A Data-Centric Approach to Missing Data Imputation: Addressing Noise, Adversarial, and Fairness Challenges

Arthur Dantas Mangussi[1,2], **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
Avenue Cesare Monsueto Giulio Lattes – 1201 – São José dos Campos – Brazil

[3]University of Coimbra – 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, pha@dei.uc.pt

Abstract. *Real-world datasets frequently suffer from quality issues, all of which can jeopardize machine learning models performance. Aligned with the Data-Centric AI paradigm, this work focuses on missing data, investigating its interaction with other data quality challenges rather than proposing new imputation methods. Specifically, the study examines how missing data behaves in the presence of noise, adversarial attacks, and fairness-related concerns. The findings reveal that such interactions notably influence imputation quality error, predictive performance, and fairness outcomes. These insights emphasize the importance of considering broader data quality factors when addressing missing data. Additionally, this research contributes to a novel Python package designed to generate missing values under various realistic settings. This tool facilitates reproducible experiments and enables more equitable benchmarking of imputation strategies, supporting future research in missing data and data-centric evaluations.*

Resumo. *Conjuntos de dados do mundo real frequentemente apresentam problemas de qualidade, os quais podem comprometer o desempenho de modelos de aprendizado de máquina. Alinhado ao paradigma de Data-Centric AI, este trabalho foca em dados ausentes, investigando sua interação com outros problemas de qualidade de dados, em vez de propor novos métodos de imputação. Especificamente, o estudo analisa como os dados faltantes se comportam na presença de ruído, ataques adversariais e questões relacionadas à fairness. Os resultados mostram que essas interações influenciam significativamente o erro na qualidade da imputação, o desempenho preditivo e os resultados de equidade. Esses achados reforçam a importância de considerar fatores mais amplos de qualidade de dados ao lidar com dados faltantes. Além disso, esta pesquisa contribui com um novo pacote em Python desenvolvido para gerar valores ausentes em diferentes cenários realistas. Essa ferramenta possibilita experimentos reprodutíveis e promove uma comparação mais justa entre estratégias de imputação, apoiando pesquisas futuras em dados faltantes e avaliações centradas nos dados.*

1. Introduction

Data mining techniques aim to uncover hidden patterns in data, extracting valuable information and transforming it into practical knowledge [Yu et al. 2013]. However, most real-world datasets used for learning often contain various data irregularities, such as imbalanced classes, overlapping instances, noise, redundant or irrelevant features, and missing values [Clemente et al. 2023]. These data quality issues can impair the generalization performance of classifiers trained on such data [Nakhaei et al. 2023].

To mitigate these problems, data preprocessing emerged as a crucial step in the Machine Learning (ML) pipeline, with the goal of improving data quality and, consequently, enhancing classifier performance in downstream tasks. Recently, there has been a growing emphasis on constructing high-quality datasets and systematically refining data to boost the effectiveness of AI systems- a movement known as *Data-Centric* AI. Studies in this area have shown that improving data quality can significantly enhance classification performance [Clemente et al. 2023].

Most ML models used for classification tasks do not handle missing data (MD) directly within the data preprocessing step, requiring this issue to be addressed beforehand [Pereira et al. 2022]. Missing data can be described as the absence of information in one or more features within a dataset [Santos et al. 2019]. The literature originally categorizes MD by their cause into three mechanisms [Rubin 1976]: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR) [Mangussi et al. 2025b]:

- **MCAR**: Missing values occur entirely at random, meaning the causes of missingness are unrelated to any observed data in the dataset;
- **MAR**: Missingness is related to the observed data, indicating a dependency between the data features and the causes of missingness;
- **MNAR**: Missingness depends on both observed and unobserved data (e.g., features not present in the dataset). Consequently, the causes of missingness are unknown.

To address MD problem, the literature proposes several strategies. According to [García-Laencina et al. 2010], four main approaches exist for pattern classification with missing data: Case Deletion, Missing Data Imputation, Model-Based Procedures, and use of Machine Learning methods that are robust to missing values. This work focuses on missing data imputation, which aims to replace missing entries with estimated values based on a predefined strategy [García-Laencina et al. 2010], and it is the state-of-the-art strategy to deal with such a problem. Imputation methods range from simple techniques, such as replacing values with the mean, median, or mode, to more advanced models, including Multiple Imputation by Chained Equations (MICE)[Buuren and Groothuis-Oudshoorn 2011], Partial Multiple Imputation with Variational Autoencoders (PMIVAE)[Pereira et al. 2022], and Generative Adversarial Imputation Nets (GAIN) [Yoon et al. 2018].

In the MD literature, most research has focused on developing new imputation algorithms, with quality assessed via standard metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Mean Absolute Percentage Error (MAPE). Some studies also consider downstream classification metrics such as accuracy and F1-score [Hasan et al. 2021, Lin and Tsai 2020], aligning with a model-centric AI approach.

However, the interaction between missing data and broader data quality challenges remains underexplored. The combined impact of missing values and other data imperfections, such as noise, adversarial perturbations, and algorithmic bias, has received limited attention. Therefore, this work adopts a data-centric perspective on missing data imputation, aiming to investigate these intersections. The main research question addressed is: *What is the interaction between missing data and other data quality inconsistencies, and to what extent do these issues impact the performance of data imputation strategies?*

Previous work has shown that the quality of observed data significantly impacts the imputation process and must be carefully addressed. Moreover, there are notable gaps in the literature regarding data quality issues in missing data environments. This work aims to fill these gaps through the following contributions:

- **Work 1**: Although previous studies have explored the use of noise-robust algorithms for imputation, they do not examine whether applying a simple preprocessing step, such as a noise filter, prior to imputation affects the results. A comprehensive exploratory analysis on the use of noise filters (NFs) before imputation is still lacking. To the best of our knowledge, our work is the first study to systematically investigate the direct and indirect effects of noise filtering on imputation quality;
- **Work 2**: To date, no study has explored the robustness of imputation strategies under adversarial settings. This work extends adversarial machine learning research by incorporating a Data-Centric AI perspective, focusing on the intersection of adversarial attacks and missing data quality;
- **Works 3 and 4**: There is a notable gap in the literature concerning the impact of missing data imputation strategies on system fairness, particularly within fairness-aware models applied to multivariate scenarios. The use of GAN-based imputation methods remains underexplored in fairness-aware machine learning. Moreover, few studies consider scenarios involving both MAR and MNAR mechanisms across multiple features, especially under high missingness rates. Recent imputation techniques like Autoencoders and robust classifiers like XGBoost are also seldom evaluated in these contexts. To the best of our knowledge, this work is the first to comprehensively address both MAR and MNAR mechanisms in multivariate settings with high missing rates, using widely adopted classifiers such as XGBoost and Random Forest. Furthermore, we present the first in-depth analysis demonstrating that integrating imputation strategies with fairness-aware models promotes more equitable outcomes under challenging data conditions.

The reminder of this document is organized as follows: Section 2 gives an overview of the methodology applied in this work, while Section 3 presents the overall results and conclusions about this research. Moreover, it highlights the achievements and academic impact of this work.

2. Methodology

Most experimental studies in this domain follow the setup described by [Santos et al. 2019], which includes four stages: (i) Data Collection, where complete datasets (i.e., without missing values) are selected; (ii) Data Amputation, where missing values are artificially introduced under specific configurations (i.e., defined

mechanisms and patterns); (iii) Data Imputation, where one or more techniques are applied to estimate the missing values; and (iv) Evaluation, which assesses imputation quality by comparing the estimated values to the original ones and/or by evaluating the performance of classifiers trained on the imputed data against those trained on the complete data.

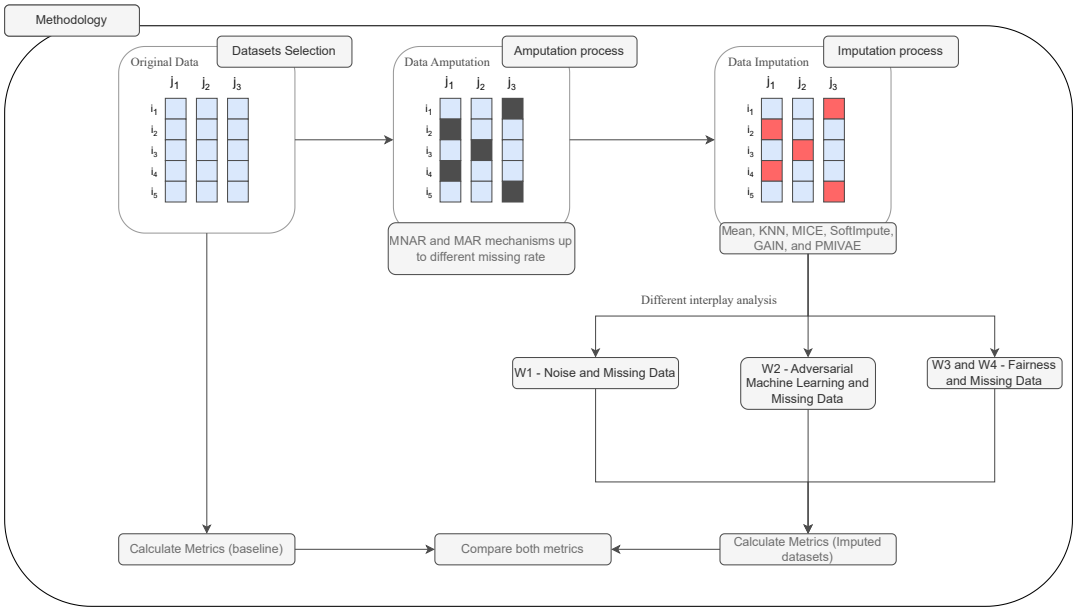

Figure 1. Overview of the methodology of this work.

An overview of the methodology applied in this work is shown in Figure 1. We followed the main steps typically adopted in traditional experimental setups within the MD field [Santos et al. 2019]. However, our work focuses on addressing noise, adversarial, and fairness challenges within a missing data environment. Furthermore, we conducted independent studies for each intersection between a data quality issue and missing data, as outlined by the achievements in Section 4. In the *Data Collection* step, we selected open-source original and complete datasets from UCI repository and/or OpenML. These datasets vary across the different studies developed; however, they all consist of binary classification tabular data and serve as baselines to compare and evaluate our research hypotheses. Next, in the *Data Amputation* step, we applied missingness using MNAR and MAR mechanisms in a multivariate scenario, with varying missing rates (5%, 10%, 20%, 40%, and 60%). This step was performed using the `mdatagen` package [Mangussi et al. 2025b], developed in this work. Then, in the *Data Imputation* step, we evaluated various imputation strategies: SimpleImputer (imputation by the mean for numerical features and mode for the categorical ones), kNN, MICE [Buuren and Groothuis-Oudshoorn 2011], missForest [Stekhoven and Bühlmann 2012], PMIVAE [Pereira et al. 2022], SoftImpute [Hastie et al. 2015], and GAIN [Yoon et al. 2018]. SimpleImputer, kNN, missForest, and MICE were used directly from the Scikit-learn library [Pedregosa et al. 2011]. The remaining algorithms are available in different GitHub repositories[1]. The specific

[1]`https://github.com/travisbrady/py-soft-impute`,

parametrization is presented in each article published/submitted.

To assess the performance of the imputation process, we used MAE to measure imputation quality. Additionally, we employed standard classification metrics- accuracy and F1-score- to evaluate each methodology in a downstream task. Since our goal was to understand the impact of each data quality issue, we compared these metrics against the baseline (original and complete datasets) and after applying our methodology. The specific methodology of each study and the most significant results are presented in the following section.

3. Results and Discussion

In this section, we present the most significant results of our work. We begin by highlighting the importance of the artificial generation of missing data step for MD studies and introduce the first Python package specifically developed to support this task. Subsequently, each subsection addresses a specific intersection between missing data and other data quality issues.

3.1. `mdatagen` package

When researching missing data, studies typically follow two main avenues: (i) developing new methods and comparing them with state-of-the-art approaches; or (ii) conducting extensive empirical studies to characterize the performance of existing methods. However, the existing literature often fails to enable fair and reproducible comparisons between imputation techniques designed for missing data scenarios. To address this gap, we developed the `mdatagen` Python package, which offers a comprehensive suite of state-of-the-art implementations for simulating missing data mechanisms, covering MCAR, MAR, and MNAR. The package is designed to enhance accessibility and usability for researchers and data scientists in the field. Furthermore, `mdatagen` supports seamless integration with Python-based ML pipelines, ensuring reproducibility and consistency in benchmark studies. This library is already used across all the following experimental designs for artificial generation of missing values step in the following sections.

3.2. Noise Filter and Data Imputation

We employed the Edited Nearest Neighbor (ENN) algorithm with k=5 to filter potentially noisy instances from the training sets. ENN cleans the dataset by removing samples whose class label differs from the majority of their k nearest neighbors, typically eliminating data points located in overlapping, borderline, or noisy regions of the feature space [Lemaître et al. 2017].

As previously mentioned, our goal was to evaluate whether applying a simple noise-cleaning preprocessing step before imputation could improve imputation quality. To this end, we applied ENN and validated its effectiveness on both synthetic and real-world datasets. Our findings demonstrate that noise negatively affects the performance of imputation methods, which is in line with the existing literature. However, our results also show that even a simple noise filter like ENN generally improves imputation quality across both dataset types, highlighting the advantage of applying noise preprocessing

`https://github.com/jsyoon0823/GAIN`
`https://github.com/ricardodcpereira/PMIVAE`

Table 1. Average MAE results obtained for each imputation method, grouped by missing rate. The highlighted bold results present the best MAE results for each missing rate.

Missing Rate	Mean	KNN	MICE	PMIVAE	SoftImpute	GAIN	missForest
5%	0.233	0.142	0.145	0.221	0.168	0.481	**0.139**
10%	0.224	0.146	0.147	0.209	0.164	0.442	**0.130**
20%	0.211	0.158	0.148	0.196	0.162	0.467	**0.132**

before imputation. Additionally, we observed a correlation between better imputation quality and improved classification performance when using MICE, kNN, and missForest, the top three performing imputation methods in our experiments as shown in Figure 2. This supports the core principle of Data-Centric AI: better data leads to better outcomes.

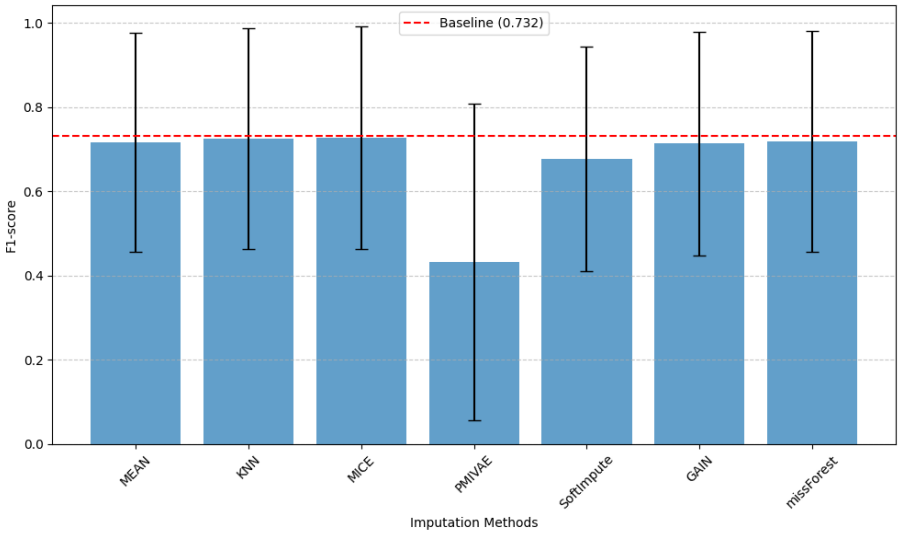

Figure 2. Overall F1-scores of all imputation methods using the Random Forest classification model, following our methodology where the ENN algorithm is applied before the imputation process. The baseline represents the F1-score obtained on the original (complete) datasets.

3.3. Robustness of Data Imputation Methodologies Against Adversarial Attacks

Adversarial Machine Learning (AML) is a field focused on studying cybersecurity attacks designed to degrade the performance of classifiers on specific tasks [Martins et al. 2020]. However, most prior research has concentrated on the impact of such attacks on classifiers, while a Data-Centric perspective remains underexplored. Our findings reveal that adversarial attack strategies also significantly affect the imputation process (Table 2). Notably, adversarial attacks influenced datasets with MAR mechanisms more than those with MNAR. To assess how adversarial attacks affect data distributions during imputation, we applied the Kolmogorov-Smirnov test for numerical features and the Chi-square test for categorical features. Results show that numerical features exhibited statistically significant differences compared to the baseline, while categorical features did not demonstrate substantial distributional changes.

Table 2. Average MAE across all datasets and missing rates for both missing data mechanisms, grouped by imputation strategies. The Baseline represents the imputation task without any adversarial attack. The (↑) and (↓) indicate an increase and decrease, respectively, compared to the baseline. It is important to note that higher (↑) MAE values indicate worse performance, as lower values are closer to the ideal (zero) [Mangussi et al. 2025a].

Adversarial attack	Imputation methods	Baseline (without attack)			Results (with attack)		
		MAR	MNAR	MCAR	MAR	MNAR	MCAR
FGSM	kNN	0.186	0.333	0.171	0.288 (↑ 55%)	0.387 (↑ 16%)	0.271 (↑ 58%)
	MICE	0.208	0.295	0.151	0.371 (↑ 78%)	0.378 (↑ 28%)	0.261 (↑ 74%)
	SoftImpute	0.247	0.243	0.224	0.320 (↑ 29%)	0.300 (↑ 24%)	0.292 (↑ 31%)
	GAIN	0.421	0.511	0.284	0.472 (↑ 12%)	**0.471 (↓ 8%)**	0.363 (↑ 28%)
PGD	kNN	0.186	0.333	0.171	0.189 (↑ 2%)	**0.316 (↓ 5%)**	0.176 (↑ 3%)
	MICE	0.208	0.295	0.151	0.213 (↑ 2%)	**0.289 (↓ 2%)**	0.155 (↑ 3%)
	SoftImpute	0.247	0.243	0.224	0.249 (↑ 1%)	0.245 (↑ 1%)	0.223 (0%)
	GAIN	0.421	0.511	0.284	**0.385 (↓ 9%)**	**0.483 (↓ 5%)**	**0.267 (↓ 6%)**
C&W	kNN	0.186	0.333	0.171	0.197 (↑ 6%)	**0.312 (↓ 6%)**	0.208 (↑ 22%)
	MICE	0.208	0.295	0.151	0.229 (↑ 10%)	0.295 (0%)	0.189 (↑ 26%)
	SoftImpute	0.247	0.243	0.224	0.262 (↑ 6%)	0.256 (↑ 5%)	0.253 (↑ 13%)
	GAIN	0.421	0.511	0.284	**0.391 (↓ 7%)**	**0.402 (↓ 21%)**	0.301 (↑ 6%)
POISON	kNN	0.186	0.333	0.171	0.308 (↑ 66%)	**0.319 (↓ 4%)**	–
	MICE	0.208	0.295	0.151	0.302 (↑ 45%)	0.388 (↑ 32%)	–
	SoftImpute	0.247	0.243	0.224	0.365 (↑ 48%)	0.244 (0%)	**0.210 (↓ 6%)**
	GAIN	0.421	0.511	0.284	**0.406 (↓ 4%)**	0.517 (↑ 1%)	–

3.4. Influence of Data Imputation in Group Fairness Metrics

Concerning missing data and fairness, our findings indicate that the selected imputation strategy, missing rate, missing data mechanism, and classifier used significantly influence system fairness as shown in Figure 3. However, we observed that improvements in fairness often come at the cost of classification performance.

Notably, MICE and MEAN were the only imputation methods in our experiments that maintained stable classification performance, with less than a 10% decrease in F1-score. Among them, MICE consistently outperformed MEAN regarding imputation quality, making it the best-case scenario overall. On the other hand, PMIVAE, likely due to its complexity and the nature of the datasets, achieved near-perfect fairness scores but suffered from poor classification performance. Given these results, we also propose a decision diagram to assist practitioners. Since missing rates of 10% and 20% are common in real-world scenarios, this diagram considers fairness, classification performance, and imputation quality to help researchers select the most suitable method based on their specific optimization goals. Based on these results, we also used Fairness-aware Machine Learning models to investigate the relationship between fairness and missing data. Our findings suggest that combining imputation strategies with fairness-aware models fosters a more equitable decision-making process. However, further analysis is needed, as we evaluated only three classification models, each exhibiting structurally different behaviors.

4. Conclusions

In this work, we investigated the interaction between missing data and other data quality issues, namely noise, fairness, and adversarial attacks, from a data-centric perspective. Missing data is a well-known challenge in real-world datasets and poses significant dif-

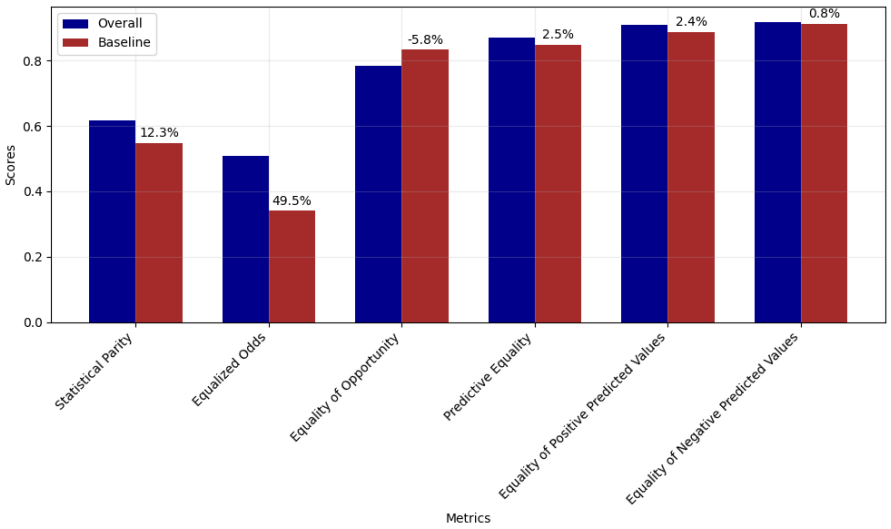

Figure 3. Comparison of fairness metrics and model performance between overall and baseline. Overall is the average fairness metrics across all datasets, missing rates, missing data mechanisms, and classifiers from our experimental setup. Percentage represent the improvement of over the baseline.

ficulties for machine learning applications. Data imputation algorithms are widely used to address this issue; however, the literature reveals significant limitations that this work aims to overcome.

Our findings show that these issues influence the performance of imputation methods, as alterations to the observed data can propagate through the imputation process. Moreover, the effectiveness of each algorithm can vary depending on the scenario. For instance, MICE demonstrated the best overall performance in terms of fairness. In noisy data settings, methods like missForest, kNN, and MICE benefited from the application of a preprocessing step, whereas this step negatively impacted SoftImpute, though SoftImpute excelled under adversarial attack conditions.

Beyond answering the main research question, we also introduced the first Python library dedicated to data amputation for missing data studies. The `mdatagen` package enables the generation of artificial missing values under configurable and realistic conditions, facilitating reproducible experiments and comprehensive benchmarking in the field of missing data.

4.1. Academic Impact

As a result of this research, the following studies have been published and/or submitted:

1. Arthur Dantas Mangussi, Ricardo Cardoso Pereira, Pedro Henriques Abreu, and Ana Carolina Lorena. Assessing Adversarial Effects of Noise in Missing Data Imputation. In *34th Brazilian Conference on Intelligent Systems 2024 (BRACIS 2024)*, Cham: Springer Nature Switzerland, 2024. p. 200-214 **(Qualis - A4)**;
2. Arthur Dantas Mangussi, Miriam Seoane Santos, Filipe Loyola Lopes, Ricardo Cardoso Pereira, Ana Carolina Lorena, and Pedro Henriques Abreu. mdatagen: A Python Library for the Artificial Generation of Missing Data. **Neurocomputing**, v. 625, p. 129478, 2025 **(Scimago - Q1)**;

3. Arthur Dantas Mangussi, Miriam Seoane Santos, Ricardo Cardoso Pereira, Ana Carolina Lorena, and Pedro Henriques Abreu. Studying the robustness of data imputation methodologies against adversarial attacks. **Computers & Security**, v. 157, p. 104574, 2025 **(Scimago - Q1)**;
4. Arthur Dantas Mangussi, Ricardo Cardoso Pereira, Miriam Seoane Santos, Ana Carolina Lorena, Mykola Pechenizkiy, Pedro Henriques Abreu, Exploring the Influence of Missing Data Imputation in Group Fairness Metrics. **Artificial Intelligence**, p.104559, 2026 **(Scimago - Q1)**;
5. Arthur Dantas Mangussi, Ricardo Cardoso Pereira, Ana Carolina Lorena, Pedro Henriques Abreu. Impact of Missing Data Imputation in Fairness-Aware Machine Learning *IEEE Transactions on Artificial Intelligence (submitted on February 24th, 2025)***(Scimago - Q1)**.

The Python package developed in this work already demonstrates its relevance within the missing data research community, the article has 8 citations in Google Scholar and the package has 20k downloads from PyPI at the date. Moreover, we expect our package to continue to be adopted in future studies, leading to additional publications indirectly related to this work.

4.2. Future Work

With the rise of Generative AI and Large Language Models, there is a growing trend to investigate whether these models can effectively perform the imputation process. Therefore, in addition to the analyses focused on imputation tasks, a key area for future research is to explore the relationship between data quality issues and how these issues interact with Large Language Models from a Data-Centric AI perspective. This work primarily focuses on tabular data, however adding the temporal component in the analysis could be another important research area to be explored in the near future.

Acknowledgements

This study was financed, in part, by the São Paulo Research Foundation (FAPESP), Brasil. Process Numbers 2021/06870-3 and 2024/23791-8.

References

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.

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.

Hastie, T., Mazumder, R., Lee, J. D., and Zadeh, R. (2015). Matrix completion and low-rank svd via fast alternating least squares. *J. Mach. Learn. Res.*, 16(1):3367–3402.

Lemaître, G., Nogueira, F., and Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5.

Lin, W.-C. and Tsai, C.-F. (2020). Missing value imputation: a review and analysis of the literature (2006–2017). *Artificial Intelligence Review*, 53:1487–1509.

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

Martins, N., Cruz, J., Cruz, T., and Henriques Abreu, P. (2020). Adversarial machine learning applied to intrusion and malware scenarios: A systematic review. *IEEE Access*, PP:1–1.

Nakhaei, A., Sepehri, M. M., and khatibi, t. (2023). A promising method for correcting class noise in the presence of attribute noise. *International Journal of Hospital Research*, 12(1):–.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pereira, R. C., Abreu, P. H., and Rodrigues, P. P. (2022). Partial multiple imputation with variational autoencoders: tackling not at randomness in healthcare data. *IEEE Journal of Biomedical and Health Informatics*, 26(8):4218–4227.

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

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.

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

Yoon, J., Jordon, J., and van der Schaar, M. (2018). Gain: Missing data imputation using generative adversarial nets. In *International Conference on Machine Learning (ICML)*, pages 5689—-5698.

Yu, Z., Fung, B., and Haghighat, F. (2013). Extracting knowledge from building-related data — a data mining framework. *Building Simulation*, 6:207–222.