Assessing Adversarial Effects of Noise in Missing Data Imputation

Arthur Dantas Mangussi, Ricardo Cardoso Pereira, Pedro Henriques Abreu, and Ana Carolina Lorena

Aeronautics Institute of Technology (ITA), Brazil
Federal University of São Paulo (UNIFESP), Brazil
University of Coimbra (UC), Portugal

34th Brazilian Conference on Intelligent Systems (BRACIS)
November 20, 2024

Introduction - Missing Data

- In real-world scenarios, a wide variety of data contain inconsistencies, such as missing values or noise;
- These inconsistencies can jeopardize the performance of Machine Learning (ML) classifiers [1];
- Missing Data (MD) refers to the absence of information in one or more variables within a dataset;
- The literature categorizes MD into three mechanisms [2]:
  - Missing Completely at Random (MCAR);
  - Missing At Random (MAR);
  - Missing Not At Random (MNAR).

Introduction - Missing Data Imputation

- To address the MD issue, the literature presents several Missing Value Imputation (MVI) strategies:
  - Mean/mode imputation;
  - Multiple imputation, as Multivariate Imputation by Chained Equation (MICE);
  - Matrix completion methods, as SoftImpute;
  - ML and Deep learning (DL) algorithms.

Introduction - Classical Experimental Setup

- In the field of MD, the classical experimental setup consists of four main steps [2]:
  - **Data Collection**: Acquiring datasets without missing values;
  - **Amputation**: Introducing artificial MD under MCAR, MAR, and/or MNAR mechanisms;
  - **Imputation**: Selecting and applying MVI strategies;
  - **Evaluation**: Assessing the imputation task. Evaluation can be categorized into two types:
    - ⋆ **Direct**: Measuring the difference between the original and imputed data;
    - ⋆ **Indirect**: Assessing the classification performance on datasets with imputed values.
- Existing literature rarely emphasizes the evaluation of MVI strategies using both direct and indirect methods [3, 4].

Introduction - Noise Data

- Noisy data is another type of data quality issue;
- The literature divided noise data into two distinct types [5]:
  - Attribute noise;
  - Class/label noise.
- Label noise is potentially more harmful than attribute noise [6, 5].

Introduction - Noise Data

- Literature outlines several methodologies for identifying potential noise and addressing it [7]:
  - Algorithms robust to noise;
  - Noise Filters (NFs);
- **Our focus in this work is on NFs**.
  - They are widely used in data pre-processing to cleanse training data [8];
  - NFs can be categorized into two types:
    - ⋆ Similarity-based filters;
    - ⋆ Ensemble-based filters.

Motivation - Related Works

The interplay of missing and noisy data has been examined in the literature through two key scenarios:

Motivation - Related Works

The interplay of missing and noisy data has been examined in the literature through two key scenarios:

1. Developing methods to predict missing values in noisy environments, such as the algorithms proposed by Zhu, He, and Liatsis [9] and Li et al. [10], which introduced, respectively:
   - Robust Imputation based on the Group Method of Data Handling (RIBG);
   - Noise-Aware Missing Data Multiple Imputation (NPMI).

Motivation - Related Works

The interplay of missing and noisy data has been examined in the literature through two key scenarios:

1. Developing methods to predict missing values in noisy environments, such as the algorithms proposed by Zhu, He, and Liatsis [9] and Li et al. [10], which introduced, respectively:
   - Robust Imputation based on the Group Method of Data Handling (RIBG);
   - Noise-Aware Missing Data Multiple Imputation (NPMI).
2. Evaluating the impact of noise on the imputation process:
   - Van Hulse and Khoshgoftaar [11] analyzes the influence of noise in the specific context of software measurement data.

Motivation - Addressing the Literature Gap

- Literature shows:
  - Quality of observed data significantly impacts the imputation task and needs to be addressed;
  - One option is to propose algorithms that are robust to noise.
- However, those works do not analyze if a simple pre-processing for treating noisy instances beforehand impacts MD imputation;
- Overall, a comprehensive exploratory analysis of NF use before the imputation task is still needed.

Objectives

Main Goal

This work aims to investigate the interaction of missing and noisy data inconsistencies and to what extent noisy data impacts the results of MVI methods.

Filling the Literature Gap

To our knowledge, this is the first study to apply a noise filter prior to imputation, examining both the direct and indirect effects of this approach on imputation quality.

Methodology: Generation of Synthetic Data

In the first step of our methodology:

- Synthetic data generation using Scikit-learn;
- Dataset configuration: 500 observations, 5 numerical input features, and two classes;
- Artificial noise was added to:
  - **Features**: Adding noise based on each feature's mean and standard deviation, ensuring values stay within the feature's minimum and maximum range;
  - **Labels**: Applying label noise by flipping class labels;
- Noise levels were introduced at 5%, 10%, and 20%.

Methodology - Flowchart for Synthetic Data

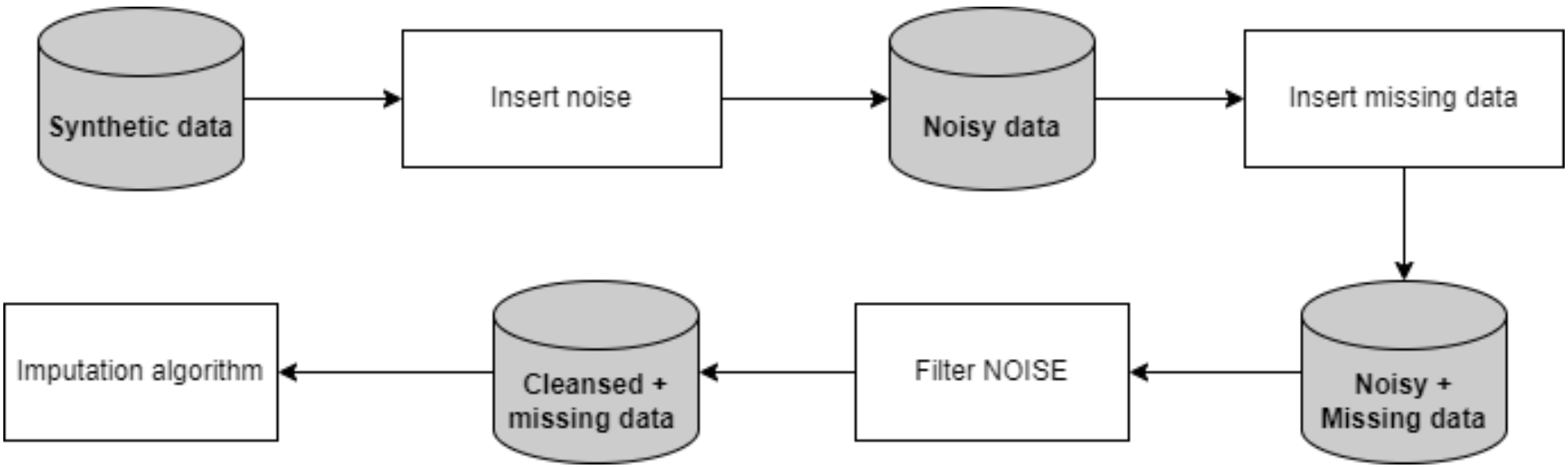

Figure: Methodology overview illustrating the sequential steps followed in the experimental design in the case of synthetic datasets.

Methodology - Overview of Experimental Setup

After inserting noise, we used the classical experimental setup for MD studies:

- **Cross-validation:** 5-fold stratified cross-validation;
- **Amputation process:** Conducted with the Python package `mdatagen`;

Methodology - Overview of Experimental Setup

After inserting noise, we used the classical experimental setup for MD studies:

- **Cross-validation:** 5-fold stratified cross-validation;
- **Amputation process:** Conducted with the Python package `mdatagen`;
- **MD Mechanism:** Artificial MNAR multivariate mechanism, removing the lowest values of features most associated with class labels;
- **Missing rates**: 5%, 10%, and 20%;

Methodology - Overview of Experimental Setup

After inserting noise, we used the classical experimental setup for MD studies:

- **Cross-validation:** 5-fold stratified cross-validation;
- **Amputation process:** Conducted with the Python package `mdatagen`;
- **MD Mechanism:** Artificial MNAR multivariate mechanism, removing the lowest values of features most associated with class labels;
- **Missing rates**: 5%, 10%, and 20%;
- **Imputation methods:** Mean Imputation, kNN, MICE, PMIVAE, missForest, SoftImpute, and GAIN;

Methodology - Overview of Experimental Setup

After inserting noise, we used the classical experimental setup for MD studies:

- **Cross-validation:** 5-fold stratified cross-validation;
- **Amputation process:** Conducted with the Python package `mdatagen`;
- **MD Mechanism:** Artificial MNAR multivariate mechanism, removing the lowest values of features most associated with class labels;
- **Missing rates**: 5%, 10%, and 20%;
- **Imputation methods:** Mean Imputation, kNN, MICE, PMIVAE, missForest, SoftImpute, and GAIN;
- **Noise Filter (NF):** Edited Nearest Neighbor (ENN) algorithm used to filter potential noise in the training sets, with $k = 5$;

Methodology - Overview of Experimental Setup

After inserting noise, we used the classical experimental setup for MD studies:

- **Cross-validation:** 5-fold stratified cross-validation;
- **Amputation process:** Conducted with the Python package `mdatagen`;
- **MD Mechanism:** Artificial MNAR multivariate mechanism, removing the lowest values of features most associated with class labels;
- **Missing rates**: 5%, 10%, and 20%;
- **Imputation methods:** Mean Imputation, kNN, MICE, PMIVAE, missForest, SoftImpute, and GAIN;
- **Noise Filter (NF):** Edited Nearest Neighbor (ENN) algorithm used to filter potential noise in the training sets, with $k = 5$;
- **Performance Evaluation:** Classification performance of new datasets measured by F1-score using Random Forest.

Methodology - Real-world Data Collection

Table: Overview of datasets characteristics.

Dataset	Instances	Features		Classes
		Continuous	Categorical	
Wiscosin	569	30	0	2
Pima diabetes	768	8	0	2
Indian liver	583	9	1	2
Parkinsons	195	22	0	2
Mammographic masses	830	1	3	2
Thoracic surgery	470	3	13	2
Diabetic retinopathy	1151	3	16	2
BC Coimbra	116	4	0	2
Thyroid recurrence	383	1	15	2
Blood transfusion	748	4	0	2
Law school	20798	6	6	2

Limitation of Dataset Selection

- Binary classification;
- Split each dataset by class;
- Ensure both classes have equal amounts of missing values;
- Prevent the noise filter from removing an entire class.

Methodology - Flowchart for Real-world Data

- We assumed that real-world datasets already contains some noise level;
- We used the same experimental setup for synthetic and real-world data.

Results: Direct Evaluation on Synthetic Data

- **Is there an impact of NF on imputation strategies?**
  - Applying NF generally improves MAE results in imputation;
  - SoftImpute is an exception, as NF does not significantly improve its performance;
  - Even a simple noise filter like ENN can enhance imputation quality.

Results: Direct Evaluation on Synthetic Data

- **Is there an impact of NF on imputation strategies?**
  - Applying NF generally improves MAE results in imputation;
  - SoftImpute is an exception, as NF does not significantly improve its performance;
  - Even a simple noise filter like ENN can enhance imputation quality.
- **Does the missing rate influence the effect of NF?**
  - Mean, MICE, PMIVAE, and missForest show greater improvements with higher missing rates.

Results: Direct Evaluation on Synthetic Data

- **Is there an impact of NF on imputation strategies?**
  - Applying NF generally improves MAE results in imputation;
  - SoftImpute is an exception, as NF does not significantly improve its performance;
  - Even a simple noise filter like ENN can enhance imputation quality.
- **Does the missing rate influence the effect of NF?**
  - Mean, MICE, PMIVAE, and missForest show greater improvements with higher missing rates.
- **Is there a specific noise type that influences the imputation task more significantly?**
  - No clear trend is observed regarding which noise type or ratio most affects MVI results.

Results: Evaluation on Real-world Data

Table: 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**

- The missForest algorithm outperforms the remaining methods, followed by KNN and MICE;
- As observed with synthetic data, ENN applied as a pre-processing step leads to more significant improvements as the missing rate increases;

Results: Evaluation on Real-world Data

- The applied methodology generally enhances imputation quality for most MVI techniques under the MNAR mechanism;
- Taking the three best imputation methods, missForest, KNN, and MICE, respectively, we also analyze the average difference between applying or not the methodology of this work, for different missing levels.

Table: Average differences of MAE grouped by missing rate for the three best-performing imputation methods. The differences represent the results without applying and after applying NF before imputation. Therefore, **positive** values indicate **better estimates** of the missing values after applying NF.

Missing Rate	**KNN**	**MICE**	**missForest**
5%	0.010	0.005	0.007
10%	0.022	0.006	0.014
20%	0.037	0.024	0.024

Results: Indirect evaluation

- We observe a correlation between better imputation quality and improved classification performance when using MICE, KNN, and missForest. A higher F1-score indicates better performance.

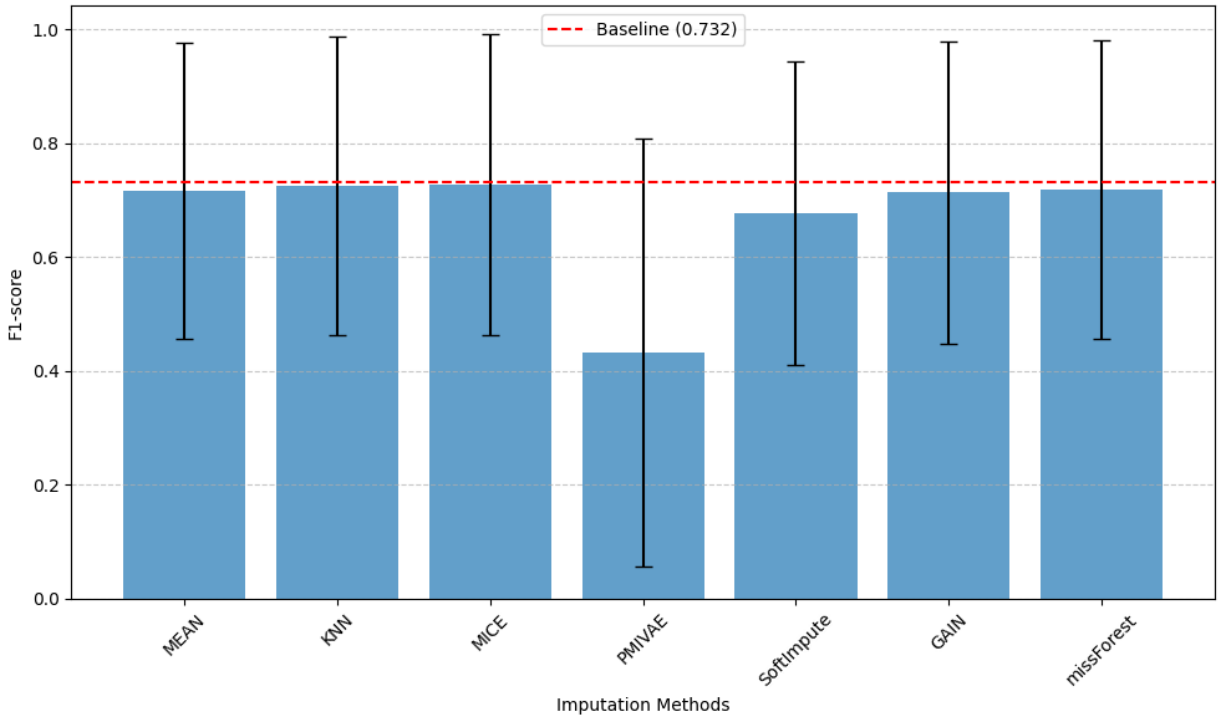

Figure: Overall F1 score of all imputation methods for Random Forest classification model.

Conclusions

- Applying the noise filter generally improved the accuracy of missing value estimates for both synthetic and real-world datasets;
- Our methodology benefited missForest, kNN, and MICE the most;
- Improvements were more pronounced for datasets with higher missing rates.

Conclusions

- Applying the noise filter generally improved the accuracy of missing value estimates for both synthetic and real-world datasets;
- Our methodology benefited missForest, kNN, and MICE the most;
- Improvements were more pronounced for datasets with higher missing rates.

Future Considerations

Extend our methodology to include:

- Larger datasets;
- MAR multivariate mechanisms;
- Additional classifiers, such as XGBoost and LightGBM;
- More complex noise filters beyond ENN.

References I

[1] Akram Nakhaei, Mohammad Mehdi Sepehri, and toktam khatibi. "A Promising Method for Correcting Class Noise in the Presence of Attribute Noise". In: *International Journal of Hospital Research* 12.1 (2023), pp. –. ISSN: 2251-8940. DOI: LBL_COMMENTED_AT/ijhr.2023.383118.1535. eprint: https://ijhr.iums.ac.ir/article_171438_ac81b13a039c851269a736c53ee3c542.pdf. URL: https://ijhr.iums.ac.ir/article_171438.html.

[2] Miriam Seoane Santos et al. "Generating synthetic missing data: A review by missing mechanism". In: *IEEE Access* 7 (2019), pp. 11651–11667.

[3] Wei-Chao Lin and Chih-Fong Tsai. "Missing value imputation: a review and analysis of the literature (2006–2017)". In: *Artificial Intelligence Review* 53 (Feb. 2020), pp. 1487–1509. DOI: 10.1007/s10462-019-09709-4.

[4] Md. Kamrul Hasan et al. "Missing value imputation affects the performance of machine learning: A review and analysis of the literature (2010–2021)". In: *Informatics in Medicine Unlocked* 27 (Nov. 2021). DOI: 10.1016/j.imu.2021.100799.

References II

[5] Xingquan Zhu and Xindong Wu. "Class Noise vs. Attribute Noise: A Quantitative Study". In: *Artif. Intell. Rev.* 22 (Nov. 2004), pp. 177–210. DOI: 10.1007/s10462-004-0751-8.

[6] José A. Sáez et al. "Analyzing the presence of noise in multi-class problems: Alleviating its influence with the One-vs-One decomposition". In: *Knowledge and Information Systems* 38 (Jan. 2014), pp. 179–206. DOI: 10.1007/s10115-012-0570-1.

[7] Benoit Frenay and Michel Verleysen. "Classification in the Presence of Label Noise: A Survey". In: *IEEE Transactions on Neural Networks and Learning Systems* 25.5 (2014), pp. 845–869. DOI: 10.1109/TNNLS.2013.2292894.

[8] José A. Sáez. "Noise Models in Classification: Unified Nomenclature, Extended Taxonomy and Pragmatic Categorization". In: *Mathematics* 10.20 (2022). ISSN: 2227-7390. DOI: 10.3390/math10203736. URL: https://www.mdpi.com/2227-7390/10/20/3736.

References III

[9] Bing Zhu, Changzheng He, and Panos Liatsis. "A robust missing value imputation method for noisy data". In: *Appl. Intell.* 36 (Jan. 2012), pp. 61–74. DOI: 10.1007/s10489-010-0244-1.

[10] Fangfang Li et al. "A Noise-Aware Multiple Imputation Algorithm for Missing Data". In: *Mathematics* 11 (Dec. 2022), p. 73. DOI: 10.3390/math11010073.

[11] Jason Van Hulse and Taghi Khoshgoftaar. "A comprehensive empirical evaluation of missing value imputation in noisy software measurement data". In: *Journal of Systems and Software* 81 (May 2008), pp. 691–708. DOI: 10.1016/j.jss.2007.07.043.

Acknowledgments

We gratefully acknowledge the Brazilian funding agencies FAPESP (Fundação Amparo à Pesquisa do Estado de São Paulo) under grants 2022/10553-6, 2023/13688-2 and 2021/06870-3. This research was also supported by the Portuguese Recovery and Resilience Plan (PRR) through project C645008882-00000055 - Center for Responsable AI.

Thank you for your attention!

contact: mangussiarthur@gmail.com

LinkedIn:

GitHub: