromised. Algorithmic bias—especially in sensitive contexts such as healthcare—can lead to indirect discrimination and exacerbate existing social inequalities, disproportionately affecting vulnerable groups. These effects may constitute violations of fundamental rights, including the principles of equality, non-discrimination, and human dignity. In this context, algorithmic fairness emerges not just as a technical consideration but as a normative imperative, reinforcing the need for artificial intelligence systems to uphold standards of fairness, transparency, accountability, and oversight. To this end, it becomes increasingly necessary that adopters of AI-based systems, especially in sensitive domains, implement algorithmic impact assessment tools. These assessments serve as preventive instruments to identify, monitor, and mitigate potential risks of bias and unfair outcomes, ensuring that the adoption of such technologies does not result in indirect discrimination or harm to fundamental rights. This requirement is aligned with emerging regulatory frameworks such as the European Union's Artificial Intelligence Act (AI Act), Brazil's General Personal Data Protection Law (LGPD), and OECD and UNESCO recommendations, all of which emphasize the need for ex ante impact assessments to ensure trustworthy and rights-respecting AI. They also contribute to a culture of ethical and legally compliant innovation, promoting responsible AI practices aligned with data protection regulations and human rights standards. Therefore, algorithmic imputation should be seen not merely as a technical task but as part of a broader decision-making process that demands an interdisciplinary, reflective, and legally informed approach. These findings reinforce the urgency of shifting the development and deployment of AI systems from a narrow focus on performance optimization to a comprehensive commitment to ethical, legal, and socially responsible outcomes. Future work should shift the focus toward developing models that optimize both algorithmic efficiency and fairness under broader missingness mechanisms, including not only MCAR but also MAR and MNAR.

References

Cao, W., Wang, D., Li, J., Zhou, H., Li, Y., and Li, L. (2018). Brits: bidirectional recurrent imputation for time series. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*.

Du, W., Côté, D., and Liu, Y. (2023). SAITS: Self-attention-based imputation for time series. *Expert Systems with Applications*.

Fortuin, V., Baranchuk, D., Raetsch, G., and Mandt, S. (2020). Gp-vae: Deep probabilistic time series imputation. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*.

Liu, M., Li, S., Yuan, H., Ong, M. E. H., Ning, Y., Xie, F., Saffari, S. E., Shang, Y., Volovici, V., Chakraborty, B., and Liu, N. (2023). Handling missing values in healthcare data: A systematic review of deep learning-based imputation techniques. *Artificial Intelligence in Medicine*.

Meng, C., Trinh, L., Xu, N., Enouen, J., and Liu, Y. (2022). Interpretability and fairness evaluation of deep learning models on MIMIC-IV dataset. *Scientific Reports*.

Mesquita, T. P., Silva, D. M. P. F., Ribeiro, A. M. N. C., Silva, I. R. R., Bastos-Filho, C. J. A., and Monteiro, R. P. (2024). A comparative analysis of deep learning-based methods for multivariate time series imputation with varying missing rates. In *2024 IEEE Eighth Ecuador Technical Chapters Meeting (ETCM)*.

Miao, X., Wu, Y., Wang, J., Gao, Y., Mao, X., and Yin, J. (2021). Generative semi-supervised learning for multivariate time series imputation. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Min, S., Asif, H., and Vaidya, J. (2025). Exploring the inequitable impact of data missingness on fairness in machine learning. *IEEE Intelligent Systems*, 40(3):28–38.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

Omar, M., Soffer, S., Agbareia, R., Bragazzi, N. L., Apakama, D. U., Horowitz, C. R., Charney, A. W., Freeman, R., Kummer, B., Glicksberg, B. S., Nadkarni, G. N., and Klang, E. (2024). Socio-demographic biases in medical decision-making by large language models: A large-scale multi-model analysis. *medRxiv*.

Pfohl, S. R., Cole-Lewis, H., Sayres, R., Neal, D., Asiedu, M., Dieng, A., Tomasev, N., Rashid, Q. M., Azizi, S., Rostamzadeh, N., et al. (2024). A toolbox for surfacing health equity harms and biases in large language models. *Nature Medicine*, 30(12):3590–3600.

Russell, S. (2020). *Artificial Intelligence A Modern Approach*. Pearson Series, 4 edition.

Silva, I., Moody, G., Mark, R., and Celi, L. A. (2012a). Predicting Mortality of ICU Patients: The PhysioNet/Computing in Cardiology Challenge 2012 (version 1.0.0). `https://www.physionet.org/content/challenge-2012/1.0.0/`. Accessed 23 June 2025.

Silva, I., Moody, G., Scott, D. J., Celi, L. A., and Mark, R. G. (2012b). Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012. *IEEE*.

Tipirneni, S. and Reddy, C. K. (2022). Self-Supervised Transformer for Sparse and Irregularly Sampled Multivariate Clinical Time-Series. *ACM Trans. Knowl. Discov. Data*.

Verma, S. and Rubin, J. (2018). Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, pages 1–7. ACM.

Wang, J., Du, W., Cao, W., Zhang, K., Wang, W., Liang, Y., and Wen, Q. (2024). Deep learning for multivariate time series imputation: A survey.

Wenjie, D. (2023). Pypots: A python toolbox for data mining on partially-observed time series. *arXiv preprint arXiv:2305.18811*.

Yoon, J., Zame, W. R., and Van Der Schaar, M. (2019). Estimating Missing Data in Temporal Data Streams Using Multi-Directional Recurrent Neural Networks. *IEEE Transactions on Biomedical Engineering*.