Missing Data Under Green AI Umbrella

Arthur Dantas Mangussi[1,2], **Ricardo Cardoso Pereira**[3], **Pedro Henriques Abreu** [3], **Ana Carolina Lorena**[1,2]

[1]Computer Science Division – Aeronautics Institute of Technology
– Praça Marechal Eduardo Gomes – 50 – São José dos Campos – Brazil

[2]Science and Technology Institute – Federal University of São Paulo – Talim
St. 330 – São José dos Campos – Brazil

[3]CISUC/LASI – Centre for Informatics and Systems of the University of Coimbra
– Department of Informatics Engineering – Pólo II – Pinhal de Marrocos –
Coimbra – Portugal

{arthuradm, aclorena}@ita.br, {rdpereira, pha}@dei.uc.pt

Abstract. *Missing data is a common issue that can undermine machine learning performance, and imputation methods have emerged as state-of-the-art solutions. However, training these methods can be costly and environmentally impactful. In this work, we investigate the missing data problem under Green AI constraints using a Data-Centric AI approach. We evaluate three missingness mechanisms, four missing rates, and ten datasets to assess both data quality and downstream performance. We also propose an optimization model to select the best-performing imputation method while considering sustainability constraints, offering a path toward more responsible and effective data imputation.*

1. Introduction

Data mining pipelines aim to extract valuable and actionable knowledge from raw data [Ali and Omer 2017]. Within this pipeline, the preprocessing step focuses on improving data quality rather than solely developing robust machine learning (ML) classifiers. This shift in focus has given rise to the paradigm of Data-Centric Artificial Intelligence (AI). However, real-world datasets often suffer from quality issues such as noise, imbalance, overlapping, and missing data [Clemente et al. 2023].

Missing data (MD) refers to the absence of information in one (univariate) or more (multivariate) features. It is typically categorized into three main types [Mangussi et al. 2025c, Santos et al. 2019]:

- **Missing Completely at Random (MCAR)**: The probability of missingness is independent of observed and unobserved data.
- **Missing at Random (MAR)**: The missingness depends only on the observed data.
- **Missing Not at Random (MNAR)**: The missingness is related to the unobserved (i.e., missing) data itself.

To address the missing data problem, the literature offers several strategies, such as case deletion (removing rows with missing values), missing data imputation (replacing missing entries using various techniques), and robust ML algorithms that can inherently handle missingness, such as ensemble methods [García-Laencina et al. 2010]. In this work, we focus specifically on missing data imputation methods.

According to [Hasan et al. 2021], imputation techniques can be broadly divided into statistical and ML-based methods. The simplest statistical approach is single imputation, which replaces all missing entries with a fixed value, often the mean, median, or mode. To overcome its limitations, the literature proposes multiple imputation techniques, which use approximate values drawn from a distribution to reflect the uncertainty of the missing data [Emmanuel et al. 2021]. Among these, Multivariate Imputation by Chained Equations (MICE) is one of the most widely used methods [Buuren and Groothuis-Oudshoorn 2011]. MICE employs a Gibbs sampling approach that iteratively estimates the posterior distributions of missing values using conditional models for each variable.

In contrast, machine learning algorithms such as Random Forest (RF) and k-Nearest Neighbors (kNN) have also been adopted for imputation. The RF-based method, known as missForest, applies an iterative scheme where a random forest is trained on observed data to predict missing values. Meanwhile, kNN imputes values by identifying the nearest neighbors and aggregating their values. More recently, deep learning approaches have also shown promise. Autoencoders (AEs) are neural networks trained to reconstruct their inputs, allowing them to learn from incomplete data and generate plausible imputations [Mangussi et al. 2025a]. A recent variant, the Siamese Autoencoder-Based Imputation (SAEI), has demonstrated promising results in the literature [Pereira et al. 2024a].

In recent years, ML models have become increasingly complex, often requiring a larger number of hyperparameters and computational layers to achieve state-of-the-art performance, as demonstrated across the AI literature. However, these advancements come at a cost: they demand significantly more computational resources for both training and inference. Training large-scale AI models typically requires vast amounts of data and computing power, resulting in high energy consumption, increased water usage for cooling data centers, and considerable greenhouse gas emissions [Bolón-Canedo et al. 2024].

As the environmental impact of AI continues to grow—almost exponentially in some cases—there has been a rising concern over its carbon footprint. This has led to the emergence of a new paradigm known as Green AI, which promotes sustainable practices throughout the model design, training, and deployment processes. Green AI aims to reduce environmental costs by optimizing algorithms, enhancing hardware efficiency, and adopting responsible data management practices. In this context, Data-Centric AI has emerged as a promising strategy to reduce energy consumption by improving the quality of the data used, rather than increasing model complexity [Salehi and Schmeink 2024].

Despite these developments, the literature still underexplores the intersection of missing data imputation and Data-Centric AI. More specifically, the environmental impact of imputation methods within Green AI principles remains largely unaddressed. This gap motivates our work: we propose a novel optimization-based formulation to select the best-performing imputation method while considering environmental constraints from a Data-Centric AI perspective.

Our experimental setup evaluates this approach using ten public benchmark datasets, four state-of-the-art imputation methods, and four levels of artificially induced missingness (5%, 10%, 20%, and 40%), generated under MCAR, MAR, and MNAR mechanisms in a multivariate scenario. We assess imputation quality using Mean Absolute Error (MAE) and classification effectiveness using F1-score, which are well-known metrics used into MD research field [Hasan et al. 2021]. Finally, we introduce a novel optimization model designed to balance imputation accuracy and sustainability, aligning with the principles of Green and Data-Centric AI. Our findings demonstrate that, in general, missForest and MICE were the optimum imputation methods, achieving the best trade-off between imputation accuracy and downstream classification performance. These findings support the effectiveness of missForest as a robust method aligned with the Data-Centric AI paradigm.

The remainder of this work is organized as follows: Section 2 presents related works on the MD field and Green AI; Section 3 introduces the problem addressed in this study; Section 4 describes the methodology and the experimental setup; The results are presented and discussed in Section 5; Section 6 demonstrates the real case study using healthcare data; Section 7 outlines the conclusions and future directions of this work.

2. Related Works

As previously mentioned, the interplay between missing data, Data-Centric AI, and Green AI remains underexplored in the literature, with only a few studies addressing their combined impact.

[Salehi and Schmeink 2024] conducted a comprehensive survey on Data-Centric Green Artificial Intelligence, emphasizing the importance of improving energy efficiency in AI systems. Their study highlights a shift in focus—from maximizing predictive accuracy to optimizing computational efficiency—as a way to reduce environmental costs. They argue that the convergence of Data-Centric AI and Green AI offers a promising direction for developing novel methodologies that prioritize sustainability in AI development.

In another relevant study, [Verdecchia et al. 2022] conducted empirical experiments using six different AI algorithms and a dataset containing 5,574 records. They applied two types of dataset modifications to assess the potential energy savings under a Data-Centric Green AI framework. Their findings reveal that modifying datasets can significantly reduce energy consumption, often with minimal or no loss in accuracy.

However, neither of these studies addresses the missing data problem, revealing a clear gap in the literature. In this context, our work contributes by proposing a novel optimization-based approach that explicitly balances imputation quality and classification performance under environmental constraints. To the best of our knowledge, this is the first study to bridge the concepts of Data-Centric AI, Green AI, and missing data imputation in a unified framework, thereby extending the insights provided by [Salehi and Schmeink 2024] and [Verdecchia et al. 2022].

3. Problem Description

As mentioned earlier, the missing data problem must be addressed before training any ML model, as it can critically affect classification performance if not treated properly. Moreover, with the rise of Green AI, the literature has yet to adequately investigate the carbon emissions associated with the imputation process - an area that deserves closer attention.

In this context, we investigate how to select an imputation method that achieves an optimal balance between the quality of imputation and the resulting classification performance, while limiting associated carbon emissions. The problem is formalized as follows:

$$\min \sum_{i=1}^{n} x_i[\alpha(1 - F_i) + (1 - \alpha)E_i] \tag{1}$$

$$\sum_{i=1}^{n} x_i = 1 \tag{2}$$

$$\sum_{i=1}^{n} x_i C_i \leq C_{\text{max}} \tag{3}$$

$$\sum_{i=1}^{n} x_i F_i \geq F_{\text{min}} \tag{4}$$

$$x_i \in \{0, 1\}, \quad i = 1, 2, \ldots, n \tag{5}$$

Here, F_i denotes the F1-score of a classification model when trained with the i-th imputed dataset, and E_i represents the quality of imputation for the i-th method (e.g., MAE). The parameter $\alpha \in [0, 1]$ balances both objectives. It can be adjusted to favor either imputation quality or classification performance, offering flexibility based on the researcher's goals. The constraints enforce selecting exactly one imputation method (Equation 2), ensuring its associated carbon emissions C_i do not exceed a maximum C_{max} (Equation 3), and its resulting F1-score F_i meets or exceeds a minimum threshold F_{min} (Equation 4). Finally, x_i denotes the decision variable that indicates whether a given method was selected (Equation 5).

The methodology for conducting the traditional experimental setup in the context of MD is described in the following section.

4. Methodology

The methodology employed in this work is illustrated in Figure 1. It follows the traditional experimental setup used in missing data (MD) studies, which consists of four main steps: *Data Collection, Data Amputation, Data Imputation,* and *Evaluation*. The process begins with *Data Collection*, where complete datasets (i.e., without missing values) are selected. In the *Data Amputation* step, artificial missing values are introduced according to a defined configuration, including specific missingness mechanisms and patterns (univariate or multivariate). The *Data Imputation* step then applies one or more imputation techniques to estimate these missing values. Finally, in the *Evaluation* step, the quality of imputation is assessed by comparing the imputed values with their original counterparts and/or by evaluating the classification performance of one or more classifiers when trained on the imputed data, relative to results obtained using the original complete data [Mangussi et al. 2025c, Santos et al. 2019].

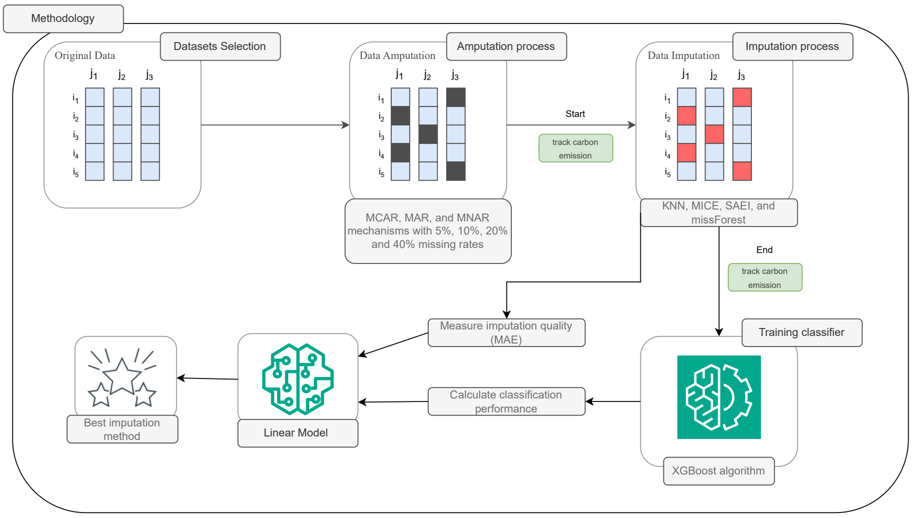

Figure 1. Illustration of the methodology applied in this work.

For the implementation of the experimental setup, we used Python version 3.11.9. All experiments were conducted on a machine equipped with an NVIDIA RTX 3060 GPU, and running Ubuntu Linux version 22.04.4. The following sections provide a more detailed description of each methodological step.

4.1. Dataset Collection

We selected 10 benchmark datasets, all of which are available from the UCI Repository[1] and/or OpenML[2]. These datasets are used for binary classification tasks and were chosen because they are complete ones (i.e., without missing values) to ensure a controlled experimental environment. An overview of datasets characteristics are summarized in Table 1.

Table 1. Overview of datasets characteristics

Dataset	Instances	Features	
		Categorical	Continuous
bc-coimbra	116	0	9
echocardiogram	61	1	8
hcv-egyptian	1385	8	20
heart-cleveland	173	8	5
mathernal-risk	1014	0	6
parkinsons	195	0	22
pima-diabetes	768	0	8
ricci	118	3	3
thoracic-surgery	470	13	3
wisconsin	569	0	30

[1] https://archive.ics.uci.edu/

[2] https://www.openml.org/

4.2. Amputation Strategy

To introduce artificial missing values (i.e., to perform the amputation process), we adopted a stratified five-fold cross-validation scheme. At each fold, the training data were first normalized to the range [0,1]. Subsequently, for each iteration, we used the Python package `mdatagen` [Mangussi et al. 2025c] to generate artificial missing data (MD) under three distinct mechanisms—MCAR, MAR, and MNAR—in a multivariate setting (i.e., with more than one feature containing missing values). These missing values were introduced independently into the training and test sets to maintain consistency across both.

In the MCAR mechanism, every data point has an equal probability of being removed, resulting in a completely random missingness pattern. In contrast, the MAR mechanism was implemented by creating pairs of correlated features, where the feature most correlated with the target variable ($\mathbf{X_{obs}}$) determines the missingness pattern in another feature ($\mathbf{X_{miss}}$), with lower $\mathbf{X_{obs}}$ values yielding higher missingness probabilities. Meanwhile, the MNAR mechanism was implemented based on the Missingness Based on Own and Unobserved Values (MBOUV) framework, randomly distributing the missing values across features for a given global missing rate. In this setup, the Missingness Based on Unobserved Values (MBUV) approach was applied to all nominal features and half of the numerical features, while the Missingness Based on Observed Values (MBOV) approach was applied to the remaining half [Pereira et al. 2024b].

For all three MD scenarios, four different levels of missingness were considered: 5%, 10%, 20%, and 40% [Mangussi et al. 2025a, Pereira et al. 2024a].

4.3. Imputation process

For the imputation process, we selected four state-of-the-art methods: k-Nearest Neighbors (kNN), Multiple Imputation by Chained Equations (MICE) [Buuren and Groothuis-Oudshoorn 2011], missForest [Stekhoven and Bühlmann 2012], and the Siamese Autoencoder-Based Imputation (SAEI) approach [Pereira et al. 2024a]. The kNN, MICE, and missForest implementations were used directly from the scikit-learn library, while SAEI was used from the author's official GitHub repository[3]. The corresponding parameter settings for each method are presented in Table 2. These parameter choices were based on prior literature, including [Mangussi et al. 2025b, Mangussi et al. 2025a, Pereira et al. 2024a].

Table 2. Parameters of each imputation method.

Imputation method	Parameters
KNN	k = 5 (neighbors), Euclidean distance metric
MICE	100 iterations, default parameters from Scikit-learn
SAEI	epochs = 200, batch size = 64, optimizer = "Adam"
missForest	criterion = "absolute_error", n_estimators=10, default parameters from Scikit-learn

The primary goal of this work is to select the best-performing imputation method considering both its imputation quality and its environmental impact, aligning with the constraints of Green AI. To this end, we measured the CO_2 emissions associated with

[3] `https://github.com/ricardodcpereira/SAEI`

training each imputer for every cross-validation fold, using the CodeCarbon Python package[4].

4.4. Evaluation Criteria

Finally, we performed the evaluation of the imputed datasets using two complementary approaches:

- **Direct Evaluation:** We measured imputation quality using the Mean Absolute Error (MAE), which quantifies the difference between the imputed values and the original, ground–truth values. This allows for a direct assessment of the imputation accuracy.
- **Indirect Evaluation:** We assessed the quality of the imputed data by its impact on a downstream classification task. To do this, we used a stratified cross-validation scheme to build complete datasets by concatenating the imputed test folds, yielding a new dataset with the original shape. An XGBoost classifier was then trained on this data, and its performance was measured using the F1-score. XGBoost was selected due to its state-of-the-art performance on tabular classification tasks, as evidenced in [Shwartz-Ziv and Armon 2022]. The XGBoost model was used with its default hyperparameters, and the stratified cross-validation splits were aligned with those used for the imputation error measurement. In addition, we monitored the CO_2 emissions associated with this training process, in line with the Green AI constraints of this work.

5. Results and Discussion

We begin by examining the overall imputation quality and classification performance, as described in Section 4, based on our experimental design. Tables 3 and 4 summarize the results of the direct and indirect evaluations, respectively. In general, MICE, kNN, and SAEI emerged as the best-performing imputation methods for MCAR, MAR, and MNAR, respectively, as shown in Table 3. Consistent with the literature, these methods achieved lower MAE values for MCAR, followed by MAR and MNAR, confirming that MNAR is the most challenging missingness mechanism.

To assess the statistical significance of the results, a Three-Way ANOVA was performed using imputation method, missing rate, and dataset as factors, with MAE as the dependent variable. Since the Anderson-Darling test with a significance level of $\alpha = 5\%$ indicated that the normality assumption was violated for all missingness mechanisms, data were rank-transformed using Ordered Quantile Normalization before applying Three-Way ANOVA on Ranks. The analysis revealed that all factors significantly affected performance, except for the imputation method under the MNAR mechanism, which was not statistically significant ($p = 0.71650$).

For a missing rate of 5%, MICE outperformed the other methods for MAR and MNAR, while missForest achieved the best performance for MCAR. At 40% missingness, the trend persisted, except that SAEI yielded the best results for MAR and MNAR. At the intermediate levels of 10% and 20%, the best-performing method varied across scenarios.

Considering the indirect evaluation (i.e., assessing the quality of imputed data via downstream classification performance), we observed that MICE, kNN, and missForest

[4] https://codecarbon.io/

Table 3. Mean Absolute Error (MAE) across all datasets, grouped by missing rate. The "Overall" column shows the average MAE across all missing rates. Highlighted results indicate the best-performing imputation methods for each missing data mechanism.

	missForest			SAEI			MICE			kNN		
Missing Rate	**MCAR**	**MAR**	**MNAR**	**MCAR**	**MAR**	**MNAR**	**MCAR**	**MAR**	**MNAR**	**MCAR**	**MAR**	**MNAR**
5	**0.125**	0.144	0.213	0.211	0.222	0.237	0.132	**0.141**	**0.210**	0.140	0.150	0.218
10	0.136	0.152	**0.214**	0.193	0.225	0.235	**0.129**	**0.143**	0.217	0.145	0.154	0.220
20	0.144	0.178	0.232	0.193	0.211	**0.218**	**0.133**	0.167	0.229	0.154	**0.165**	0.239
40	**0.158**	0.223	0.294	0.194	**0.189**	**0.179**	0.164	0.281	0.306	0.168	0.195	0.293
Overall	0.141	0.174	0.238	0.198	0.212	**0.217**	**0.139**	0.183	0.241	0.152	**0.166**	0.243

achieved the best F1-scores across our experimental design, as shown in Table 4. Notably, for the MCAR mechanism, MICE outperformed all other methods across all missing rates. In contrast, the SAEI approach generated imputations that degraded classification performance, representing the worst-case scenario. For MNAR, missForest performed best at lower missing rates (5% and 10%), while MICE was superior at higher missing rates (20% and 40%). In general, as the proportion of missing data increased, the F1-score decreased across all scenarios. We hypothesize that this trend occurs because, with higher levels of missingness, the data available for imputation no longer captures the original structure and patterns, thereby degrading classification performance. As a reference point, the F1-score for the complete dataset was 0.708, indicating that across all scenarios, the imputed datasets led to lower classification performance than the baseline.

From a Data-Centric AI perspective, one of the key findings is that MICE and kNN, which demonstrated the best imputation quality, also yielded the best classification results, confirming that better data quality translates into improved model performance. However, this trend was not observed for the MNAR mechanism, which is the most challenging scenario. Interestingly, although SAEI achieved superior imputation quality, its generated data failed to improve classification performance. We speculate that this may be due to the complexity of the autoencoder model relative to the characteristics of the dataset, suggesting a potential misalignment between the SAEI method and the underlying data structure. While methods like MICE and kNN may simplify the decision boundary, SAEI might introduce complexity that hampers classification.

Table 4. Overall F1-score across all datasets grouped by missing rate. The "Overall" column shows the average F1-score across all missing rates. Highlighted results indicate the best-performing imputation methods for each missing data mechanism.

	missForest			SAEI			MICE			kNN		
Missing Rate	**MCAR**	**MAR**	**MNAR**	**MCAR**	**MAR**	**MNAR**	**MCAR**	**MAR**	**MNAR**	**MCAR**	**MAR**	**MNAR**
5%	0.667	0.658	**0.659**	0.513	0.567	0.531	**0.669**	0.651	0.652	0.656	**0.671**	0.645
10%	0.657	0.636	**0.653**	0.422	0.501	0.478	**0.670**	**0.656**	0.650	0.657	0.651	0.634
20%	0.643	**0.640**	0.601	0.373	0.429	0.460	**0.646**	0.629	**0.637**	0.629	0.637	0.627
40%	0.624	0.597	0.613	0.341	0.427	0.424	**0.647**	**0.622**	**0.628**	0.630	0.615	0.609
Overall	0.645	0.626	**0.633**	0.395	0.445	0.451	**0.653**	0.631	**0.633**	0.639	**0.640**	0.623

Building upon the previous results, our goal is to identify the best-performing algorithm by simultaneously considering imputation quality and classification performance within a unified framework that respects Green AI constraints. This approach is aligned with the Data-Centric AI paradigm, where the focus is on selecting the algorithm that improves data quality and, consequently, enhances downstream task performance. To this end, Equations 1-4 formalize our optimization model. In this work, we balance both objectives by setting $\alpha = 0.5$ and vary the parameter $F_{\min}$ to accept baseline performance deterioration levels of 5%, 10%, 15%, 20%, and 25%. This procedure allows for a sensitivity analysis across different $F_{\min}$ thresholds. As shown in Table 5, the objective function (OF) values remained stable even as we set stricter thresholds for F1-score deterioration. Consequently, with an allowed deterioration of -15%, the solver successfully identified the optimal imputation method for each MD mechanism. Based on this result, we fixed $F_{\min}$=0.602.

We solved the optimization model independently for the MNAR, MCAR, and MAR scenarios using AMPL in Python with the HiGHS 1.10 solver. According to [Bolón-Canedo et al. 2024], two of the most popular tools for estimating the carbon footprint of ML experiments are CarbonTrack and CodeCarbon, both of which we used in this study. CarbonTrack [Lacoste et al. 2019] calculates CO_2 emissions using the formula: Power Consumption × Time × Carbon Emission Factor, which depends on the local energy grid. In our case, we assumed the South American region using an RTX 3090 GPU, with a carbon factor of 0.2 kg CO_2/kWh. Based on this, our estimation resulted in 0.07 kg CO_2. Therefore, we set $C_{\max}$=0.01166 kg CO_2, corresponding to a maximum execution time of 10 minutes.

Table 5. Objective function (OF) values and selected optimal imputation methods for each missing data (MD) mechanism, aggregated across all missing rates. Negative values in the F1-score column reflect the outcomes of the sensitivity analysis conducted with respect to the $F_{\min}$ constraint.

F1-score		MCAR		MAR		MNAR	
		OF	Optimal Imputation Method	OF	Optimal Imputation Method	OF	Optimal Imputation Method
-5%	0.673	-	-	-	-	-	-
-10%	0.638	0.243	MICE	0.263	KNN	-	-
-15%	0.602	0.243	MICE	0.263	KNN	0.3025	missForest
-20%	0.567	0.243	MICE	0.263	KNN	0.3025	missForest
-25%	0.531	0.243	MICE	0.263	KNN	0.3025	missForest

In the aggregated scenario encompassing all missing rates, our optimization model selected MICE, kNN, and missForest as the best-performing imputation methods under MCAR, MAR, and MNAR mechanisms, respectively, as shown in Table 5. To further explore how the missing rate influences the optimal selection, we present the corresponding results in Table 6.

Table 6. Objective function values (OF) and corresponding optimal imputation method for each MD mechanism grouped by missing rate.

Missing Rates	MCAR		MAR		MNAR	
	OF	Optimal Imputation Method	OF	Optimal Imputation Method	OF	Optimal Imputation Method
5%	0.2290	missForest	0.2395	KNN	0.2770	missForest
10%	0.2295	MICE	0.2435	MICE	0.2805	missForest
20%	0.2505	missForest	0.2960	MICE	0.3060	KNN
40%	0.2585	MICE	0.299	KNN	0.3405	missForest

In general, missForest and MICE demonstrated the best overall performance across different missing rates and MD mechanisms. missForest was the top-performing method for both MCAR and MNAR scenarios at the 5% missing rate. However, as the missing rate increases, the optimal selection varies depending on the mechanism and level of missingness. An interesting pattern is that missForest was never selected under the MAR mechanism, regardless of the missing rate, whereas for MNAR, it was the best-performing method at all rates except 20%. Conversely, MICE achieved the best performance at the 10% missing rate under both MCAR and MAR, but it was not selected under MNAR.

Tables 3 and 4 present the results for imputation quality and classification performance, respectively. However, relying solely on these tables makes it difficult to intuitively determine the optimal imputation method for each experimental scenario. In this context, our optimization model proves to be a valuable tool for guiding researchers and practitioners by balancing imputation quality and classification performance under Green AI constraints, in line with the Data-Centric AI perspective.

6. Real Case Study in Healthcare Data

Encouraged by the promising results obtained in the experiments from Section 5, we aimed to assess the practical applicability and generalizability of our approach in a real-world context. To this end, we conducted a case study using an open-source healthcare dataset related to COVID-19, enabling us to investigate the behavior of our optimization framework under realistic conditions and domain-relevant constraints. The dataset comprises 19,000 observations and 17 features, including 16 categorical variables indicating the presence or absence of specific symptoms and one numerical variable representing patient age. The target variable is binary, indicating whether a patient's condition worsened due to COVID-19.

For this dataset, we applied the amputation process under MCAR, MAR, and MNAR mechanisms, introducing up to 10% missingness to simulate realistic scenarios where data are incomplete. We then evaluated the four imputation methods within this context. We also use the parameter $F_{\min}$ = 0.583, which represent a 10% of degradation in F1-score for baseline (i.e., the original data) accepting $C_{\max}$ = 0.01166 CO_2kg/kwh. The goal was to assess the effectiveness of our optimization approach from a Data-Centric AI perspective, subject to Green AI constraints in real-world healthcare data.

After applying our optimization model in this case study, our findings demonstrate that MICE was the best-performing imputation method for MCAR, MAR, and MNAR mechanisms, as seen in Table 7.

Table 7. Best-performing imputation method for healthcare case study. The column "OF" outlines the result of Equation 1.

Mechanism	OF	Best Imputation Method
MCAR	0.2594	MICE
MAR	0.2381	MICE
MNAR	0.3222	MICE

7. Conclusion

In this work, we proposed a novel optimization formulation to select the best-performing imputation method by balancing imputation quality and classification performance under environmental constraints. Our experimental setup involved 10 public benchmark datasets, four state-of-the-art imputation methods, and missing data generated under MCAR, MAR, and MNAR mechanisms with missing rates of 5%, 10%, 20%, and 40%. The results indicate that, overall, missForest and MICE were the most frequently selected methods, offering the best trade-off between imputation accuracy and downstream classification performance. These findings highlight the effectiveness of both methods as robust approaches aligned with the principles of the Data-Centric AI paradigm.

Additionally, we applied our optimization model to a real-world healthcare dataset. Using the same methodology, the results contrasted with the benchmark experiments: MICE was identified as the optimal method across all missing data mechanisms. This case study reinforces the importance of context-aware method selection and demonstrates how our approach can assist practitioners in choosing imputation strategies under Green AI constraints.

As a continuation of this research, we plan to extend our methodology by exploring the use of Large Language Models (LLMs) for missing data imputation. Although recent studies have shown that LLMs can achieve high performance in various tasks, they also come with a significant carbon footprint. For example, training GPT-3, with 175 billion parameters, is estimated to produce approximately 1,216,952 lbs of CO_2 emissions [Salehi and Schmeink 2024]. This raises an important research question: *Can LLM-based imputation methods offer competitive performance while still respecting Green AI constraints?*

A promising direction for future work is to extend the proposed methodology to other data modalities and complex tasks, including multiclass, multilabel, and domain-diverse datasets. This would allow broader evaluation under sustainability constraints and may require adapting the objective function, particularly when applying the F1 constraint in more complex scenarios.

Acknowledgements

This study was financed, in part, by the São Paulo Research Foundation (FAPESP), Brasil. Process Numbers 2021/06870-3 and 2024/23791-8. This work was also financed through national funds by FCT - Fundação para a Ciência e a Tecnologia, I.P., in the framework of the Project UIDB/00326/2025 and UIDP/00326/2025. Additionally, it was supported by the Portuguese Recovery and Resilience Plan (PRR) through project C645008882-00000055-Center for Responsable AI.

References

Ali, N. A. and Omer, Z. M. (2017). Improving accuracy of missing data imputation in data mining. *Kurdistan Journal of Applied Research*, pages 66–73.

Bolón-Canedo, V., Morán-Fernández, L., Cancela, B., and Alonso-Betanzos, A. (2024). A review of green artificial intelligence: Towards a more sustainable future. *Neurocomputing*, 599:128096.

Buuren, S. and Groothuis-Oudshoorn, C. (2011). Mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67.

Clemente, F., Ribeiro, G. M., Quemy, A., Santos, M. S., Pereira, R. C., and Barros, A. (2023). ydata-profiling: Accelerating data-centric ai with high-quality data. *Neurocomputing*, 554:126585.

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., and Tabona, O. (2021). A survey on missing data in machine learning. *Journal of Big Data*, 8(1).

García-Laencina, P. J., Sancho-Gómez, J.-L., and Figueiras-Vidal, A. R. (2010). Pattern classification with missing data: a review. *Neural Computing and Applications*, 19(2):263–282.

Hasan, M. K., Alam, M. A., Roy, S., Dutta, A., Jawad, M. T., and Das, S. (2021). Missing value imputation affects the performance of machine learning: A review and analysis of the literature (2010–2021). *Informatics in Medicine Unlocked*, 27:100799.

Lacoste, A., Luccioni, A., Schmidt, V., and Dandres, T. (2019). Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Mangussi, A. D., Pereira, R. C., Abreu, P. H., and Lorena, A. C. (2025a). Assessing adversarial effects of noise in missing data imputation. In Paes, A. and Verri, F. A. N., editors, *Intelligent Systems*, pages 200–214, Cham. Springer Nature Switzerland.

Mangussi, A. D., Pereira, R. C., Lorena, A. C., Santos, M. S., and Abreu, P. H. (2025b). Studying the robustness of data imputation methodologies against adversarial attacks. *Computers Security*, 157:104574.

Mangussi, A. D., Santos, M. S., Lopes, F. L., Pereira, R. C., Lorena, A. C., and Abreu, P. H. (2025c). mdatagen: A python library for the artificial generation of missing data. *Neurocomputing*, 625:129478.

Pereira, R. C., Abreu, P. H., and Rodrigues, P. P. (2024a). Siamese autoencoder architecture for the imputation of data missing not at random. *Journal of Computational Science*, 78:102269.

Pereira, R. C., Abreu, P. H., Rodrigues, P. P., and Figueiredo, M. A. (2024b). Imputation of data missing not at random: Artificial generation and benchmark analysis. *Expert Systems with Applications*, 249:123654.

Salehi, S. and Schmeink, A. (2024). Data-centric green artificial intelligence: A survey. *IEEE Transactions on Artificial Intelligence*, 5(5):1973–1989.

Santos, M. S., Pereira, R. C., Costa, A. F., Soares, J. P., Santos, J., and Abreu, P. H. (2019). Generating synthetic missing data: A review by missing mechanism. *IEEE Access*, 7:11651–11667.

Shwartz-Ziv, R. and Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90.

Stekhoven, D. and Bühlmann, P. (2012). Missforest?non-parametric missing value imputation for mixed-type data. *Bioinformatics (Oxford, England)*, 28:112–8.

Verdecchia, R., Cruz, L., Sallou, J., Lin, M., Wickenden, J., and Hotellier, E. (2022). Data-Centric Green AI An Exploratory Empirical Study . In *2022 International Conference on ICT for Sustainability (ICT4S)*, pages 35–45, Los Alamitos, CA, USA. IEEE Computer Society.