VAE-BRIDGE: Variational Autoencoder Filter for Bayesian Ridge Imputation of Missing Data

Ricardo Cardoso Pereira
Centre for Informatics and Systems of the University of Coimbra
Dept. of Informatics Engineering
Coimbra, Portugal
rdpereira@dei.uc.pt

Pedro Henriques Abreu
Centre for Informatics and Systems of the University of Coimbra
Dept. of Informatics Engineering
Coimbra, Portugal
pha@dei.uc.pt

Pedro Pereira Rodrigues
Center for Health Technology and Services Research
University of Porto
Porto, Portugal
pprodrigues@med.up.pt

***Abstract*—The missing data issue is often found in real-world datasets and it is usually handled with imputation strategies that replace the missing values with new data. Recently, generative models such as Variational Autoencoders have been applied for this imputation task. However, they were always used to perform the entire imputation, which has presented limited results when comparing to other state-of-the-art methods. In this work, a new approach called Variational Autoencoder Filter for Bayesian Ridge Imputation is introduced. It uses a Variational Autoencoder at the beginning of the imputation pipeline to filter the instances that are later fitted to a Bayesian ridge regression used to predict the new values. The approach was compared to four state-of-the-art imputation methods using 10 datasets from the healthcare context covering clinical trials, all injected with missing values under different rates. The proposed approach significantly outperformed the remaining methods in all settings, achieving an overall improvement between 26% and 67%.**

***Index Terms*—missing data, variational autoencoder, bayesian ridge, data imputation, healthcare data**

I. Introduction

Missing data affects most real-world datasets and it can compromise their many uses. As an example, most machine learning models can not be trained with data containing missing values, and the ones that can usually suffer decreases in their performance. The missing values can be categorized into three mechanisms according to its relation with the data [1], [2]:

- Missing Completely At Random (MCAR), where the missing values are not related to any features;
- Missing At Random (MAR), where the missing values are related to the features available on the dataset;
- Missing Not At Random (MNAR), where the missing values are related with themselves or with other unobserved features.

This missing data issue is usually handled in a preprocessing stage through imputation strategies that generate new plausible values to replace the ones that are missing. Different strategies may present different results for the three missing mechanisms, although in general the results tend to be better for MCAR and MAR [3], [4]. However, in most real-world contexts the mechanism most frequently found is MNAR (e.g., healthcare data), being therefore imperative to improve the imputation results for this missing data type. State-of-the-art approaches for the imputation task include methods such as Multiple Imputation by Chained Equations (MICE), K-Nearest Neighbors (KNN) and specific types of neural networks like the Autoencoders (AE) [5].

A standard AE is a neural network that learns a compressed representation of the data through an unsupervised process, being for that reason widely used for dimensionality reduction. Among its variants, the Denoising Autoencoder (DAE) has been used for missing data imputation. This last one works as a standard AE that learns from a noisy version of the data, being the missing values the noise in this context [6]. However, a recent trend is the use of generative models for the imputation task, particularly Variational Autoencoders (VAE). The VAE is just another AE variant that learns the multi-dimensional parameters of the probability distribution from the input data, namely the mean and variance of a Gaussian function. By sampling from these Gaussian parameters, the model is able to generate new instances following the same distribution, having for that reason generative capabilities [7]. VAEs have been used to address missing data by performing the entire imputation. However, the results are limited and lack significance when comparing to other state-of-the-art methods, with stronger evidence in structured data [8], [9].

In this work a new approach called Variational Autoencoder Filter for Bayesian Ridge Imputation (VAE-BRIDGE) is introduced, and it comprises two main parts:

- A VAE is used in the initial steps of the imputation pipeline to filter the instances that will be considered for the generation of new values (therefore performing a kind of instance selection);
- The final imputation is performed by a Bayesian ridge regression fitted with the filtered instances.

The approach is compared with other state-of-the-art methods in an experiment that uses 10 public datasets from clinical trials that were injected with missing values under the MNAR mechanism. This data context and mechanism were used be-

This work was supported in part by the Portuguese Foundation for Science and Technology (FCT) Research Grant SFRH/BD/149018/2019.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

cause healthcare studies suffer frequently from missing values, and these ones are usually under the MNAR assumptions. Moreover, four different missing rates were considered in the study (10%, 20%, 30% and 40%). The results from the experiment show that the VAE-BRIDGE approach outperformed all the remaining state-of-the-art methods with an overall improvement of 26% and 67% comparing to the second best and worst methods of each dataset, respectively. These results were proved to have a statistically significant of 5% through the Three-Way ANOVA and post-hoc Tukey's HSD tests. To the best of the authors' knowledge, the proposed approach is novel in the missing data field since VAEs were never used for this purpose.

The remainder of the paper is organized in the following way: Section II presents related work of missing data imputation; Section III presents the background concepts needed for this work; Section IV describes the VAE-BRIDGE approach here proposed; Section V describes the experiment and the obtained results; and Section VI shows the conclusions and future directions of this work.

II. Related Work

The use of VAEs to address missing data imputation is rather recent, and just a couple of works have yet used this method. McCoy *et al.* [8] conducted an experiment where the imputation of missing values was performed using a VAE, together with the mean of the feature and the Principal Component Analysis (PCA) method. Both the VAE and PCA were used with a multiple imputation strategy, where they would make predictions iteratively until the reconstruction error was below a given threshold. The Root Mean Squared Error (RMSE) was used as the reconstruction metric. The methods were applied to a synthetic nonlinear dataset and to a simulated milling circuit dataset, with missing rates of 20% and 90% (the missing mechanisms were not specified). The VAE outperformed the remaining methods in all settings, although achieving RMSE differences of less than 0.1 when compared to the mean imputation in the simulated dataset.

Boquet *et al.* [9] proposed a method that uses a VAE for imputation and connects its output directly to a standard neural network for regression, aiming to solve the missing data issue before performing traffic forecasting. The VAE is compared with a standard AE and the PCA method, but the imputation is not assessed directly. Instead, only the forecast error is evaluated through the RMSE metric and the Mean Absolute Percentage Error (MAPE). The experiments used real traffic data from the freeway Performance Measurement System of the California Department of Transportation, and missing values under the MNAR and MCAR mechanisms were injected (the last one with rates of 10%, 20% and 40%). The forecasting results when the VAE was used outperformed the remaining methods, with improvements of at least 40% for MNAR and 17% for MCAR.

When considering strategies to perform instance filtering with the goal of improving the imputation results, no works were yet developed using VAEs. Tsai and Chang [10] published an experiment where the DROP3 algorithm was used to perform instance selection before and after the missing values were imputed with the KNN algorithm (four different workflows were tested). However, the instance selection is not related with the imputation procedure and does not aim to improve it. In fact, the results are assessed only through the impact of the entire workflow on the accuracy of two classifiers (Support Vector Machine and KNN). The experiment was conducted on 29 public datasets from different contexts, covering all attribute types. The results show that the improvements were limited, with differences of less than 1% in several settings, and with KNN presenting the best accuracy for numerical and mixed datasets. Huang *et al.* [11] presented a very similar study, with only a few differences: the instance selection is made by two methods (DROP3 e IB3) and is always performed before the imputation. Moreover, only the KNN classifier is considered. The experiment used eight public datasets from the medical context, containing once again all attribute types. The accuracy results of the KNN classifier were in general better when the instance selection was used, with the improvements varying between less than 1% and 18%.

In conclusion, only a couple of works have yet used VAEs to solve the missing data problem, and both present issues: the imputation was not properly assessed since the baseline methods used for comparison were not state-of-the-art [8] or the imputation error was not even measured [9]. Moreover, the use of VAEs for filtering purposes prior to the final imputation was never done, which is one of the main novel aspects from this work.

III. Background

The VAE-BRIDGE approach here proposed is based on two existing concepts: Variational Autoencoders and Bayesian ridge regressions. Both methods are described in the following sections.

A. Variational Autoencoder

A VAE is a type of Autoencoder that has generative capabilities. While a standard AE learns a compressed representation of the input data, the VAE is able to learn the parameters of a Gaussian distribution which describes the data. Therefore, both variants present similar characteristics with a few key differences [7]:

- The latent space of an AE is the output of a regular hidden layer, usually with less units than the input layer, which makes this model useful for dimensionality reduction. On the other hand, a VAE latent space is the multidimensional parameters of a Gaussian distribution (mean and variance), which are then used to generate new samples with the same characteristics (see Figure 1);
- While the AE loss function is only the reconstruction error, the VAE adds another term for regularization (see Equation 1, where $q(z|X)$ is the encoder output, $p(X|z)$ is the decoder output, X is the input data and z represents the new samples from the learned distribution). The

reconstruction error is needed for the model to learn how to reconstruct the data, and can be, for example, the Mean Squared Error. The regularizer is the Kullback-Leibler divergence between the encoder and decoder distributions, and is needed to ensure the latent space is correctly structured (i.e., similar input data should have similar latent space representations).

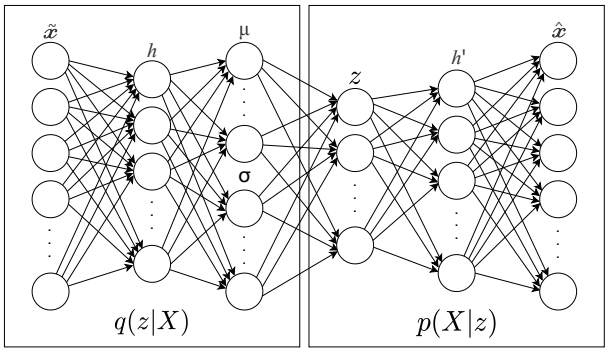

Fig. 1. General architecture of a Variational Autoencoder.

$$L(X) = -E_{z\sim Q(z|X)}[log\, p(X|z)] + KL(q(z|X) \parallel p(z)) \quad (1)$$

B. Bayesian Ridge Regression

The frequentist approach to perform linear regression is based on assigning a weight/coefficient to each independent variable, reflecting therefore their effect on the dependent variable. An error term is also considered to account for noise and other external factors. The model is usually fitted through the Ordinary Least Squares (OLS) approach, which minimizes the residual sum of squares (see Equation 2, were y is the dependent variable, w the coefficients and x_i one of the M instances) [12]. In other words, the goal is to minimize the model error by adjusting the w coefficients.

$$min \sum_{i=1}^{M} (y_i - w^T x_i) \quad (2)$$

A problem found in this type of regression is overfitting, particularly when data displays multicollinearity patterns. To mitigate this issue regularization is often used. A common strategy is to penalize the size of the w coefficients with a L2 regularizer, creating the so-called ridge regression [12], [13].

A limitation of the OLS method is the w coefficient values having single point estimates, meaning these values are the ones most likely to be correct given the training data. However, the uncertainty surrounding the model results are not accounted at all.

The Bayesian approach addresses this issue by modeling the regression with probability distributions instead of single value estimates. Assuming the use of a Gaussian distribution, the formulation of the Bayesian regression is presented in Equation 3 (using the same definitions from Equation 2, with X being a matrix with all M instances) [12], [13].

$$y \sim \mathcal{N}(w^T X, \sigma^2) \quad (3)$$

The outcome will therefore be the posterior distribution of the w coefficients, instead of their exact values. This change allows the use of priors, which can be helpful if relevant information about the model is already known. Moreover, the model uncertainty is accounted for, since the posterior distribution gives a range of possible w coefficients based on the data and the prior [12]. This last aspect is particularly relevant when the amount of instances used to fit the model is reduced. In fact, when the number of instances increases the w coefficients converge to the ones obtain from the OLS method, since the level of uncertainty is decreasing. Moreover, when the prior also follows a Gaussian distribution, the L2 regularization is implicitly applied, creating the Bayesian ridge regression concept [12], [14].

One of the most common ways to fit a Bayesian regression is drawing samples from the posterior distribution to improve and approximate it, using, for example, Monte Carlo methods. Another common approach is to use the Maximum A Posterior (MPA) method [12], [14].

IV. Variational Autoencoder Filter for Bayesian Ridge Imputation

The VAE-BRIDGE approach starts by training a VAE with all data instances that are complete, but excluding the feature containing missing values (meaning that no pre-imputation is required). The model will learn the multidimensional parameters of the Gaussian distribution that represents the data, which will then be used for filtering purposes. Afterwards, each instance having missing values is encoded with the previously trained VAE, and its multidimensional Gaussian parameters are compared to the ones from each complete instance. The goal is to obtain the k percent instances that are described by the most similar Gaussian distributions, following the formula from Equation 4. This distance is an adaptation of the euclidean metric to include both Gaussian parameters (mean and variance).

$$d_{p,q} = \sqrt{\sum_{i=1}^{n} (\mu_{p_i} - \mu_{q_i})} + \sqrt{\sum_{i=1}^{n} (\sigma^2{}_{p_i} - \sigma^2{}_{q_i})} \quad (4)$$

The selected k percent instances are finally used to fit a Bayesian ridge regression. Any imputation model could be used in this step, but the Bayesian ridge is known to provide better long term predictions through regularization strategies that deal with overfitting (see Section III-B), being for that reason used by state-of-the-art methods such as MICE [15]. An additional aspect of this regression model is the easiness to interpret its results, something that is often important in sensitive contexts (e.g., healthcare). For datasets with more than two features this regression will be multivariate, with all the features without missing values being the independent variables.

For a proper generalization of this method the following aspects must also be considered:

- If the missing data scenario is multivariate (i.e., two or more features have missing values), the described process must be repeated individually for each of these features;
- To avoid issues with the domain of the features and to speed up the VAE training convergence, all features should be normalized within $[0, 1]$. Consequently, the VAE output layer should use sigmoid as the activation function;
- Categorical features must be transformed to binary ones through a one-hot encoding process, otherwise the VAE training procedure and the distance formula from Equation 4 will not be valid. The imputation of these features will be a real value within $[0, 1]$ (as the previous point states), which can be converted to a binary value assuming a fixed threshold (e.g., 0.5).

The complete VAE-BRIDGE approach is summarized in Figure 2.

Input: $complete\_rows$, $incomplete\_rows$, k
Output: $imputed\_rows$

Normalize all data within $[0, 1]$
Apply one-hot encoding to the categorical features
for each feature having missing values (md_i) **do**
  Train a VAE with $complete\_rows \setminus md_i$
  $enc\_data \leftarrow$ Encode $complete\_rows \setminus md_i$ with VAE
  for each instance z in $incomplete\_rows$ **do**
    $enc\_z \leftarrow$ Encode $z \setminus md_i$ with VAE
    $sim\_z \leftarrow$ Find k similar rows to enc_z from enc_data using the distance formula from $Eq. 1$
    $br \leftarrow$ Fit a Bayesian ridge regression with sim_z
    $imputed\_rows \leftarrow$ Predict missing data in z with br
  end for
end for
return $imputed\_rows$

Fig. 2. Pseudocode of the VAE-BRIDGE algorithm.

The key aspect of this method is the filtering step based on the VAE encoding capabilities. By choosing only the k percent instances that have similar distribution parameters to the one that is being imputed, the method ensures that noisy data not relevant for the imputation task is ignored by the Bayesian ridge regression during the final prediction step. Therefore, for this reason, different k values will have a major impact on the approach results. To properly understand this impact a sensitivity analysis study was conducted. The experiment used the well-known Breast Cancer Wisconsin dataset, with missing values under MNAR with different rates (10%, 20% and 30%). Only this dataset was used given the number of variables to consider in the study, and the choice was based on its popularity among the healthcare public datasets. The focus on the MNAR mechanism is justified by the fact that most missing values in healthcare are usually under MNAR assumptions, as previously stated. The results were evaluated through the Mean Absolute Error (MAE). Figure 3 presents the conclusions of the study, where each bar shows the average number of times each k value presented the best MAE results for the three missing rates. The study considered 10 different percentages for k (10% to 100%, with 10% steps).

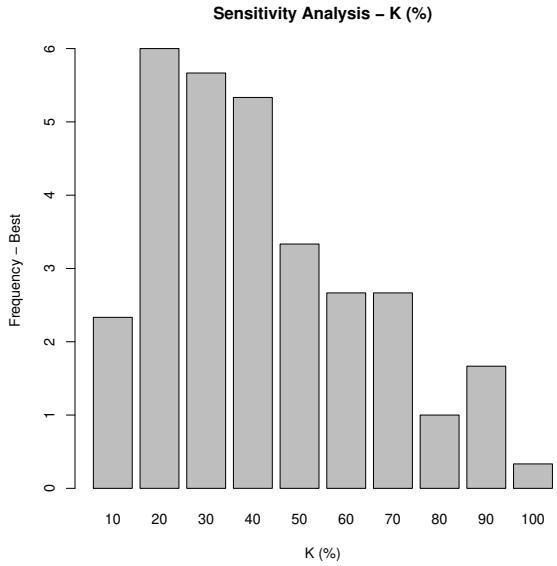

Fig. 3. Sensitivity analysis of the k parameter. Each bar shows the average number of times each k value presented the best MAE results for the three missing rates.

From the obtained results, the k values $\{20, 30, 40\}$ clearly outperformed the remaining ones, with $k = 20$ presenting the best overall score. Such results not only show the best k values to be used, but also prove the impact of using only relevant data for the imputation. In general, when k increases over the 40% threshold, the results tend to be worse. This can be explained by the fact that the Bayesian ridge regression is using more data that is not relevant to the imputation, which creates additional noise in the fitting process. Using $k = 10$ also presented bad results since the filter probably discarded more information than it should, leading to the lost of data that was in fact relevant. Therefore, the used k value must neither be too small or too big, with percentages between 20 and 40 showing a good balance for this criteria. Nevertheless, to have accurate k values this study should be conducted for the different datasets. However, given the computation complexity of this task, only the best values obtained for the Wisconsin dataset ($k = \{20, 30, 40\}$) were used for the remaining ones in the experiments.

V. Experimental Results

To evaluate the quality of the imputation performed by the VAE-BRIDGE method an experiment was conducted, aiming to compare it to the following state-of-the-art methods:

- Standard VAE, which performed the entire imputation [8];
- Denoising Autoencoder (DAE), which is a basic discriminative Autoencoder that learns a compressed represen-

tation of the input data with noise (in this context the missing values are the noise) [6];
- Multiple Imputation by Chained Equations (MICE), which uses a multiple imputation approach to fit regression models using the features with missing values as the dependent variables [15];
- K-Nearest Neighbors (KNN) Imputation with $K = 5$, which finds the five nearest neighbors of an instance through a distance metric (usually the Euclidean one) and uses their average values to impute the missing data [16].

Regarding the Autoencoder-based methods (VAE-BRIDGE, VAE and DAE), they all used a similar architecture:

- A single hidden layer with 100 units and the ReLU activation function for the DAE, while the VAEs had two "parallel" layers (mean and variance) with 100 units and linear activation;
- The optimization algorithm used was Adam with a learning rate of 0.001, batches of 64 instances and 200 epochs;
- To avoid overfitting each layer applies the L2 regularizer with a factor of 0.01;
- The weights of the layers are initialized using the glorot normal approach, which follows a truncated normal distribution centered on zero;
- The output layer always uses the sigmoid activation function, as explained in the previous section.

Moreover, the VAE-BRIDGE approach used $k = \{20, 30, 40\}\%$, since these were the best overall values obtained from the sensitivity analysis study presented in the previous section. However, the best results were once again obtained for $k = 20$, being therefore the ones presented here.

The experiment considered 10 public datasets from the healthcare context, covering clinical studies of different pathologies. The choice of this context lies in the fact that it often suffers from missing data, which compromises severely the studies' results [17]. All datasets were obtained from the UCI repository (available at https://archive.ics.uci.edu/ml/datasets.php), and they have different sizes and both continuous and categorical features, as Table I shows.

TABLE I
CHARACTERISTICS OF THE DATASETS USED IN THE EXPERIMENTS.

Dataset	# Instances	# Features	
		Continuous	Categorical
wisconsin	569	31	0
ctg	2126	21	2
pima	768	9	0
liver	583	10	1
hcv-egy	1385	19	10
parkinsons	195	23	0
bc-coimbra	116	10	0
thoracic-surgery	470	14	3
spine	310	13	0
mammographic-masses	830	2	4

All datasets were complete and were latter injected with missing values under the MNAR mechanism, using the method from [18] (the lowest values are set to be missing). This missing mechanism was used since it is the one more often found in healthcare contexts, and it is also the one that poses more challenges for the imputation task [17]. Each feature of each dataset was iteratively injected with missing data and imputed, with the imputation quality being assessed through the Mean Absolute Error (MAE) metric, calculated between the ground truth and the imputed data. The final MAE of a dataset is the average MAE of all its features.

The data was normalized within $[0, 1]$ and split in train, validation and test sets (with 60%-20%-20% proportions) for all methods except the KNN, since it does not require training. Four missing rates were considered (10%, 20%, 30% and 40%), with the missing values being pre-imputed with the mean for the DAE and standard VAE methods. Notice that the missing values were injected independently for the train, validation and test sets, in order to ensure the same missing rate and MNAR assumptions for each one. To mitigate bias and stochastic behaviors, the experiment was executed five independent times, with the datasets being shuffled in each run. Moreover, the average results from the runs were considered for comparison. A graphical representation of the experimental setup here described is presented in Figure 4.

The results from the experiment are presented in Table II. The VAE-BRIDGE approach outperformed all state-of-the-art methods for every dataset and missing rate. The overall improvement from the second best and worst methods to VAE-BRIDGE was 26% and 67%, respectively. The second best method varied between VAE and MICE depending on the dataset, while the DAE presented the overall worst results.

To assess the statistical significance of the results, the Three-Way ANOVA test was applied with a significance level of 5%. The factors considered were the dataset, the missing rate and the algorithm, while the dependent variable was the MAE. This statistical test can only be applied if the data follows a normal distribution, which was confirmed visually through the Q-Q plot presented in Figure 5. Moreover, the same analysis was conducted on the data subgroups. The *p*-values from the test are presented in Table III, and they show the results are statistically significant for the datasets and algorithms ($p = 2e - 16$). However, no evidence of sensitivity to the missing rates was found ($p = 0.365$).

To conclude if the VAE-BRIDGE approach outperformed the remaining algorithms with a statistical significance of 5%, the post-hoc Tukey's HSD test was applied for this particular factor. The *p*-values from this test are displayed in Table IV. The VAE-BRIDGE significantly outperformed VAE ($p = 0.006732$), DAE ($p < 0.000001$), MICE ($p = 0.000001$) and KNN ($p < 0.000001$).

VI. CONCLUSIONS

In this work a new approach for missing data imputation called Variational Autoencoder Filter for Bayesian Ridge Imputation (VAE-BRIDGE) is introduced. It uses a VAE to filter the instances that are relevant for the imputation, while a Bayesian ridge regression is fitted with them and predicts the new values. The method relies on the fact that instances that are not relevant for the imputation may compromise its results

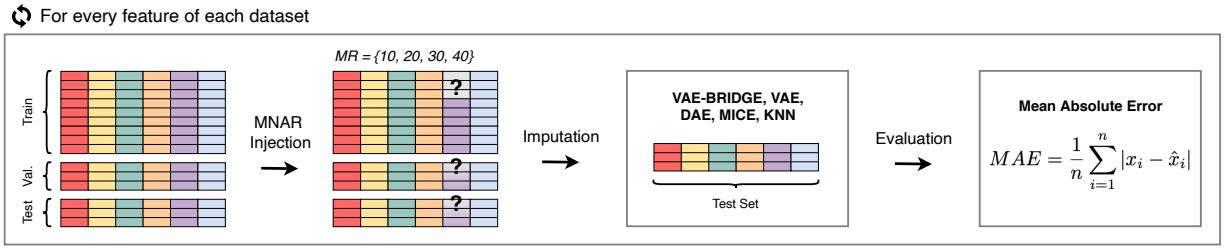

Fig. 4. Graphical representation of the experimental setup used in this work.

TABLE II

RESULTS FROM THE EXPERIMENT. THE FIRST TWO COLUMNS IDENTIFY THE DATASET AND THE MISSING RATE PERCENTAGE. THE NEXT FIVE COLUMNS PRESENT THE MAE VALUES (MEAN AND STANDARD DEVIATION) FOR THE USED METHODS. THE LAST COLUMN SHOWS THE PERCENTAGE IMPROVEMENT FROM THE BEST METHOD TO THE SECOND BEST. THE BEST RESULTS FOR EACH DATASET AND MISSING RATE ARE BOLDED.

Dataset	MR (%)	VAE-BRIDGE	VAE	DAE	MICE	KNN	↑ %
wisconsin	10	**0.019 ± 0.019**	0.078 ± 0.037	0.213 ± 0.049	0.03 ± 0.034	0.074 ± 0.037	37%
	20	**0.017 ± 0.018**	0.062 ± 0.028	0.189 ± 0.045	0.028 ± 0.034	0.079 ± 0.037	39%
	30	**0.017 ± 0.017**	0.055 ± 0.023	0.173 ± 0.044	0.029 ± 0.036	0.085 ± 0.038	41%
	40	**0.017 ± 0.017**	0.052 ± 0.021	0.158 ± 0.045	0.032 ± 0.042	0.094 ± 0.037	47%
ctg	10	**0.028 ± 0.048**	0.051 ± 0.063	0.198 ± 0.149	0.043 ± 0.072	0.06 ± 0.092	35%
	20	**0.025 ± 0.043**	0.042 ± 0.05	0.179 ± 0.133	0.043 ± 0.076	0.065 ± 0.086	40%
	30	**0.024 ± 0.041**	0.037 ± 0.045	0.161 ± 0.106	0.066 ± 0.139	0.091 ± 0.142	35%
	40	**0.023 ± 0.038**	0.035 ± 0.042	0.148 ± 0.09	0.065 ± 0.118	0.094 ± 0.125	34%
pima	10	**0.161 ± 0.077**	0.176 ± 0.075	0.237 ± 0.078	0.195 ± 0.093	0.177 ± 0.1	9%
	20	**0.136 ± 0.052**	0.142 ± 0.043	0.208 ± 0.065	0.189 ± 0.067	0.174 ± 0.069	4%
	30	**0.124 ± 0.051**	0.127 ± 0.032	0.191 ± 0.069	0.203 ± 0.073	0.188 ± 0.071	2%
	40	**0.112 ± 0.05**	0.117 ± 0.024	0.173 ± 0.074	0.216 ± 0.082	0.199 ± 0.069	4%
liver	10	**0.096 ± 0.19**	0.158 ± 0.196	0.28 ± 0.223	0.117 ± 0.219	0.131 ± 0.201	18%
	20	**0.09 ± 0.187**	0.139 ± 0.185	0.263 ± 0.225	0.127 ± 0.25	0.145 ± 0.235	29%
	30	**0.083 ± 0.178**	0.123 ± 0.171	0.238 ± 0.203	0.198 ± 0.321	0.218 ± 0.31	33%
	40	**0.073 ± 0.153**	0.108 ± 0.146	0.212 ± 0.171	0.167 ± 0.241	0.189 ± 0.234	32%
hcv-egy	10	**0.141 ± 0.202**	0.155 ± 0.181	0.347 ± 0.163	0.156 ± 0.224	0.303 ± 0.142	9%
	20	**0.127 ± 0.181**	0.132 ± 0.159	0.333 ± 0.156	0.162 ± 0.229	0.331 ± 0.137	4%
	30	**0.113 ± 0.16**	0.117 ± 0.142	0.319 ± 0.152	0.167 ± 0.234	0.372 ± 0.142	3%
	40	**0.098 ± 0.138**	0.106 ± 0.126	0.303 ± 0.151	0.169 ± 0.237	0.441 ± 0.181	8%
parkinsons	10	**0.066 ± 0.089**	0.134 ± 0.131	0.284 ± 0.104	0.096 ± 0.145	0.108 ± 0.092	31%
	20	**0.059 ± 0.078**	0.117 ± 0.112	0.254 ± 0.094	0.103 ± 0.162	0.132 ± 0.122	43%
	30	**0.053 ± 0.068**	0.103 ± 0.089	0.231 ± 0.079	0.107 ± 0.156	0.15 ± 0.145	49%
	40	**0.049 ± 0.06**	0.094 ± 0.074	0.21 ± 0.07	0.105 ± 0.133	0.159 ± 0.12	48%
bc-coimbra	10	**0.142 ± 0.106**	0.188 ± 0.109	0.269 ± 0.069	0.185 ± 0.132	0.172 ± 0.116	17%
	20	**0.131 ± 0.088**	0.16 ± 0.085	0.243 ± 0.054	0.184 ± 0.125	0.183 ± 0.108	18%
	30	**0.122 ± 0.079**	0.15 ± 0.079	0.229 ± 0.047	0.197 ± 0.129	0.197 ± 0.113	19%
	40	**0.11 ± 0.071**	0.141 ± 0.07	0.208 ± 0.041	0.211 ± 0.133	0.212 ± 0.112	22%
thoracic-surgery	10	**0.084 ± 0.166**	0.114 ± 0.162	0.259 ± 0.211	0.096 ± 0.191	0.122 ± 0.194	13%
	20	**0.077 ± 0.148**	0.099 ± 0.142	0.254 ± 0.208	0.1 ± 0.192	0.139 ± 0.217	22%
	30	**0.07 ± 0.127**	0.087 ± 0.117	0.244 ± 0.191	0.135 ± 0.233	0.159 ± 0.237	20%
	40	**0.062 ± 0.108**	0.079 ± 0.097	0.232 ± 0.171	0.158 ± 0.249	0.178 ± 0.248	22%
spine	10	**0.274 ± 0.219**	0.326 ± 0.154	0.363 ± 0.137	0.297 ± 0.231	0.359 ± 0.191	8%
	20	**0.25 ± 0.205**	0.289 ± 0.157	0.326 ± 0.14	0.31 ± 0.251	0.362 ± 0.213	13%
	30	**0.219 ± 0.186**	0.251 ± 0.136	0.285 ± 0.134	0.323 ± 0.277	0.368 ± 0.227	13%
	40	**0.191 ± 0.162**	0.218 ± 0.117	0.249 ± 0.12	0.313 ± 0.252	0.36 ± 0.201	12%
mammographic-masses	10	**0.025 ± 0.072**	0.05 ± 0.064	0.242 ± 0.175	0.056 ± 0.161	0.072 ± 0.154	50%
	20	**0.022 ± 0.065**	0.044 ± 0.058	0.229 ± 0.153	0.045 ± 0.116	0.062 ± 0.111	50%
	30	**0.02 ± 0.062**	0.041 ± 0.059	0.223 ± 0.149	0.042 ± 0.108	0.059 ± 0.1	51%
	40	**0.02 ± 0.059**	0.04 ± 0.059	0.219 ± 0.149	0.045 ± 0.119	0.071 ± 0.113	50%

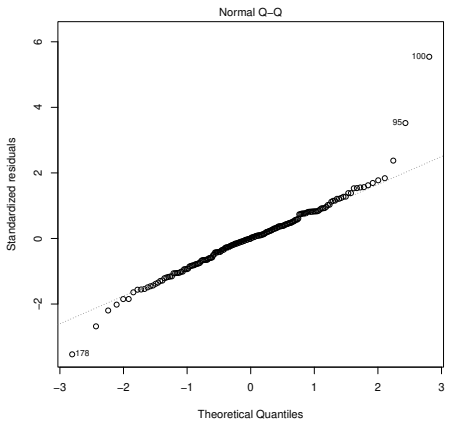

Fig. 5. Normal Q-Q plot showing that the data follows a normal distribution.

TABLE III
P-VALUES OF THE THREE-WAY ANOVA TEST.

Factor	***P*-value**
Algorithm	2e-16
Missing Rate	0.365
Dataset	2e-16

TABLE IV
P-VALUES OF THE TUKEY'S HSD POST-HOC TEST.

	VAE	**DAE**	**MICE**	**KNN**
VAE-BRIDGE	0.006732	0.000000	0.000001	0.000000

by adding unnecessary noise. The approach was compared with four state-of-the-art methods (standard VAE, DAE, MICE and KNN) in an experiment with 10 datasets from clinical trials that were injected with missing values under MNAR (including 10%, 20%, 30% and 40% missing rates). The VAE-BRIDGE approach outperformed all the remaining methods, achieving an overall improvement between 26% and 67%. The results were validated with a statistical significance of 5%.

In the future new experiments will be conducted to test the VAE-BRIDGE approach with datasets from other contexts (some containing pre-existent missing values) and with the remaining missing data mechanisms (MCAR and MAR), considering in this last scenario the impact of having different mechanisms in the train, validation and test sets. Moreover, a study on the impact of the imputation in classification tasks will also be conducted.

REFERENCES

[1] M. S. Santos, J. P. Soares, P. Henriques Abreu, H. Araújo, and J. Santos, "Influence of Data Distribution in Missing Data Imputation," in *International Conference on Artificial Intelligence in Medicine in Europe*, 2017, pp. 285–294.

[2] M. S. Santos, R. C. Pereira, A. F. Costa, J. P. Soares, J. Santos, and P. H. Abreu, "Generating Synthetic Missing Data: A Review by Missing Mechanism," *IEEE Access*, vol. 7, pp. 11 651–11 667, 2019.

[3] S. M. van Kuijk, W. Viechtbauer, L. L. Peeters, and L. Smits, "Bias in regression coefficient estimates when assumptions for handling missing data are violated: a simulation study," *Epidemiology, Biostatistics and Public Health*, vol. 13, no. 1, 2016.

[4] U. Garciarena and R. Santana, "An extensive analysis of the interaction between missing data types, imputation methods, and supervised classifiers," *Expert Systems with Applications*, vol. 89, pp. 52–65, 2017.

[5] W.-C. Lin and C.-F. Tsai, "Missing value imputation: a review and analysis of the literature (2006–2017)," *Artificial Intelligence Review*, pp. 1–23, 2019.

[6] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th International Conference on Machine learning*, 2008, pp. 1096–1103.

[7] D. P. Kingma and M. Welling, "Auto-encoding Variational Bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[8] J. T. McCoy, S. Kroon, and L. Auret, "Variational autoencoders for missing data imputation with application to a simulated milling circuit," *IFAC-PapersOnLine*, vol. 51, no. 21, pp. 141–146, 2018.

[9] G. Boquet, J. L. Vicario, A. Morell, and J. Serrano, "Missing data in traffic estimation: A variational autoencoder imputation method," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2882–2886.

[10] C.-F. Tsai and F.-Y. Chang, "Combining instance selection for better missing value imputation," *Journal of Systems and Software*, vol. 122, pp. 63–71, 2016.

[11] M.-W. Huang, W.-C. Lin, C.-W. Chen, S.-W. Ke, C.-F. Tsai, and W. Eberle, "Data preprocessing issues for incomplete medical datasets," *Expert Systems*, vol. 33, no. 5, pp. 432–438, 2016.

[12] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT press, 2012.

[13] M. H. Kutner, C. J. Nachtsheim, J. Neter, W. Li *et al.*, *Applied linear statistical models*. McGraw-Hill Irwin New York, 2005, vol. 5.

[14] M. E. Tipping, "Sparse bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, no. Jun, pp. 211–244, 2001.

[15] S. v. Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in r," *Journal of Statistical Software*, pp. 1–68, 2010.

[16] M. S. Santos, P. H. Abreu, P. J. García-Laencina, A. Simão, and A. Carvalho, "A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients," *Journal of Biomedical Informatics*, vol. 58, pp. 49–59, 2015.

[17] R. C. Pereira, M. S. Santos, P. P. Rodrigues, and P. H. Abreu, "MNAR Imputation with Distributed Healthcare Data," in *EPIA Conference on Artificial Intelligence*. Springer, 2019, pp. 184–195.

[18] B. Twala, "An empirical comparison of techniques for handling incomplete data using decision trees," *Applied Artificial Intelligence*, vol. 23, no. 5, pp. 373–405, 2009.