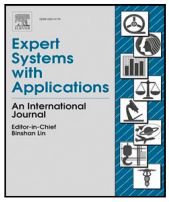
Expert Systems With Applications 249 (2024) 123654

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Imputation of data Missing Not at Random: Artificial generation and benchmark analysis

Ricardo Cardoso Pereira [a,*], Pedro Henriques Abreu [a], Pedro Pereira Rodrigues [b], Mário A.T. Figueiredo [c,d]

[a] University of Coimbra, Centre for Informatics and Systems of the University of Coimbra, Department of Informatics Engineering, 3030-290 Coimbra, Portugal
[b] University of Porto, Center for Health Technology and Services Research, Faculty of Medicine (MEDCIDS), 4200-319 Porto, Portugal
[c] Instituto Superior Técnico, University of Lisbon, 1049-001 Lisbon, Portugal
[d] Instituto de Telecomunicações, 1049-001 Lisbon, Portugal

ARTICLE INFO

Keywords:
Missing data
Missing Not at Random
Imputation
Artificial generation
Benchmark analysis

ABSTRACT

Experimental assessment of different missing data imputation methods often compute error rates between the original values and the estimated ones. This experimental setup relies on complete datasets that are injected with missing values. The injection process is straightforward for the Missing Completely At Random and Missing At Random mechanisms; however, the Missing Not At Random mechanism poses a major challenge, since the available artificial generation strategies are limited. Furthermore, the studies focused on this latter mechanism tend to disregard a comprehensive baseline of state-of-the-art imputation methods. In this work, both challenges are addressed: four new Missing Not At Random generation strategies are introduced and a benchmark study is conducted to compare six imputation methods in an experimental setup that covers 10 datasets and five missingness levels (10% to 80%). The overall findings are that, for most missing rates and datasets, the best imputation method to deal with Missing Not At Random values is the Multiple Imputation by Chained Equations, whereas for higher missingness rates autoencoders show promising results.

1. Introduction

Missing data appears in many real-world domains and can have a major impact in most data science tasks. For example, several machine learning classifiers cannot be trained on datasets with missing values, and those that can show a significant accuracy degradation (Pereira et al., 2020a). Adopting a case deletion strategy is typically a bad solution since it may cause the dataset to become biased and valuable data could be lost. Therefore, one of the most common approaches to deal with missing data is to perform imputation, which consists in replacing the missing values with plausible estimates thereof. However, the imputation procedure can be affected differently by the nature of the missing data, which can be caused by three types of mechanisms (Rubin, 1976). One type of mechanisms is known as Missing Completely At Random (MCAR), since there is no dependency between the missingness causes and the underlying values. MCAR mechanisms satisfy

$$Pr(R = 0|Y, \psi) = Pr(R = 0|\psi) \tag{1}$$

where R is the binary matrix indicating which values are missing, Y is the underlying complete data, and ψ are the parameters of the missing data model (Van Buuren, 2018). Another class of mechanisms is called Missing At Random (MAR), where the probability of missing data is constant only within specific observed groups (e.g., specific values of a feature). In other words, the missing values are related to some part of the underlying values. MAR mechanisms obey

$$Pr(R = 0|Y, \psi) = Pr(R = 0|Y_{\text{obs}}, \psi) \tag{2}$$

since they depend on the observed fraction of the data, Y_{obs}. Finally, if neither the MCAR nor the MAR assumptions hold, the mechanism is in the Missing Not At Random (MNAR) class. In other words, the etiology of missing data is unknown and, therefore, may depend on the unobserved data. In fact, the MNAR mechanism can have two different settings: the missingness may depend on the missing values themselves or on other features that are not available in the dataset.

To achieve the best possible imputation results, it is vital to select the most suitable method according to the data characteristics and domain. This decision often requires an experimental study in

* Corresponding author.
E-mail addresses: rdpereira@dei.uc.pt (R.C. Pereira), pha@dei.uc.pt (P.H. Abreu), pprodrigues@med.up.pt (P.P. Rodrigues), mario.figueiredo@tecnico.ulisboa.pt (M.A.T. Figueiredo).

https://doi.org/10.1016/j.eswa.2024.123654
Received 17 July 2022; Received in revised form 12 February 2024; Accepted 8 March 2024
Available online 15 March 2024

which several methods are used to estimate the missing values. The method yielding the lowest value of some estimation error is then selected (Pereira et al., 2020a). However, to calculate this error, the dataset must be complete (i.e., it cannot have missing values) since ground truth values are required. Therefore, the complete dataset must be artificially injected with missing data under some specific mechanism. This process must accurately mimic the assumed missingness mechanism, otherwise the obtained results will be biased. This is easy to achieve for the MCAR and MAR mechanisms, but the existing MNAR approaches are rather limited and they are mostly applicable to numeric features, mimicking scenarios where the missingness of each value only depends on itself (Santos et al., 2019). Such limitations become a major issue given that MNAR is the most prevalent mechanism in several critical contexts, such as healthcare (Peek & Rodrigues, 2018).

In this work, we address these limitations by introducing four novel MNAR strategies, which mimic real-world scenarios that follow the assumptions behind this type of mechanism. Overall, the proposed strategies can be used with numeric and categorical features (including nominal ones), and are able to yield missing values related to themselves and to other unavailable/unobserved data. To the best of the authors' knowledge, this is the first proposal of MNAR artificial generation strategies with such realistic characteristics.

Furthermore, a benchmark study is conducted to compare the imputation error obtained by a comprehensive baseline of six state-of-the-art methods. The main goal is to evaluate how these methods perform when dealing with the different MNAR characteristics introduced by the proposed generation strategies. The conclusions obtained from this benchmark study help to understand which imputation methods are more suitable for the different MNAR settings, leading to a set of guidelines that can be used by other researches and practitioners.

The experimental setup uses 10 public medical datasets, chosen due to frequently suffering from missing data under MNAR assumptions (Peek & Rodrigues, 2018). Also, since all datasets are public, the study can be easily replicated. The setup considers five levels of missingness (10%, 20%, 40%, 60%, and 80%) and the results are evaluated individually for numeric, categorical, and mixed data types, using the mean absolute error (MAE) metric. Finally, a detailed discussion about the conclusions of the experiment is presented, with a set of guidelines as a take-home message.

To summarize, this work has two main contributions:

- New strategies to artificially generate MNAR values, which are essential to properly conduct missing data studies with this mechanism. These new strategies replicate real-world scenarios of MNAR;
- A benchmark study comparing different state-of-the-art imputation methods when used to estimate the missing values generated according to the new proposed strategies.

The remainder of the paper is organized as follows: Section 2 presents related work about generation and imputation of missing values under MNAR assumptions; Section 3 introduces the new artificial MNAR strategies; Section 4 describes the experimental setup used in the benchmark study; Section 5 presents the experimental results; finally, Section 6 discusses the results and the conclusions extracted from them.

2. Related work

The most common strategy to deal with MNAR missing data is to use a resilient imputation method and perform sensitivity analysis with it. For example, if a feature has MNAR missing values, it is reasonable to assume that its distribution based on the available data is shifted. Sensitivity analysis would allow a clearer understanding of the shift magnitude, and how much it can vary, without compromising the imputation (Austin et al., 2020). However, this procedure requires previous knowledge about the features' distribution and domain, which frequently makes it unfeasible to apply. In these cases, the solution tends to be to use an imputation procedure that is more resilient to the MNAR assumptions. The most common one is multiple imputation, where the idea is to impute the data several times while varying different parameters to better account for uncertainty. The imputation results are then individually analyzed and combined towards a final result. The best-known method adopting this procedure is the Multiple Imputation by Chained Equations (MICE) (White et al., 2011). More recently, autoencoders (AE) have been used for missing data imputation under MNAR, particularly its denoising (DAE) and variational (VAE) versions (Pereira et al., 2020b). Since DAEs and VAEs are, respectively, more resilient to noise in the data and able to generate new estimates based on a learned distribution, they present valuable characteristics to address the MNAR scenario. Nevertheless, other imputation methods not so suitable for MNAR are sometimes included in the experimental baseline, such as k-Nearest Neighbors (kNN), matrix completion approaches (e.g., singular value decomposition (SVD)), and the common mean/mode imputation. The following works report results for most of these imputation methods.

Beaulieu-Jones et al. (2018) presented a study where electronic health records (EHR) are injected with missing values under three types of mechanisms and also through real data modeling, and different imputation approaches are used: random sample, mean, median, kNN, SVD, SoftImpute, and several MICE variants. They used a dataset of patient records (e.g., age, sex, body mass index, and laboratory measures) from the Geisinger Health System in Pennsylvania. The results showed that the lowest root mean squared error (RMSE) was obtained with the MICE variant that uses predictive mean matching with five nearest neighbors. However, these results are consistently worse for MNAR: when comparing with MAR and MCAR, in some cases the RMSE values are double that of MNAR.

Pereira et al. (2019) proposed an approach to use different information sources from the same healthcare context to improve the imputation quality and classification performance for datasets with MNAR missing data. The strategy divides a complete dataset into N smaller subsets, each being combined gradually with the remaining ones. Because of the unique features that each dataset has, the combination of two or more datasets generates missing values under MNAR that are imputed through five methods: mean imputation, MICE, kNN, SoftImpute and support vector regression (SVR). The experiment was conducted with 15 public medical datasets, and the results were evaluated through the RMSE metric. The results showed that the MICE algorithm outperformed the remaining methods.

Qiu et al. (2020) proposed the use of VAEs to perform the imputation of MNAR missing values in genomic data. The authors proposed an extension to the vanilla VAE where a shift correction is applied to the decoder output in order to account for the probable distribution bias caused by the MNAR mechanism. The experiment was conducted with two datasets, namely the pan-cancer RNA sequencing data from the Cancer Genome Atlas and DNA methylation data, and different levels of missingness were considered: 5%, 10%, 30%, and 50%. The VAE and its extension were compared with the kNN and SVD methods, with the results being reported through the RMSE metric. While the VAE and the SVD presented similar results, the VAE with shift correction outperformed the remaining imputation methods for all missingness rates.

Boquet et al. (2019, 2020) applied a two-layer neural network for regression with the Caltrans Performance Measurement System (PeMS) dataset. The data contains missing values (10%, 20%, and 40% rates), which were imputed with a VAE, a vanilla AE, and the PCA method. The obtained results showed that the VAE outperformed the remaining methods, achieving RMSE improvements of at least 40% for MNAR.

Beaulieu-Jones and Moore (2017) presented a study where EHR with missing values are imputed with a DAE; the results were compared with iterative SVD, kNN, SoftImpute, and mean/median imputation. The experiment was conducted with the public dataset ALS Pooled

Resource Open-Access Clinical Trials (PRO-ACT). The results, evaluated through the RMSE metric over missingness rates between 10% and 50%, showed that the DAE obtains the best results for the MNAR mechanism, although very close to those obtained by kNN and SoftImpute imputation.

Gondara and Wang (2017) compared DAEs for missing data imputation with MICE, using accuracy to evaluate binary imputation and RMSE to assess time series imputation. Their experiment considered ten simulated and four real-world datasets, mostly from the healthcare context. The results showed that the DAE consistently outperforms MICE in all datasets, for both metrics, sometimes achieving improvements over 20%. The work only considers the MCAR and MNAR mechanisms, and the imputation error is consistently better for MNAR in the simulated datasets. However, with the real-word data, the results were mixed for both mechanisms. The authors extended the experiment in a latter work (Gondara & Wang, 2018), exploring the use of multiple imputation with DAEs and considering two different scenarios: all the features have missing values (uniform synthetic generation); only half of the features are set to be missing (random synthetic generation). The experiment considered 15 real-world datasets from different contexts, and the results showed that the DAE outperforms MICE for all the uniform scenarios and in seven cases for the random scenario (five under the MNAR mechanism).

Bruni et al. (2021) proposed the concept of donor imputation where a complete observation is chosen to pair with each record containing missing values. The imputation is then performed by replacing the missing values with the data available in the paired complete observations (hence the donor term). The authors suggest that the method works for both categorical and numeric features, while being valid for all missing data mechanisms (including MNAR). The reported tests focused on the European Tertiary Education Register database, which contains several characteristics of higher education institutions in the European Union. The results show that reconstructed values are very similar to the original data.

Regarding the artificial generation of MNAR values, the existing approaches are rather limited (Santos et al., 2019). The most common approach is to remove the lowest values of a selected feature up to a predefined missing rate (Twala, 2009). Alternatively, the highest values of the feature can be removed instead (Xia et al., 2017). Since this approach is univariate (i.e., the missing values depend on a single feature), the feature most correlated with the class labels is often chosen (Twala, 2009). This strategy poses a major limitation: since it finds the lowest or highest values of the feature, only numeric and ordinal data types are supported, not nominal features. For multivariate scenarios, Garciarena and Santana (2017) proposed an extension of the described univariate approach to work on several features, called Missingness depending on its Value Itself (MIV). The strategy is essentially the same: the lowest values are removed, but now for several previously selected features. The missingness rate is global for all features, being therefore divided by the selected ones. The same authors also proposed another mechanism called Missing depending on unobserved Variables (MuOv), where the generated missing values are related to a feature not included in the dataset. The approach randomly selects N patterns to be missing up to a certain rate (since the causative features are unknown), and the values from these patterns are removed. Twala (2009) used the same approach, but the features are aggregated in pairs or triplets and only one feature from each set is injected with missing values. Ali and Omer (2017) proposed a different multivariate strategy where each feature is split into two groups: one with all values no larger than the feature's median and another with the remaining values. One of the groups is then randomly chosen to have values removed in an amount equal to twice the missing rate, therefore respecting the quotas needed to achieve the global missing rate defined for the dataset. Zhu et al. (2012) and Pan et al. (2015) followed the same strategy for numeric and ordinal features, but extended it to nominal data by dividing the categories into two equally-sized groups. Since these approaches work only with half of the observations (considering the split into two groups), the global missing rate is limited to 50%.

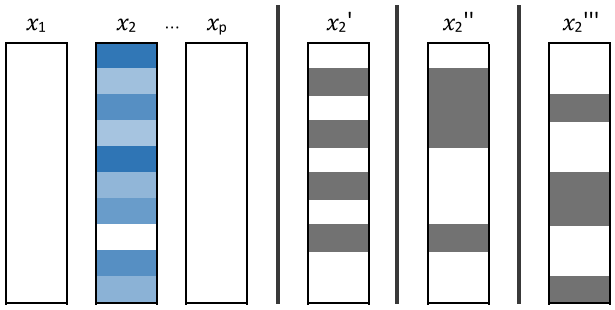


Fig. 1. Example of the MBOV variants with a 40% missing rate. Gray instances of x_2 are missing values based on lower-values removal with $p = 0$ (x_2') and $p = 0.25$ (x_2'') values under MCAR, and middlemost-values removal (x_2'''). The magnitude of x_2 values is represented by a gradient of the blue color.

3. Artificial generation of MNAR

In most missing data experimental studies, the imputation quality is assessed via some error measure between the estimated values and the real ones. As a consequence, such experiments require access to the ground truth, which means all datasets must be complete and missing values must be artificially generated. The procedure used for this generation needs to obey the characteristics of the mechanism under study. As previously discussed, the existing artificial MNAR generation methods are rather limited and are not realistic for this mechanism. Therefore, four new strategies are introduced in this work: three are entirely novel (Sections 3.2, 3.3, and 3.4), and the remaining one comprises several new variants (Section 3.1).

3.1. Missingness Based on Own Values (MBOV)

Missingness Based on Own Values (MBOV) is the most common strategy for the MNAR mechanism. As the name implies, MBOV assumes that the missingness probability of a value depends on itself. In this strategy, the lower values of a selected feature are removed to achieve a certain missing rate (Twala, 2009). Alternatively, the higher values can be chosen instead (Xia et al., 2017). Because they lack a notion of order (see Section 2), MBOV cannot be applied to nominal features.

Besides the default strategy, two variants are introduced in this work to better mimic real-world scenarios. The first variant adds a stochastic component controlled by a parameter p that represents the fraction of the missing values to be randomly chosen, therefore following the MCAR mechanism. The remaining missing values are generated under MNAR assumptions, with the lower values being removed. The idea is to add a certain amount of randomness to an otherwise deterministic MNAR mechanism. The parameter p controls the level of randomness: for $p = 1$, the generation is entirely random, which means the missing values would all be under the MCAR mechanism; on the other hand, for $p = 0$, the generation follows the default deterministic MNAR behavior. Therefore, to have most of the missing values under MNAR assumptions, this value should satisfy $0 < p < 0.5$. In the second variant, the values of a selected feature which are close to the center of the distribution are removed instead of the lowest or highest values. The goal is to simulate scenarios where extreme values are more often reported, while the most common values are more commonly kept out. As an example, in an healthcare context, physicians tend to report values considered to be out of the expected domain, but often disregard values within the normal range. An example of these variants is available in Fig. 1.

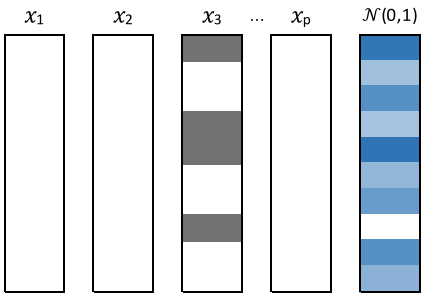


Fig. 2. Example of the MBUV strategy with a 40% missing rate. Gray instances of x_3 are missing values, which are based on the edge values of the new random continuous feature (rightmost one). The magnitude of its values is represented by a gradient of the blue color.

3.2. Missingness Based on Unobserved Values (MBUV)

Missingness Based on Unobserved Values (MBUV) is based on a common MAR generation strategy, where the missing values are related with an observed feature of the dataset. To transpose this strategy for a MNAR context, this relation must be established with an unobserved feature. Therefore, a new random numeric feature is generated, following a normal distribution, and its extreme instances (i.e., those with the lowest and highest values) are set to be missing values on the target observed feature. The approach is similar to the Missing depending on unobserved Variables (MuOv) from Garciarena and Santana (2017), with two main differences: MuOv provides multivariate generation by following the described procedure for several features, while MBUV is univariate (although it can be easily extended for multivariate scenarios, as seen in Section 3.4); MuOv uses a random process to choose the missing instances while MBUV divides the desired missing rate in half and selects the edge instances upon this threshold. This strategy is applicable to any data type (including nominal features), since it only requires the unobserved feature to be ordered. An example of this strategy is available in Fig. 2.

3.3. Missingness Based on Intra-Relation (MBIR)

Missingness Based on Intra-Relation (MBIR) is a novel strategy that aims to identify a likely MAR relation and transform it to MNAR. The approach uses two features: the one that will have the missing values (f_{miss}) and another one used to establish the MAR relation (f_{obs}). The f_{miss} feature is selected a priori and can be of any type, but f_{obs} is selected using an iterative process. The instances containing the lowest values of each numeric and ordinal feature in the dataset (with the exception of f_{miss}) are set to be missing for the f_{miss} feature, achieving a certain missing rate. An auxiliary missing indicator binary feature (f_{ind}) is created for f_{miss}, being set to 1 for the instances where the values are missing. The differences between the values of f_{obs} for both groups ($f_{\text{ind}} = 1$ and $f_{\text{ind}} = 0$) are then assessed through statistical tests and, if found to be significant, the generated missingness is likely to be MAR. From all the iterations that meet the previous criteria, the one with the most significant differences is chosen and the respective f_{obs} feature is deleted. As stated above, the goal is to transform a likely MAR scenario into MNAR, which is achieved by deleting f_{obs} since the relation found with an existent feature of the dataset becomes a relation with an unobserved feature. Pseudo code of the described MBIR procedure is presented in Algorithm 1.

The MBIR strategy relies on statistical tests to identify scenarios that are likely to be MAR. In order to make the identification more robust, two different statistical approaches are considered: a frequentist and a Bayesian one. Both approaches are for two-sample scenarios and evaluate whether they come from significantly different distributions. The frequentist approach is the Mann–Whitney rank test (Lin et al., 2021), used to evaluate if the f_{obs} samples for $f_{\text{ind}} = 1$ and $f_{\text{ind}} = 0$ are sufficiently different with a statistical significance level of $\alpha = 0.05$. This test is non-parametric and not based on a normality assumption, which makes it more generally applicable than the well-known t-test. The Bayesian approach used is a Bayes factor for a two-sample design (Morey et al., 2011), used on the same two groups of f_{obs} values. Following the significance scale proposed by Lee and Wagenmakers (2014), Bayes factors ≥ 10 provide strong evidence of significant difference between the samples.

3.4. Missingness Based on Own and Unobserved Values (MBOUV)

Missingness Based on Own and Unobserved Values (MBOUV) is a combination of the MBOV and MBUV approaches to perform multivariate generation of missing values. The desired missingness rate is global for the entire dataset and randomly distributed among all its features, leading to different levels of missingness for different features. Afterwards, the generation follows two rules: MBUV is applied to all nominal features and to half of the numeric ones, while MBOV (with lowest values removal) is applied to the remaining half of the features. Both approaches are applied iteratively, and the split of numeric features is random.

4. Benchmark setup

The benchmark study uses a comprehensive baseline of state-of-the-art imputation methods, which are briefly described in following list:

- Denoising Autoencoder (DAE) is a neural network that learns a compressed representation of the input data, aiming to output a reconstruction of that same data (Charte et al., 2018). It can be a shallow or deep model, depending on the number of hidden layers, and its parameters are optimized in the same way as any other neural network, that is, usually by stochastic gradient descent or one of its variants. The loss function measures the error between the data at the input and output layers. A DAE is trained with a corrupted version of the input data, created by adding noise. In the missing data context, the corruption corresponds to the absence of values (Vincent et al., 2008; Choudhury & Pal, 2019).
- Variational Autoencoder (VAE) is an autoencoder trained to learn a Gaussian multivariate latent representation of the input data, generating new values by sampling from this latent Gaussian representation (Kingma & Welling, 2013; McCoy et al., 2018). A VAE is a generative model trained by minimizing a loss function corresponding to the reconstruction error, plus a Kullback–Leibler divergence encouraging the learned distribution in the latent space to be close to the prior distribution.
- k-Nearest Neighbors (kNN) imputation with $k = 5$ finds the k nearest neighbors of the record with missing values (using the Euclidean distance) and imputes with the average values (for numeric features) or the most common ones (for categorical features) from the k neighbors (Santos et al., 2015).
- Mean/Mode Imputation uses the mean (for numeric features) or the mode (for categorical features) to impute.
- Multiple Imputation by Chained Equations (MICE) iteratively fits several regression models through a multiple imputation strategy, assuming as the dependent features those containing missing data (Buuren & Groothuis-Oudshoorn, 2010). The procedure follows the following steps:

    1. The missing values are pre-imputed with a simple method, usually the mean/mode.

Algorithm 1 Pseudocode for the MBIR procedure.

Input: $dataset$, f_{miss}, mr
Output: $amputed\_dataset$
1: **for** each f_{obs} in continuous and ordinal features of $dataset \setminus \{f_{\text{miss}}\}$ **do**
2: Find the instances $inst_{\text{mr}}$ with the lower $mr\%$ values of f_{obs}
3: Set the f_{miss} values to be missing on the $inst_{\text{mr}}$ instances
4: Create the missing indicator f_{ind} for f_{miss} ($f_{\text{ind}} = 1$ when missing, $f_{\text{ind}} = 0$ otherwise)
5: Find if the differences between f_{obs} values for $f_{\text{ind}} = 1$ and $f_{\text{ind}} = 0$ are statistically significant
6: **if** differences are statistically significant **then**
7: Save f_{obs} to $MAR_{f_{\text{obs}}}$
8: **end if**
9: **end for**
10: Select the f_{obs} from $MAR_{f_{\text{obs}}}$ associated with the iteration that presented the most statistically significant differences
11: Create $amputed\_dataset$ by applying lines 2 and 3 to the $dataset$
12: Delete the selected f_{obs} feature from $amputed\_dataset$
13: **return** $amputed\_dataset$

Table 1
Architecture details of the autoencoder-based methods.

Aspect	Configuration
DAE Structure	Two hidden layers with 10 units and ReLU activation function.
VAE Structure	Two hidden layers with 10 units, plus the three layers needed for the generative process: two "parallel" layers (mean and variance) with 10 units, connected to another one that samples new points from the latent Gaussian distribution.
Optimization Algorithm	Adam, with the MSE loss and a learning rate of 0.001.
Training Procedure	Batches of 64 instances, for a maximum of 2000 epochs, early stopping, and learning rate reduction of 80% when the validation loss shows no improvements in 100 consecutive epochs.
Overfitting Mitigation	Each hidden layer under L2 regularization with 0.01 weight.
Network Initialization	Network weights are randomly sampled from a truncated normal distribution centered at zero.
Output	Sigmoid activation function, since the data is normalized to the range $[0, 1]$.

2. Each feature originally containing missing values is modeled using regression (linear or not), with all the remaining features as the independent variables, including the complete and pre-imputed observations. This is applied iteratively to each feature.
3. The result is a complete dataset imputed by the multiple regressions. However, many factors such as the features' order may create bias in this dataset. To mitigate it, the procedure is repeated several times until the parameters of the regressions stabilize.

- SoftImpute is a matrix completion process, achieved through a nuclear-norm minimization approach, where the missing data is imputed with estimated values obtained with a soft-thresholded SVD (Mazumder et al., 2010).

Furthermore, the autoencoder-based methods (DAE and VAE) require the choice of neural network architectures and hyper-parameters, which are described in Table 1.

The experiments were conducted on 10 public medical datasets, in different domains of clinical research containing routinely collected data. As previously stated, this type of data is ideal since it often suffers from MNAR missing data (Peek & Rodrigues, 2018). All these datasets are publicly available at the UCI Machine Learning Repository and their details are presented in Table 2.

These datasets are complete (i.e., do not have missing data) and, during the experiment, they are injected with missing values under MNAR mechanisms using the approaches described in Section 3. This approach was applied iteratively to each feature, excepted for the MBOUV method since it already provides a multivariate process. The imputation results are assessed through the mean absolute error (MAE) between the original dataset and the imputed one. For nominal features, this metric is applied to the one-hot encoded vectors, and the results are then transformed into error rates in the same scale of the numeric and ordinal features. For aggregation purposes, the overall MAE of a dataset is computed by averaging the MAE of all the features. Other metrics such as the mean squared error and root mean squared error are also computed, but the results and respective conclusions are the same as for MAE. For that reason, and to avoid redundant information, only the MAE results are presented. A scheme of the experimental procedure is shown in Fig. 3.

Table 2
Details of the medical datasets used in the study.

Dataset	# Instances	# Features	
		Categorical	Continuous
bc-coimbra	116	1	9
cleveland	303	7	7
cmc	1473	8	2
ctg	2126	2	23
pima	768	1	8
saheart	462	2	8
thyroid	7200	16	6
transfusion	748	1	4
vertebral	310	1	5
wisconsin	569	1	30

All features are normalized to the interval $[0, 1]$ and split into train (50%), validation (20%), and test (30%) sets, which are used by the methods that require a training procedure. The data split is performed in a stratified manner to ensure equal missingness rates and MNAR assumptions for all sets. One-hot encoding is applied to the categorical features after missing data injection. The benchmark study covers five missingness rates (10%, 20%, 40%, 60%, and 80%), allowing for a comprehensive analysis of the imputation results. The autoencoder-based methods require a pre-imputation of the missing values, which is performed with the mean/mode method.

The experiment was independently executed 30 times. The datasets were shuffled in each execution, and the final results are the average over these 30 runs. Furthermore, the results are provided with 95% confidence intervals for significance evaluation: if in a given experiment, two imputation methods have MAE results with overlapping confidence intervals, the difference between those results is not considered statistically significant.

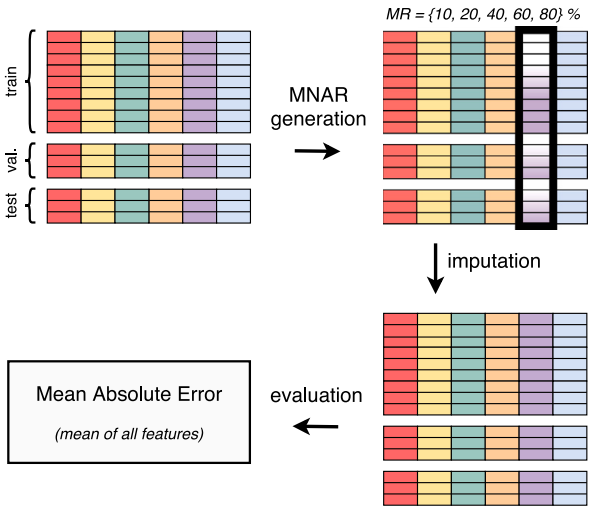


Fig. 3. High-level representation of the benchmark setup.

5. Experimental results

To study the imputation patterns in different types of data, the analysis is first performed separately for numeric (i.e., continuous) and categorical features (Sections 5.1 and 5.2, respectively) and, afterwards, for all features together (Section 5.3). The results obtained for each applicable MNAR strategy are discussed individually in each of the subsections, while the general conclusions are in Section 6. The results are presented using charts where the MAE values for each missingness rate are plotted for all imputation approaches. The 95% confidence intervals of the results are displayed as translucent grey areas behind the plotted points and lines. When these intervals overlap for different imputation methods, the grey areas become more opaque, meaning that the differences between the results of those methods are not statistically significant. For a more detailed view of the results, the raw data is provided in Appendix.

5.1. Numeric features

5.1.1. Missingness Based on Own Values (MBOV)

Fig. 4 shows the results for the MBOV approach with the lowest values removed. While SoftImpute, followed by MICE, present the best results for lower rates (minimum MAEs of 0.084 and 0.1, respectively), AE-based imputation, particularly the VAE, performs better at higher rates (minimum MAE of 0.075), with the inflection point around the 40% rate.

Fig. 5 shows MBOV results but with the highest values removed. The main conclusions regarding the best and worst imputation approaches are essentially the same obtained for MBOV with lowest value removal. The exception is the SoftImpute method, which achieved significantly worse results in this setting (only the mean imputation presented worst results). Therefore, MICE obtained the best MAE (0.152) for lower missingness rates, while AE-based imputation achieved again the best results for higher rates (minimum MAE of 0.087 for the VAE). The inflection point is around the 30% rate in this setting. There is also an inversion of the MAE trend when the missing rate increases. For MBOV with lowest value removal, the MAE tends to increase gradually with the missing rate, except for the AE methods where the opposite happens. For MBOV with higher value removal, the MAE tends to decrease with the missing rate for all imputation methods. Although this change of behavior is probably related to the features' distribution, it is important to notice that the autoencoder methods are both consistent in terms of results and behavior. In both MBOV settings, the worst results were obtained by the mean imputation (maximum MAE of 0.332).

Fig. 6 presents the results for the MBOV approach when a quarter of the missing values are randomly chosen ($p = 0.25$), therefore being under MCAR assumptions. Here, the SoftImpute method is consistently the best (minimum MAE of 0.08), being only slightly surpassed by the VAE for the 80% rate (MAE of 0.088). The worst results were again obtained by mean imputation (maximum MAE of 0.201). The pattern of results is similar to the MBOV with lowest value removal, but the 25% MCAR values lead to a significant improvement of the SoftImpute results.

Fig. 7 shows the results for the MBOV approach with the middlemost values being removed. This is the MBOV setting where the pattern at the results is most different. Up to 40% missingness rate, the results do not present significant differences among most imputation methods. The exception is kNN imputation, which is consistently the worst for all rates (maximum MAE of 0.132). Nevertheless, for the 10% and 20% rates, mean imputation shows the best results (minimum MAE of 0.038). Since the values removed are the middlemost ones, there is a good chance they will be close to the features' mean, which can justify these results. However, above the 40% rate, the different imputation methods tend to produce results with more significant differences. For the 40%, 60%, and 80% rates, VAE outperforms the remaining methods (minimum MAE of 0.052).

5.1.2. Missingness Based on Unobserved Values (MBUV)

Fig. 8 presents the results for the MBUV missingness mechanism. In this setting, all imputation approaches exhibit consistent MAE values for all rates, except the SoftImpute, which tends to perform worse at higher rates. The best imputation was achieved by the MICE method (minimum MAE of 0.062) and the worst by the mean imputation (maximum MAE of 0.119).

5.1.3. Missingness Based on Intra-Relation (MBIR)

For MBIR, Figs. 9 and 10 show the results for the frequentist and Bayesian tests (see Section 3.3), respectively. The results obtained using both tests are essentially the same, with small MAE differences that can be accounted by the stochastic nature of some methods. The best and worst methods were respectively the MICE (minimum MAE of 0.065) and mean imputation (maximum MAE of 0.161), consistently throughout all missing rates. There is a slight increase in the MAE values when the missing rate grows, except for the AE-based methods, where the behavior is the opposite. Furthermore, the differences between the MAE values of the kNN and SoftImpute methods are not significant since their confidence intervals often overlap.

5.2. Categorical features

5.2.1. Missingness Based on Unobserved Values (MBUV)

Fig. 11 presents the results for the MBUV approach. An immediately noticeable aspect is the large confidence intervals displayed by all imputation methods, which leads to an absence of significantly best and worst methods when considering the amount of overlap. The clearest differences are between the mode imputation and the SoftImpute methods, both appearing to be significantly worse than the others. Nevertheless, mode imputation shows the worst MAE up to the 60% rate, being surpassed by SoftImpute after that rate (maximum MAEs of 0.188 and 0.205, respectively). The best MAE values are obtained with the MICE method, which is only slightly surpassed by the VAE after the 60% rate (minimum MAEs of 0.125 and 0.134, respectively).

5.2.2. Missingness Based on Intra-Relation (MBIR)

Figs. 12 and 13 present the MBIR results for the frequentist and Bayesian tests, respectively. Once again, there are no significant differences between the MAE values obtained with both statistical tests. Moreover, the exact same patterns of results found for the MBUV mechanism are visible here: large confidence intervals that often overlap, although this overlap tends to fade away above the 60% rate;

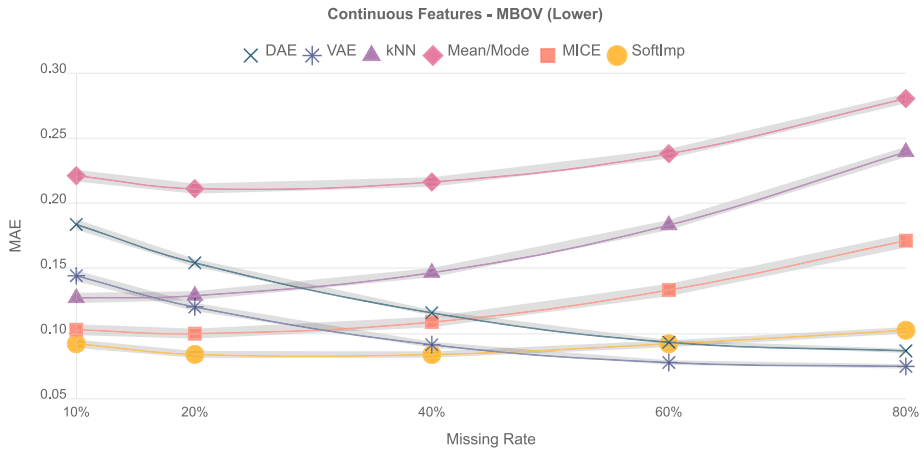


Fig. 4. Results for MBOV with lowest values being removed, applied to continuous features.

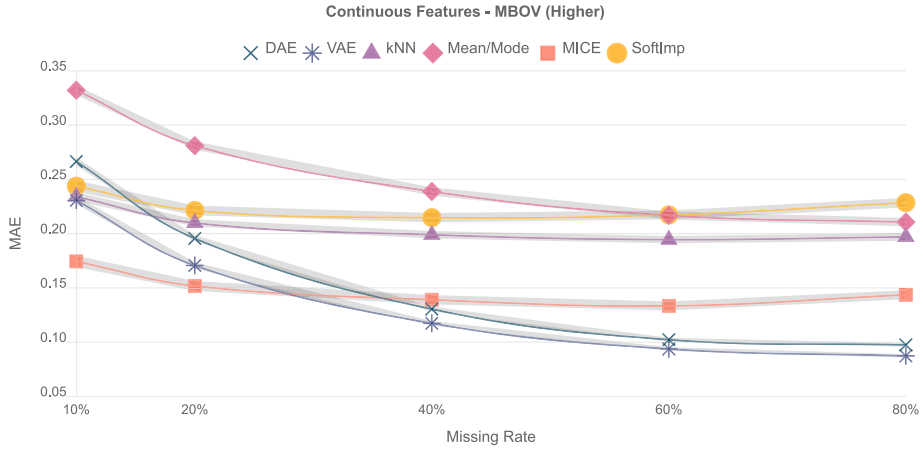


Fig. 5. Results for MBOV with highest values being removed, applied to continuous features.

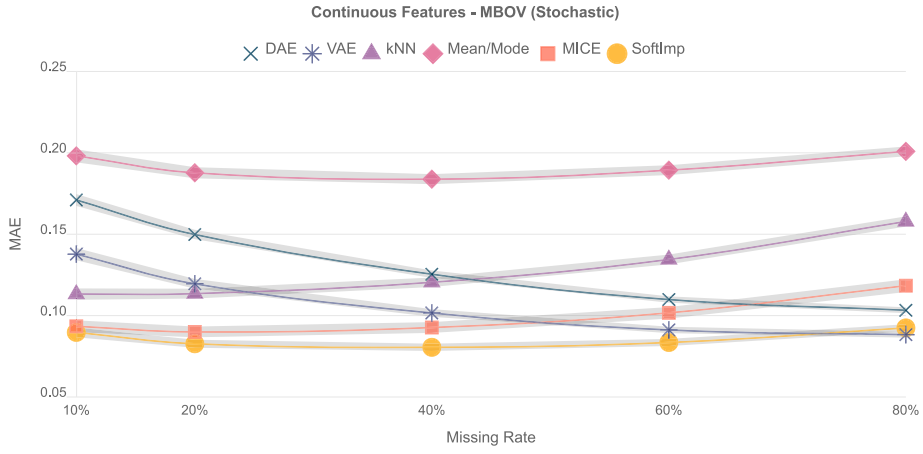


Fig. 6. Results for MBOV with partial random missingness, applied to continuous features.

the mode imputation and the SoftImpute methods still provide the worst MAE values, although the mode achieved worse results in this setting (maximum MAE of 0.221); MICE is generally the best method (minimum MAE of 0.111), although now it is surpassed by both AE-based methods beyond the 60% missingness rate (minimum MAE of 0.131 for the VAE).

5.3. Mixed features

5.3.1. Missingness Based on Unobserved Values (MBUV)

Fig. 14 shows the results for MBUV. The conclusions are the same as those obtained for this generation approach when only numeric features are considered: the MAE values are consistent for all missing rates,

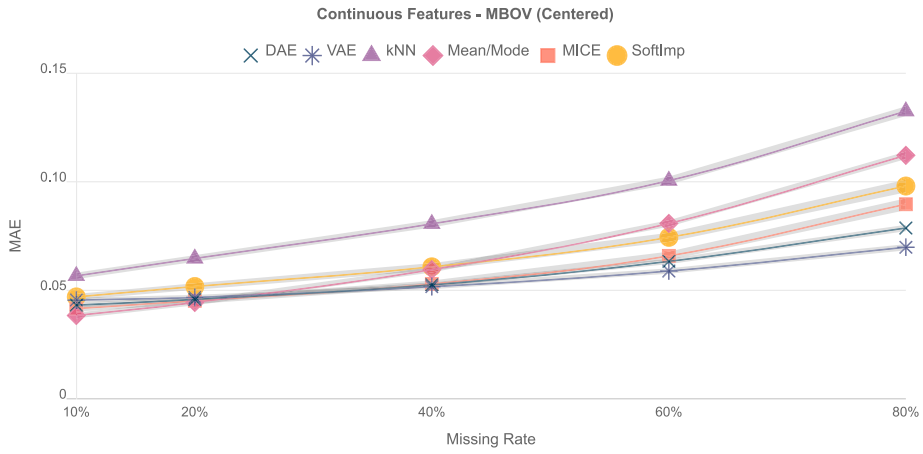


Fig. 7. Results for MBOV with middlemost values removal, applied to continuous features.

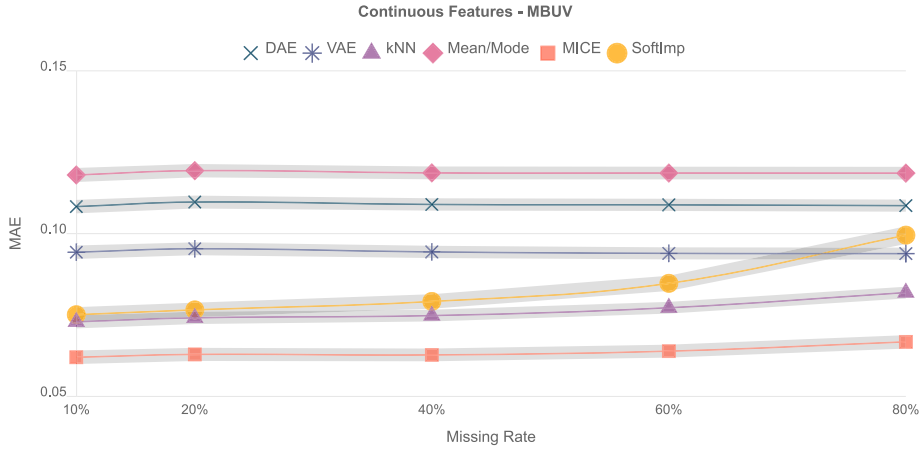


Fig. 8. Results for MBUV applied to continuous features.

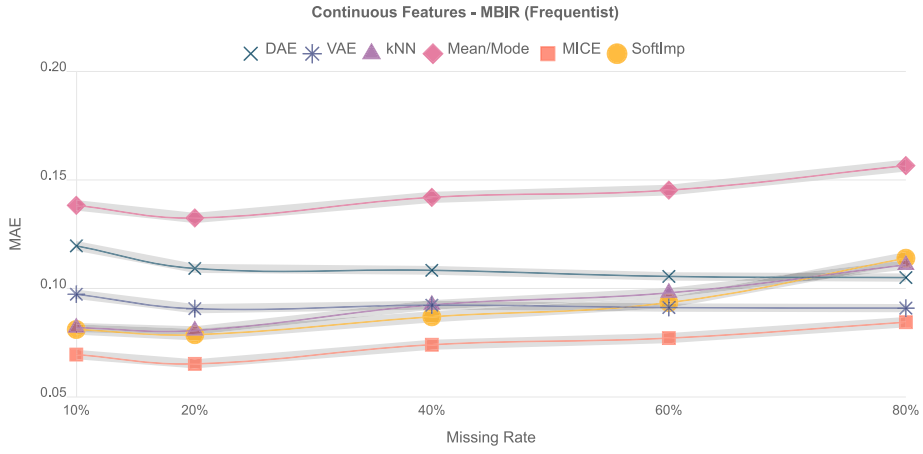


Fig. 9. Results for MBIR with the frequentist test, applied to continuous features.

with the single exception of SoftImpute that performs worse with rates above 60%. The best imputation results were achieved by the MICE method (minimum MAE of 0.076) and the worst by the mean/mode imputation (maximum MAE of 0.131). In fact, comparing with numeric features only, the MAE values in this mixed features setting are shifted towards higher errors. This is a clear indicator that the categorical features have a negative impact on the results, suggesting that all these methods tend to perform better with numeric features.

5.3.2. *Missingness Based on Intra-Relation (MBIR)*

The MBIR results for the frequentist and Bayesian tests are, respectively, available on Figs. 15 and 16. Following the pattern found with the numeric and categorical cases, there are no significant differences between both statistical tests. In fact, similarly to the MBUV results, the conclusions for MBIR are the same as for the numeric features alone: the best and worst methods continue to be the MICE (minimum MAE of 0.076) and mean/mode imputation (maximum MAE of 0.17) for all

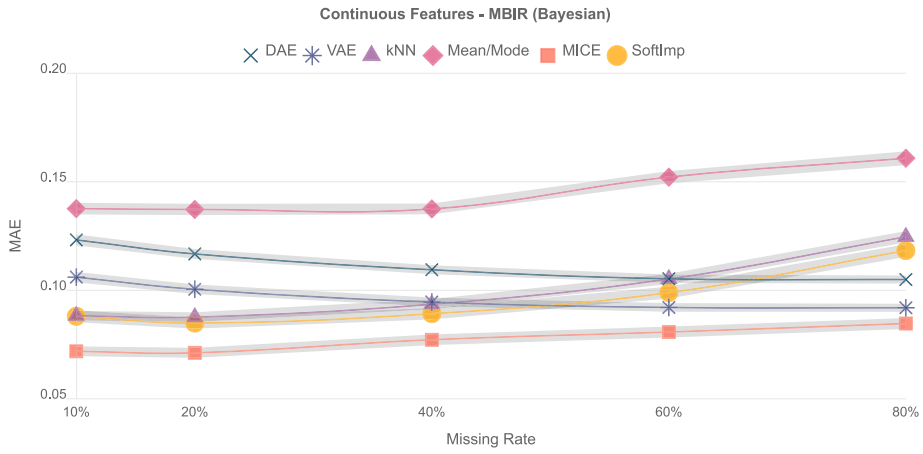


Fig. 10. Results for MBIR with the Bayesian test, applied to continuous features.

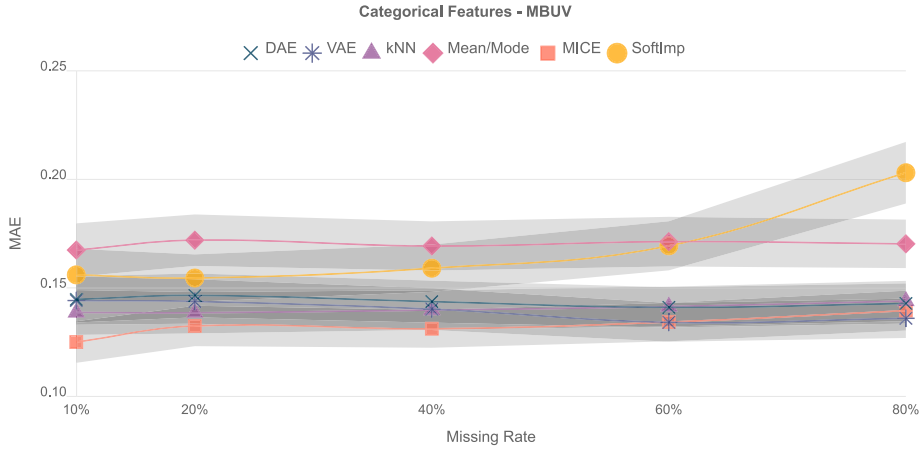


Fig. 11. Results for MBUV applied to categorical features.

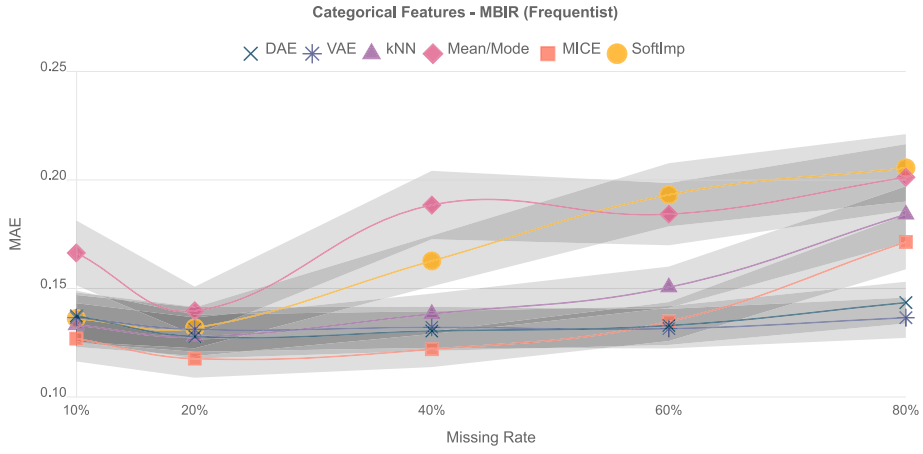


Fig. 12. Results for MBIR with the frequentist test, applied to categorical features.

missing rates, expect for the 80% rate where the VAE surpasses MICE (minimum MAE of 0.101). The kNN and SoftImpute methods continue to not show significantly different results. Moreover, as previously found for the MBUV mechanism, the MAE results are shifted towards worse values in this mixed setting when comparing to the continuous features setting.

5.3.3. Missingness Based on Own and Unobserved Values (MBOUV)

Fig. 17 presents the results for the MBOUV mechanism. Up to the 40% rate, there are no significant differences among the MAE values of the imputation methods, with the exception of mean/mode imputation, which clearly yields worse results (maximum MAE of 0.212). The MICE and kNN methods appear to provide slightly better results for

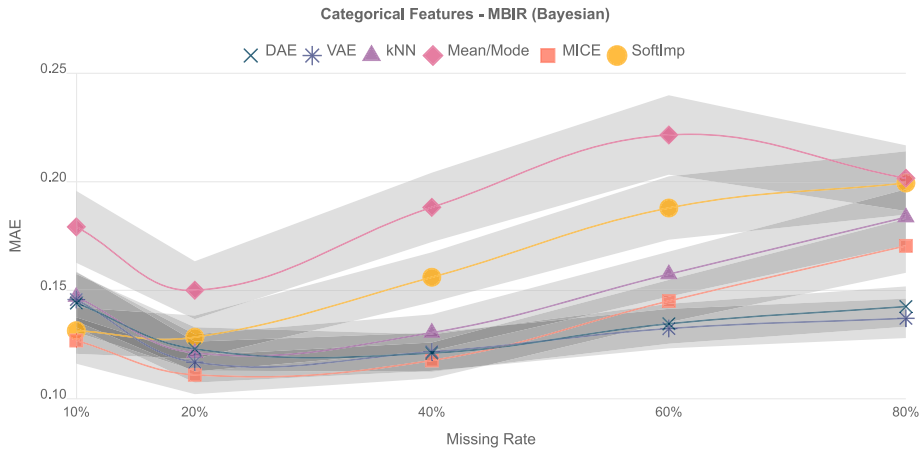


Fig. 13. Results for MBIR with the Bayesian test, applied to categorical features.

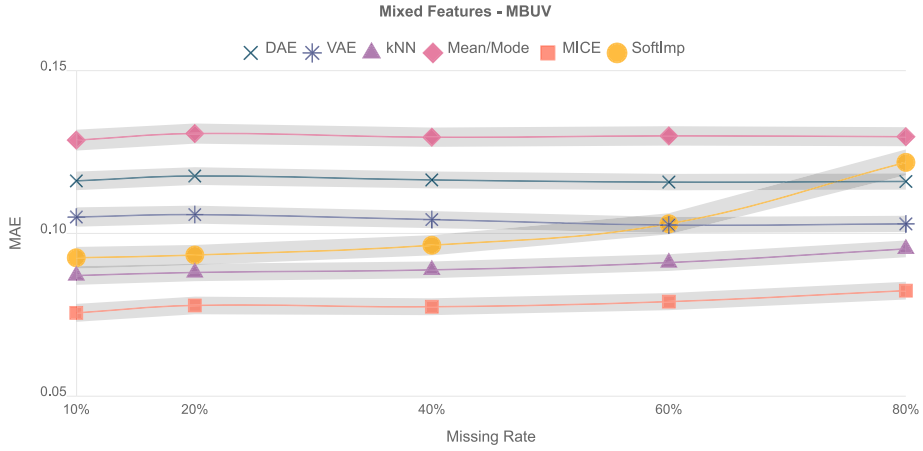


Fig. 14. Results for MBUV applied to all (mixed) features.

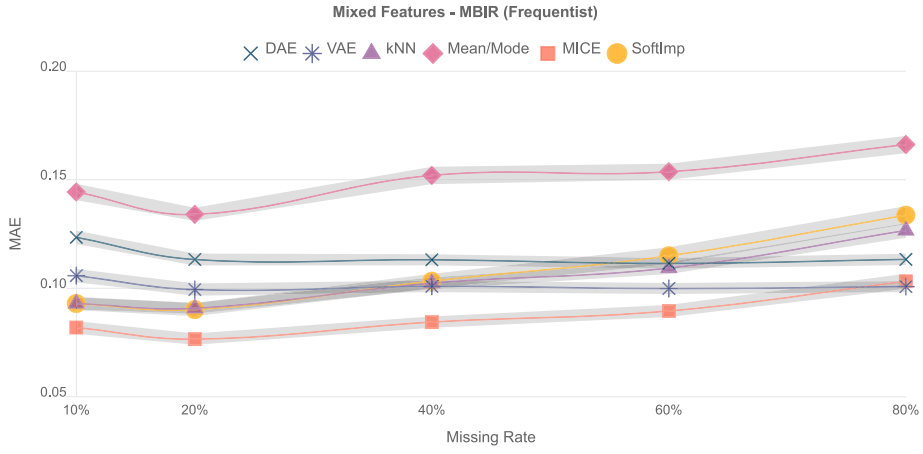


Fig. 15. Results for MBIR with the frequentist test, applied to all (mixed) features.

the 10% and 20% rates, but the improvements are not significant (minimum MAEs of 0.113 and 0.114, respectively). However, for rates above 40%, there is a disparity between the imputation methods, with the AEs gradually showing significant improvements relatively to the remaining approaches (minimum MAE of 0.139 for the DAE). An interesting aspect is that for higher rates, MICE presents the worst results by a considerable margin (the MAE value for the 80% rate was clipped). This behavior differs from previous experiments where MICE presented consistently good results and small confidence intervals for the univariate scenarios.

6. Discussion and future directions

From all the results previously analyzed, there are several main conclusions that can be drawn. To help this discussion, Table 3 presents a summary of the results by showing the best imputation method for each

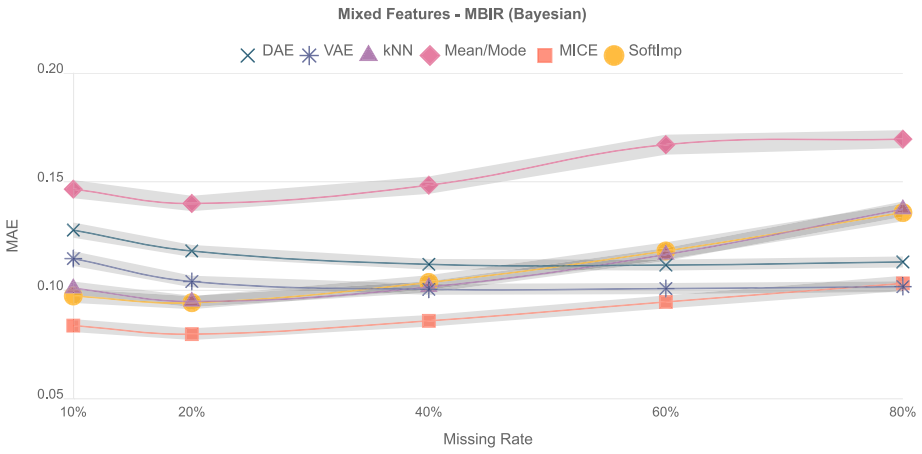


Fig. 16. Results for MBIR with the Bayesian test, applied to all (mixed) features.

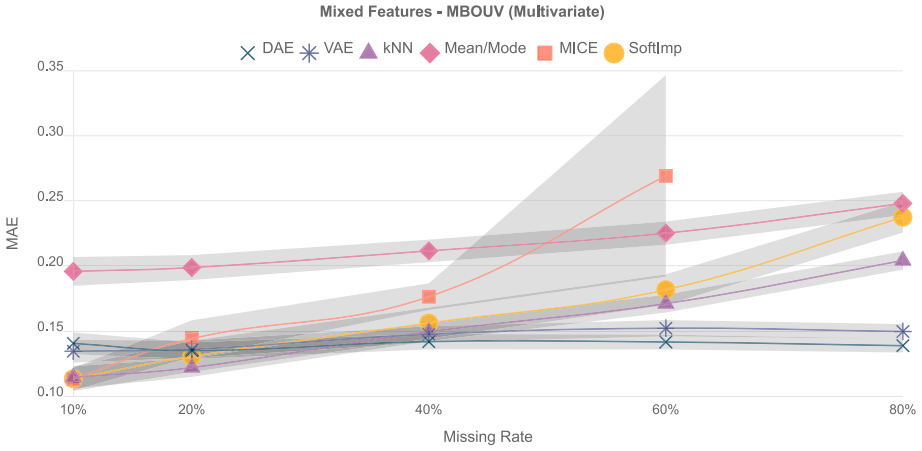


Fig. 17. Results for MBOUV applied to all (mixed) features.

Table 3
Best imputation method for each missing rate, MNAR generation and data type. The methods which presented statistically significant improvements are marked with an asterisk.

Data type	MNAR generation	Missing rate				
		10%	20%	40%	60%	80%
Continuous	MBOV (Lower)	SoftImp*	SoftImp*	SoftImp*	VAE*	VAE*
	MBOV (Higher)	MICE*	MICE*	VAE*	VAE*	VAE*
	MBOV (Stochastic)	SoftImp	SoftImp*	SoftImp*	SoftImp*	VAE*
	MBOV (Centered)	Mean/Mode*	Mean/Mode	VAE	VAE*	VAE*
	MBUV	MICE*	MICE*	MICE*	MICE*	MICE*
	MBIR (Frequentist)	MICE*	MICE*	MICE*	MICE*	MICE*
	MBIR (Bayesian)	MICE*	MICE*	MICE*	MICE*	MICE*
Categorical	MBUV	MICE	MICE	MICE	VAE	VAE
	MBIR (Frequentist)	MICE	MICE	MICE	VAE	VAE
	MBIR (Bayesian)	MICE	MICE	MICE	VAE	VAE
Mixed	MBUV	MICE*	MICE*	MICE*	MICE*	MICE*
	MBIR (Frequentist)	MICE*	MICE*	MICE*	MICE*	VAE
	MBIR (Bayesian)	MICE*	MICE*	MICE*	MICE*	VAE
	MBOUV (Multivariate)	MICE	KNN	DAE	DAE	DAE

missingness rate, MNAR mechanism, and data type. Methods marked with * presented statistically significant improvements compared to all the remaining ones.

For low and medium rates (up to 60%), MICE is often the best imputation method. However, for half of the MBOV settings, SoftImpute showed better results, although MICE immediately follows very closely. The single exception is for MBOV with the middlemost values being removed, where the mean imputation surpassed these methods. This is an expected result considering that the middlemost values will be naturally close to the mean value, and it would only be useful in a real-world scenario in which the missing data would have these characteristics. Therefore, the superior consistency of the MICE results is undeniable with the low and medium levels of missingness. However, for higher rates (above 60%), MICE loses accuracy and is often surpassed by the AE-based methods (VAE and DAE), which are the best ones for these high levels of missingness with the MBOV, MBOUV, and all categorical features settings. A plausible justification is that, with lower missing

rates, the amount of information available is still enough for methods usually applied in MAR contexts (such as MICE and SoftImpute) to still achieve the best results. In such settings, the MNAR assumptions are weaker, which leads to a lower level of uncertainty while performing the imputation, since this process is eased by the available information. This is visible in the MBOV setting with a stochastic component, where 25% of the missing values are under MCAR (which weakens the MNAR assumptions) and the results for SoftImpute and MICE improve in all missing rates. Furthermore, the multiple imputation strategy followed by the MICE method helps to deal with the more noticeable uncertainty in the medium levels of missingness, which is another justification for its superiority.

For higher rates, the AE-based methods show higher resilience to noise (i.e., corrupted data) and better generalization, leading to better results. For example, in several MNAR scenarios for numeric data, their results tend to improve when the missing rate increases. This is a clear indicator of how well AEs deal with corrupted data, to the point where they improve their results with more noise mostly because it leads to less overfitting and improved generalization (Vincent et al., 2010). Besides, AEs can also capture more complex patterns since they rely on non-linear activation functions, while methods such as MICE use linear models. This is particularly relevant with high missing rates since the mechanism describing the missing values is expected to be more complex and uncertain. Moreover, AEs do not assume that the data follow a specific distribution, which is also helpful since the MNAR mechanism introduces deviations in the data distributions, which tend to increase when the missing rates are high. However, AEs also present a major limitation: they need to be trained with complete data in order to be applied to incomplete data. In many real-world scenarios, this may be unfeasible, since the datasets may not contain enough complete data to properly train the network.

Analyzing the significance of the differences between imputation methods, while the results for the numeric features have low variance and are significantly different, the opposite is found for categorical features: the confidence intervals are large for all imputation methods and consistently overlap. This pattern of results, associated with higher MAE values, show that all the imputation methods are more suitable for numeric features. Regarding the consistency of the results across different missingness rates, there is a natural tendency for the imputation methods to gradually perform worse for higher levels of missingness. However, as previously explained, the AE-based methods show the opposite trend for most scenarios with numeric features, which makes them a natural candidate for these high levels of missingness.

Finally, the approaches herein proposed for artificial generation of MNAR provide an ample experimental setup that mimics different characteristics of this type of mechanism. Although there is some consistency on the best and worst imputation approaches within different generation approaches, the results tend to be different both in terms of trend and error magnitude. The single exception are the frequentist and Bayesian tests in the MBIR approach, since both lead to similar results. Apart from this exception, the experimental results show the importance of properly simulating the missing data scenarios, with realistic strategies, in order to obtain meaningful and trustworthy results.

The take-home message from these benchmark results can be summarized in the following guidelines:

- For low and medium levels of missingness in univariate settings, the MICE and SoftImpute methods should be the go-to solution. For multivariate scenarios, MICE should be avoided;
- For higher missing rates, AE-based methods should be considered, particularly when there is enough complete data to properly train the network. Otherwise, MICE and SoftImpute are good choices;
- These three methods (MICE, SoftImpute, and AE) deal well with departures from the MNAR assumptions, which makes them suitable for scenarios with mixed missing data mechanisms;
- The kNN and mean/mode imputation are still among the most widely used methods (Garciarena & Santana, 2017). However, they should be avoided in MNAR settings. This conclusion comes without surprise, since they are often referred to as methods for MAR and MCAR;
- All these methods can be applied to mixed data types, but they are more suitable for numeric data. Better approaches for categorical features are needed. Although the imputation of categorical features can be transformed into a classification problem, this does not provide a global imputation strategy for the dataset;
- It is not possible to state that a specific generation strategy is better at mimicking the MNAR nature since they all cover different characteristics of this type of mechanism. Instead, a mix of these strategies should be used to obtain more representative and reliable results.

In the future, several research directions should be explored. Benchmark studies should be further extended to accommodate settings where the missing mechanism is a composition of MNAR with MCAR and/or MAR, a setting often found in real-world data. Furthermore, the impact of post-processing sometimes applied to MNAR imputation should also be studied and considered in the benchmark (e.g., sensitivity analysis and the application of the delta-adjustment method). Regarding the artificial generation of MNAR, there are specific settings that can be incorporated into new strategies. For example, for data domains where sensitive attributes exist, it is reasonable that the mechanism may only be applied to those types of attributes, while other mechanisms can be used for the remaining features. This would create a concept of sensitive-aware MNAR generation.

CRediT authorship contribution statement

Ricardo Cardoso Pereira: Conceptualization, Methodology, Software, Investigation, Validation, Writing – original draft. **Pedro Henriques Abreu:** Conceptualization, Methodology, Validation, Writing – review & editing, Supervision. **Pedro Pereira Rodrigues:** Conceptualization, Methodology, Validation, Writing – review & editing, Supervision. **Mário A.T. Figueiredo:** Conceptualization, Methodology, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data used in this research is public.

Acknowledgments

This work is supported in part by the FCT - Foundation for Science and Technology, I.P., Research Grant SFRH/BD/149018/2019. This work is also funded by the FCT - Foundation for Science and Technology, I.P./MCTES through national funds (PIDDAC), within the scope of CISUC R&D Unit - UIDB/00326/2020 or project code UIDP/00326/2020.

Appendix. Supplementary tables

For a more detailed view of the experimental results, the raw data is attached in Tables A.1–A.3 for the continuous, categorical, and mixed analysis, respectively. Each table presents the mean absolute error (MAE) values and their 95% confidence intervals for each imputation method, missing rate and Missing Not At Random (MNAR) generation strategy.

Table A.1
Results for continuous features. The MAE values and their 95% confidence intervals are presented for each imputation method.

MNAR	Missing	DAE		VAE		kNN		Mean/Mode		MICE		SoftImp	
generation	rate	μ	(95% CI)	μ	(95% CI)	μ	(95% CI)	μ	(95% CI)	μ	(95% CI)	μ	(95% CI)
	10%	0.184	(0.18;0.19)	0.144	(0.14;0.15)	0.127	(0.12;0.13)	0.221	(0.22;0.23)	0.103	(0.10;0.11)	0.092	(0.09;0.10)
	20%	0.154	(0.15;0.16)	0.120	(0.12;0.12)	0.129	(0.13;0.13)	0.211	(0.21;0.22)	0.100	(0.10;0.10)	0.084	(0.08;0.09)
MBOV (Lower)	40%	0.116	(0.11;0.12)	0.091	(0.09;0.09)	0.147	(0.14;0.15)	0.216	(0.21;0.22)	0.109	(0.10;0.11)	0.084	(0.08;0.09)
	60%	0.093	(0.09;0.09)	0.078	(0.08;0.08)	0.183	(0.18;0.19)	0.238	(0.23;0.24)	0.133	(0.13;0.14)	0.092	(0.09;0.09)
	80%	0.087	(0.08;0.09)	0.075	(0.07;0.08)	0.239	(0.24;0.24)	0.280	(0.28;0.28)	0.171	(0.17;0.18)	0.102	(0.10;0.10)
	10%	0.266	(0.26;0.27)	0.230	(0.23;0.23)	0.234	(0.23;0.24)	0.332	(0.33;0.34)	0.174	(0.17;0.18)	0.244	(0.24;0.25)
	20%	0.195	(0.19;0.20)	0.170	(0.17;0.17)	0.210	(0.21;0.21)	0.281	(0.28;0.28)	0.152	(0.15;0.16)	0.221	(0.22;0.23)
MBOV (Higher)	40%	0.130	(0.13;0.13)	0.117	(0.12;0.12)	0.199	(0.20;0.20)	0.239	(0.23;0.24)	0.139	(0.13;0.14)	0.215	(0.21;0.22)
	60%	0.102	(0.10;0.10)	0.094	(0.09;0.10)	0.194	(0.19;0.20)	0.217	(0.21;0.22)	0.133	(0.13;0.14)	0.217	(0.21;0.22)
	80%	0.097	(0.10;0.10)	0.087	(0.09;0.09)	0.197	(0.19;0.20)	0.210	(0.21;0.21)	0.143	(0.14;0.15)	0.228	(0.22;0.23)
	10%	0.171	(0.17;0.17)	0.138	(0.13;0.14)	0.113	(0.11;0.12)	0.198	(0.19;0.20)	0.093	(0.09;0.10)	0.090	(0.09;0.09)
	20%	0.150	(0.15;0.15)	0.119	(0.12;0.12)	0.113	(0.11;0.12)	0.188	(0.18;0.19)	0.090	(0.09;0.09)	0.083	(0.08;0.09)
MBOV (Stochastic)	40%	0.125	(0.12;0.13)	0.101	(0.10;0.10)	0.120	(0.12;0.12)	0.184	(0.18;0.19)	0.093	(0.09;0.10)	0.080	(0.08;0.08)
	60%	0.110	(0.11;0.11)	0.091	(0.09;0.09)	0.134	(0.13;0.14)	0.189	(0.19;0.19)	0.102	(0.10;0.11)	0.083	(0.08;0.09)
	80%	0.103	(0.10;0.10)	0.088	(0.09;0.09)	0.158	(0.15;0.16)	0.201	(0.20;0.20)	0.118	(0.11;0.12)	0.092	(0.09;0.09)
	10%	0.043	(0.04;0.04)	0.045	(0.04;0.05)	0.057	(0.06;0.06)	0.038	(0.04;0.04)	0.041	(0.04;0.04)	0.047	(0.05;0.05)
	20%	0.046	(0.04;0.05)	0.047	(0.05;0.05)	0.065	(0.06;0.07)	0.044	(0.04;0.05)	0.045	(0.04;0.05)	0.052	(0.05;0.05)
MBOV (Centered)	40%	0.052	(0.05;0.05)	0.052	(0.05;0.05)	0.080	(0.08;0.08)	0.059	(0.06;0.06)	0.053	(0.05;0.05)	0.061	(0.06;0.06)
	60%	0.063	(0.06;0.06)	0.059	(0.06;0.06)	0.100	(0.10;0.10)	0.081	(0.08;0.08)	0.066	(0.06;0.07)	0.074	(0.07;0.08)
	80%	0.079	(0.08;0.08)	0.070	(0.07;0.07)	0.132	(0.13;0.13)	0.112	(0.11;0.11)	0.090	(0.09;0.09)	0.098	(0.09;0.10)
	10%	0.108	(0.11;0.11)	0.094	(0.09;0.10)	0.073	(0.07;0.07)	0.118	(0.12;0.12)	0.062	(0.06;0.06)	0.075	(0.07;0.08)
	20%	0.110	(0.11;0.11)	0.095	(0.09;0.10)	0.074	(0.07;0.08)	0.119	(0.12;0.12)	0.063	(0.06;0.06)	0.077	(0.07;0.08)
MBUV	40%	0.109	(0.11;0.11)	0.094	(0.09;0.10)	0.075	(0.07;0.08)	0.119	(0.12;0.12)	0.063	(0.06;0.06)	0.079	(0.08;0.08)
	60%	0.109	(0.11;0.11)	0.094	(0.09;0.10)	0.077	(0.08;0.08)	0.119	(0.12;0.12)	0.064	(0.06;0.07)	0.085	(0.08;0.09)
	80%	0.109	(0.11;0.11)	0.094	(0.09;0.10)	0.082	(0.08;0.08)	0.118	(0.12;0.12)	0.067	(0.06;0.07)	0.099	(0.10;0.10)
	10%	0.120	(0.12;0.12)	0.097	(0.10;0.10)	0.082	(0.08;0.08)	0.138	(0.14;0.14)	0.069	(0.07;0.07)	0.081	(0.08;0.08)
	20%	0.109	(0.11;0.11)	0.091	(0.09;0.09)	0.081	(0.08;0.08)	0.132	(0.13;0.13)	0.065	(0.06;0.07)	0.079	(0.08;0.08)
MBIR (Frequentist)	40%	0.108	(0.11;0.11)	0.092	(0.09;0.09)	0.092	(0.09;0.09)	0.142	(0.14;0.14)	0.074	(0.07;0.08)	0.087	(0.08;0.09)
	60%	0.105	(0.10;0.11)	0.091	(0.09;0.09)	0.098	(0.10;0.10)	0.145	(0.14;0.15)	0.077	(0.07;0.08)	0.093	(0.09;0.10)
	80%	0.105	(0.10;0.11)	0.091	(0.09;0.09)	0.111	(0.11;0.11)	0.156	(0.15;0.16)	0.084	(0.08;0.09)	0.114	(0.11;0.12)
	10%	0.123	(0.12;0.13)	0.106	(0.10;0.11)	0.088	(0.09;0.09)	0.138	(0.13;0.14)	0.072	(0.07;0.07)	0.088	(0.09;0.09)
	20%	0.117	(0.11;0.12)	0.100	(0.10;0.10)	0.088	(0.09;0.09)	0.137	(0.13;0.14)	0.071	(0.07;0.07)	0.085	(0.08;0.09)
MBIR (Bayesian)	40%	0.109	(0.11;0.11)	0.094	(0.09;0.10)	0.093	(0.09;0.10)	0.137	(0.13;0.14)	0.077	(0.07;0.08)	0.089	(0.09;0.09)
	60%	0.105	(0.10;0.11)	0.092	(0.09;0.09)	0.105	(0.10;0.11)	0.152	(0.15;0.15)	0.081	(0.08;0.08)	0.099	(0.10;0.10)
	80%	0.105	(0.10;0.11)	0.092	(0.09;0.09)	0.125	(0.12;0.13)	0.161	(0.16;0.16)	0.085	(0.08;0.09)	0.118	(0.12;0.12)

Table A.2
Results for categorical features. The MAE values and their 95% confidence intervals are presented for each imputation method.

MNAR	Missing	DAE		VAE		kNN		Mean/Mode		MICE		SoftImp	
generation	rate	μ	(95% CI)	μ	(95% CI)	μ	(95% CI)	μ	(95% CI)	μ	(95% CI)	μ	(95% CI)
	10%	0.145	(0.13;0.16)	0.144	(0.13;0.16)	0.139	(0.13;0.15)	0.167	(0.16;0.18)	0.125	(0.12;0.13)	0.156	(0.14;0.17)
	20%	0.146	(0.14;0.16)	0.144	(0.13;0.15)	0.138	(0.13;0.15)	0.172	(0.16;0.18)	0.132	(0.12;0.14)	0.154	(0.14;0.17)
MBUV	40%	0.144	(0.13;0.15)	0.140	(0.13;0.15)	0.140	(0.13;0.15)	0.169	(0.16;0.18)	0.131	(0.12;0.14)	0.159	(0.15;0.17)
	60%	0.141	(0.13;0.15)	0.134	(0.13;0.14)	0.141	(0.13;0.15)	0.171	(0.16;0.18)	0.134	(0.13;0.14)	0.169	(0.16;0.18)
	80%	0.143	(0.13;0.15)	0.136	(0.13;0.14)	0.144	(0.13;0.15)	0.170	(0.16;0.18)	0.139	(0.13;0.15)	0.203	(0.19;0.22)
	10%	0.137	(0.13;0.15)	0.136	(0.12;0.15)	0.133	(0.12;0.14)	0.166	(0.15;0.18)	0.127	(0.12;0.14)	0.136	(0.13;0.15)
	20%	0.128	(0.12;0.14)	0.131	(0.12;0.14)	0.128	(0.12;0.14)	0.140	(0.13;0.15)	0.118	(0.11;0.13)	0.132	(0.12;0.14)
MBIR (Frequentist)	40%	0.130	(0.12;0.14)	0.132	(0.12;0.14)	0.138	(0.13;0.15)	0.188	(0.17;0.20)	0.122	(0.11;0.13)	0.163	(0.15;0.17)
	60%	0.133	(0.12;0.14)	0.131	(0.12;0.14)	0.150	(0.14;0.16)	0.184	(0.17;0.20)	0.135	(0.13;0.14)	0.193	(0.18;0.21)
	80%	0.143	(0.13;0.15)	0.136	(0.13;0.15)	0.184	(0.17;0.20)	0.201	(0.19;0.22)	0.171	(0.16;0.18)	0.205	(0.19;0.22)
	10%	0.144	(0.13;0.16)	0.146	(0.13;0.16)	0.147	(0.14;0.16)	0.179	(0.16;0.20)	0.127	(0.12;0.14)	0.131	(0.12;0.14)
	20%	0.123	(0.11;0.13)	0.117	(0.11;0.13)	0.121	(0.11;0.13)	0.150	(0.14;0.16)	0.111	(0.10;0.12)	0.128	(0.12;0.14)
MBIR (Bayesian)	40%	0.121	(0.11;0.13)	0.122	(0.11;0.13)	0.130	(0.12;0.14)	0.188	(0.17;0.20)	0.118	(0.11;0.13)	0.156	(0.14;0.17)
	60%	0.134	(0.13;0.14)	0.132	(0.12;0.14)	0.157	(0.15;0.17)	0.221	(0.20;0.24)	0.145	(0.14;0.15)	0.188	(0.17;0.20)
	80%	0.142	(0.13;0.15)	0.137	(0.13;0.15)	0.184	(0.17;0.20)	0.202	(0.19;0.22)	0.170	(0.16;0.18)	0.199	(0.18;0.21)

Table A.3
Results for all (mixed) features. The MAE values and their 95% confidence intervals are presented for each imputation method.

MNAR	Missing	DAE		VAE		kNN		Mean/Mode		MICE		SoftImp	
generation	rate	μ	(95% CI)	μ	(95% CI)	μ	(95% CI)	μ	(95% CI)	μ	(95% CI)	μ	(95% CI)
MBUV	10%	0.116	(0.11;0.12)	0.105	(0.10;0.11)	0.087	(0.08;0.09)	0.129	(0.13;0.13)	0.076	(0.07;0.08)	0.092	(0.09;0.10)
	20%	0.118	(0.11;0.12)	0.106	(0.10;0.11)	0.088	(0.09;0.09)	0.131	(0.13;0.13)	0.078	(0.08;0.08)	0.093	(0.09;0.10)
	40%	0.116	(0.11;0.12)	0.104	(0.10;0.11)	0.089	(0.09;0.09)	0.129	(0.13;0.13)	0.077	(0.07;0.08)	0.096	(0.09;0.10)
	60%	0.116	(0.11;0.12)	0.103	(0.10;0.10)	0.091	(0.09;0.09)	0.130	(0.13;0.13)	0.079	(0.08;0.08)	0.103	(0.10;0.11)
	80%	0.116	(0.11;0.12)	0.103	(0.10;0.11)	0.095	(0.09;0.10)	0.130	(0.13;0.13)	0.082	(0.08;0.09)	0.122	(0.12;0.13)
MBIR (Frequentist)	10%	0.123	(0.12;0.13)	0.106	(0.10;0.11)	0.093	(0.09;0.10)	0.144	(0.14;0.15)	0.082	(0.08;0.08)	0.093	(0.09;0.10)
	20%	0.113	(0.11;0.12)	0.099	(0.10;0.10)	0.091	(0.09;0.09)	0.134	(0.13;0.14)	0.076	(0.07;0.08)	0.090	(0.09;0.09)
	40%	0.113	(0.11;0.12)	0.101	(0.10;0.10)	0.102	(0.10;0.10)	0.152	(0.15;0.16)	0.084	(0.08;0.09)	0.103	(0.10;0.11)
	60%	0.111	(0.11;0.11)	0.100	(0.10;0.10)	0.109	(0.11;0.11)	0.154	(0.15;0.16)	0.089	(0.09;0.09)	0.115	(0.11;0.12)
	80%	0.113	(0.11;0.12)	0.101	(0.10;0.10)	0.127	(0.12;0.13)	0.166	(0.16;0.17)	0.103	(0.10;0.11)	0.134	(0.13;0.14)
MBIR (Bayesian)	10%	0.128	(0.12;0.13)	0.114	(0.11;0.12)	0.101	(0.10;0.10)	0.146	(0.14;0.15)	0.084	(0.08;0.09)	0.097	(0.09;0.10)
	20%	0.118	(0.12;0.12)	0.104	(0.10;0.11)	0.095	(0.09;0.10)	0.140	(0.14;0.14)	0.080	(0.08;0.08)	0.094	(0.09;0.10)
	40%	0.112	(0.11;0.11)	0.100	(0.10;0.10)	0.101	(0.10;0.10)	0.148	(0.14;0.15)	0.086	(0.08;0.09)	0.104	(0.10;0.11)
	60%	0.112	(0.11;0.11)	0.101	(0.10;0.10)	0.116	(0.11;0.12)	0.167	(0.16;0.17)	0.095	(0.09;0.10)	0.118	(0.11;0.12)
	80%	0.113	(0.11;0.12)	0.102	(0.10;0.10)	0.137	(0.13;0.14)	0.170	(0.17;0.17)	0.103	(0.10;0.11)	0.136	(0.13;0.14)
MBOUV (Multivariate)	10%	0.141	(0.13;0.15)	0.134	(0.13;0.14)	0.114	(0.11;0.12)	0.196	(0.18;0.21)	0.113	(0.10;0.12)	0.113	(0.10;0.12)
	20%	0.135	(0.13;0.14)	0.136	(0.13;0.14)	0.122	(0.11;0.13)	0.199	(0.19;0.21)	0.144	(0.13;0.16)	0.131	(0.12;0.14)
	40%	0.142	(0.14;0.15)	0.148	(0.14;0.15)	0.149	(0.14;0.16)	0.212	(0.20;0.22)	0.176	(0.17;0.19)	0.156	(0.14;0.17)
	60%	0.142	(0.14;0.15)	0.152	(0.15;0.16)	0.171	(0.16;0.18)	0.225	(0.22;0.23)	0.269	(0.19;0.35)	0.182	(0.17;0.19)
	80%	0.139	(0.13;0.14)	0.150	(0.14;0.16)	0.204	(0.20;0.21)	0.248	(0.24;0.26)	1.017	(−0.25;2.29)	0.237	(0.23;0.25)

References

Ali, N. A., & Omer, Z. M. (2017). Improving accuracy of missing data imputation in data mining. *Kurdistan Journal of Applied Research*, *2*(3), 66–73.

Austin, P. C., White, I. R., Lee, D. S., & van Buuren, S. (2020). Missing data in clinical research: A tutorial on multiple imputation. *Canadian Journal of Cardiology*.

Beaulieu-Jones, B. K., Lavage, D. R., Snyder, J. W., Moore, J. H., Pendergrass, S. A., & Bauer, C. R. (2018). Characterizing and managing missing structured data in electronic health records: data analysis. *JMIR Medical Informatics*, *6*(1), Article e11.

Beaulieu-Jones, B. K., & Moore, J. H. (2017). Missing data imputation in the electronic health record using deeply learned autoencoders. In *Pacific symposium on biocomputing 2017* (pp. 207–218).

Boquet, G., Morell, A., Serrano, J., & Vicario, J. L. (2020). A variational autoencoder solution for road traffic forecasting systems: Missing data imputation, dimension reduction, model selection and anomaly detection. *Transportation Research Part C (Emerging Technologies)*, *115*, Article 102622.

Boquet, G., Vicario, J. L., Morell, A., & Serrano, J. (2019). Missing data in traffic estimation: A variational autoencoder imputation method. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing* (pp. 2882–2886). IEEE.

Bruni, R., Daraio, C., & Aureli, D. (2021). Imputation techniques for the reconstruction of missing interconnected data from higher educational institutions. *Knowledge-Based Systems*, *212*, Article 106512.

Buuren, S. v., & Groothuis-Oudshoorn, K. (2010). Mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 1–68.

Charte, D., Charte, F., García, S., del Jesus, M. J., & Herrera, F. (2018). A practical tutorial on autoencoders for nonlinear feature fusion: Taxonomy, models, software and guidelines. *Information Fusion*, *44*, 78–96.

Choudhury, S. J., & Pal, N. R. (2019). Imputation of missing data with neural networks for classification. *Knowledge-Based Systems*, *182*, Article 104838.

Garciarena, U., & Santana, R. (2017). An extensive analysis of the interaction between missing data types, imputation methods, and supervised classifiers. *Expert Systems with Applications*, *89*, 52–65.

Gondara, L., & Wang, K. (2017). Recovering loss to followup information using denoising autoencoders. In *2017 IEEE international conference on big data (big data)* (pp. 1936–1945).

Gondara, L., & Wang, K. (2018). Mida: Multiple imputation using denoising autoencoders. In *Pacific-Asia conference on knowledge discovery and data mining* (pp. 260–272).

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational Bayes. arXiv preprint arXiv:1312.6114.

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Lin, T., Chen, T., Liu, J., & Tu, X. M. (2021). Extending the mann-whitney-wilcoxon rank sum test to survey data for comparing mean ranks. *Statistics in Medicine*, *40*(7), 1705–1717.

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

McCoy, J. T., Kroon, S., & Auret, L. (2018). Variational autoencoders for missing data imputation with application to a simulated milling circuit. *IFAC-PapersOnLine*, *51*(21), 141–146.

Morey, R. D., Rouder, J. N., Pratte, M. S., & Speckman, P. L. (2011). Using MCMC chain outputs to efficiently estimate Bayes factors. *Journal of Mathematical Psychology*, *55*(5), 368–378.

Pan, R., Yang, T., Cao, J., Lu, K., & Zhang, Z. (2015). Missing data imputation by k nearest neighbours based on grey relational structure and mutual information. *Applied Intelligence*, *43*(3), 614–632.

Peek, N., & Rodrigues, P. P. (2018). Three controversies in health data science. *International Journal of Data Science and Analytics*, *6*(3), 261–269.

Pereira, R. C., Abreu, P. H., & Rodrigues, P. P. (2020). VAE-BRIDGE: Variational autoencoder filter for Bayesian ridge imputation of missing data. In *2020 international joint conference on neural networks* (pp. 1–7).

Pereira, R. C., Santos, M. S., Rodrigues, P. P., & Abreu, P. H. (2019). MNAR imputation with distributed healthcare data. In *EPIA conference on artificial intelligence* (pp. 184–195). Springer.

Pereira, R. C., Santos, M. S., Rodrigues, P. P., & Abreu, P. H. (2020). Reviewing autoencoders for missing data imputation: Technical trends, applications and outcomes. *Journal of Artificial Intelligence Research*, *69*, 1255–1285.

Qiu, Y. L., Zheng, H., & Gevaert, O. (2020). Genomic data imputation with variational auto-encoders. *GigaScience*, *9*(8).

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

Santos, M. S., Abreu, P. H., García-Laencina, P. J., Simão, A., & Carvalho, A. (2015). A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients. *Journal of Biomedical Informatics*, *58*, 49–59.

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

Twala, B. (2009). An empirical comparison of techniques for handling incomplete data using decision trees. *Applied Artificial Intelligence*, *23*(5), 373–405.

Van Buuren, S. (2018). *Flexible imputation of missing data*. Chapman and Hall/CRC.

Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on machine learning* (pp. 1096–1103).

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.-A., & Bottou, L. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion.. *Journal of Machine Learning Research*, *11*(12).

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine*, *30*(4), 377–399.

Xia, J., Zhang, S., Cai, G., Li, L., Pan, Q., Yan, J., & Ning, G. (2017). Adjusted weight voting algorithm for random forests in handling missing values. *Pattern Recognition*, *69*, 52–60.

Zhu, B., He, C., & Liatsis, P. (2012). A robust missing value imputation method for noisy data. *Applied Intelligence*, *36*(1), 61–74.