Journal of King Saud University – Computer and Information Sciences 35 (2023) 101562

HOSTED BY

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

Addressing the Curse of Missing Data in Clinical Contexts: A Novel Approach to Correlation-based Imputation

Isabel Curioso [a,b], Ricardo Santos [a,b,*], Bruno Ribeiro [a], André Carreiro [a], Pedro Coelho [c,d], José Fragata [c,d], Hugo Gamboa [a,b]

[a] Associação Fraunhofer Portugal Research, Rua Alfredo Allen 455/461, 4200-135 Porto, Portugal
[b] Laboratory for Instrumentation, Biomedical Engineering and Radiation Physics (LIBPhys-UNL), Physics Department, NOVA School of Science and Technology, 2829-516 Caparica, Portugal
[c] Comprehensive Health Research Center, NOVA Medical School, Campo Mártires da Pátria, 130, 1169-056 Lisboa, Portugal
[d] Hospital de Santa Marta, Centro Hospitalar Universitário Lisboa Central, Rua de Santa Marta, 50, 1169-023 Lisboa, Portugal

ARTICLE INFO

Article history:
Received 13 December 2022
Revised 16 March 2023
Accepted 13 April 2023
Available online 25 April 2023

Keywords:
Missing data
Missing data imputation
Correlation
Clinical data
Machine learning

ABSTRACT

Clinical data are essential in the medical domain. However, their heterogeneous nature leads to many data quality problems, notably missing values, which undermine the performance of Machine Learning-based clinical systems. Hence, there has been a growing interest in strategies that address this challenge in order to build trustworthy systems to improve the quality of care and benefit clinical decision-making. In particular, missing value imputation is a common approach. This paper proposes three novel imputation techniques that leverage correlation in an innovative manner by exploring the relationship between values and missingness patterns. Experiments were carried out on three publicly available datasets, under three missingness mechanisms with different missing rates, and on two real-world medical datasets. The imputation precision and the classification performance of the proposed techniques were evaluated in a comprehensive comparative study, which included diverse existing methods. The developed techniques outperformed state-of-the-art methods on several assessments while overcoming current flaws shared by correlation-based imputation strategies in real-world medical problems.

1. Introduction

A growing and ageing population entails a broad amount of clinical information that needs to be organised and easily accessed by professionals. Since the analysis of paper-based data is time-consuming, digitalisation has emerged as a vital process towards optimising health care management, accompanied by the rise of valuable enablers such as Electronic Health Records (EHRs) (Ambinder, 2005).

An EHR contains thorough clinical information and can thus facilitate knowledge extraction. However, establishing relationships within the data to formulate a medical diagnosis still mostly relies on the physicians' experience. Artificial Intelligence (AI) is suitable for discovering patterns in vast datasets, an ability that could support and benefit clinical decision-making. Therefore, AI-based clinical systems will become an important instrument to assist professionals, leveraging all available information from the patient's journey.

Moreover, EHRs mirror the heterogeneous nature of clinical data, often collected through different procedures and stored in distinct formats. Unfortunately, with this variability also comes inconsistency. In fact, most real-world datasets are incomplete, which yields deleterious effects on Machine Learning (ML) models built thereon. In healthcare, reliable ML-based systems must be able to cope with missing values since their performance may influence clinical decision-making (Iranfar et al., 2021; Kang and Tian, 2018). This concern, along with the ubiquity of missing data

* Corresponding author at: Associação Fraunhofer Portugal Research, Rua Alfredo Allen 455/461, 4200-135 Porto, Portugal.
E-mail addresses: isabel.curioso@fraunhofer.pt (I. Curioso), ricardo.santos@fraunhofer.pt (R. Santos), bruno.ribeiro@fraunhofer.pt (B. Ribeiro), andre.carreiro@fraunhofer.pt (A. Carreiro), pedro.coelho@chlc.min-saude.pt (P. Coelho), jose.fragata@nms.unl.pt (J. Fragata), hugo.gamboa@fraunhofer.pt (H. Gamboa).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2023.101562
1319-1578/

in real-world databases, prompted a growing interest in developing strategies that address this challenge, particularly missing value imputation techniques.

In recent years, some authors have developed techniques that account for correlation when imputing missing values, stating that such a choice is beneficial. Even though correlation cannot assure a causal relationship between missingness and its source, it may still provide helpful insights for predicting missing data. The concept of correlation gains relevance in clinical datasets, where distinct features are frequently different manifestations of the same physiological event or medical condition, consequently exhibiting a significant level of dependency. Although promising, the study of correlation within the context of missing value imputation is still scarce in biomedical research.

Therefore, this work exploits the correlation between attributes to address the challenges posed by missing data in real-world medical datasets, aiming to improve the robustness of ML-based systems. This paper contributes to the State of the Art with three novel correlation-based imputation techniques, which leverage not only the correlation between values but also the correlation between values and missingness patterns, an innovative and unique strategy. These techniques overcome the limitations of existing imputation methods in terms of their dependency on a complete data subset. Furthermore, a comprehensive comparative study was conducted to evaluate the effectiveness of the developed techniques.

The remaining of this paper is organised as follows. Related works, particularly state-of-the-art correlation-based imputation techniques, are briefly presented in Section 2. Section 3 covers materials and methods, including missingness mechanisms and correlation. Section 4 provides detailed descriptions of the proposed imputation techniques. Section 5 introduces the five datasets used throughout this work and the experimental setup. The obtained results are presented and discussed in Section 6. Lastly, Section 7 concludes this paper by reviewing its main findings along with perspectives for future work.

2. Related work

There are two main approaches to address the challenges imposed by missing data, namely deletion and imputation.

The deletion methods, or techniques for ignoring missing data, are straightforward procedures based on completely recorded samples, like the one performed by Zhou et al. (2022). Despite its convenience, deletion should be used heedfully since it may introduce bias in the analysis (Enders, 2022).

As for imputation methods, their key purpose is to fill in, i.e. replace, the missing elements with predicted values, usually estimated from the observed data. The most common statistical and ML-based methods include mean imputation, regression imputation, stochastic regression imputation, hot-deck imputation, and K-Nearest Neighbours (KNN) imputation (Enders, 2022). Moreover, various works resort to multiple imputation approaches, such as the Multivariate Imputation by Chained Equations (MICE) algorithm developed by Van Buuren and Groothuis-Oudshoorn (2011). Furthermore, the so-called data splitting-based techniques have also gained popularity in literature, as stated by Bhagat and Singh (2022). These authors presented the Nullify the Missing Values before Imputation (NMVI) method, which divides the dataset into classes and defines an upper limit for each class that will then be included in the imputation procedure.

As previously mentioned, several authors have recently turned their attention to correlation-based imputation techniques. Mishra et al. (2021) proposed an imputation method that replaces the missing elements in an attribute with predictions from regression models trained with features that are highly correlated with the incomplete attribute. Sefidian and Daneshpour (2020) also presented regression-based algorithms, called Correlation Maximisation-based Imputation Methods (CMIM), which attempted to maximise the correlation between the missing attributes and the remaining ones. Liu et al. (2019) developed the Correlation-based Hierarchical KNN (CoHiKNN) method, a KNN-based algorithm that utilises the correlation between attributes as weights to compute the distance between each incomplete record and all complete records. Khan et al. (2022) proposed the Convolutional Neural Network Imputation (CNNI) approach, which identifies existing correlations within a dataset to train a convolutional kernel that will replace all missing values. In regards to approaches applied to real-world clinical datasets, Yoon et al. (2019) developed a DL model that exploits the relationship among features, particularly the correlation within and across data streams, to impute missing values. Tabarestani et al. (2020) presented a multitask learning method that handled missing data by projecting the high-dimensional input feature space into multiple low-dimensional and less-sparse spaces.

Nevertheless, state-of-the-art imputation approaches still face limitations that ought to be overcome, such as the often unfeasible requirement for a complete subset, i.e., a subset without missing elements (Mishra et al., 2021; Sefidian and Daneshpour, 2020; Liu et al., 2019). Besides, apart from the works developed by Yoon et al. (2019) and Tabarestani et al. (2020), the remaining imputation methods were only validated on datasets with a controlled and synthetic missingness, which does not fully reflect the entropy of a real-world scenario. Therefore, this paper proposes three novel imputation techniques that aim to tackle these drawbacks while exploiting correlation in an innovative manner.

3. Materials and methods

3.1. Missingness mechanisms

A nearly universal classification system for missing data problems was established by Rubin (1976), who pioneered the study on how the processes that cause missingness affect data analysis. Within this scope, Little and Rubin (2019) categorized the missingness mechanisms as Missing Completely at Random (MCAR), Missing at Random (MAR), or Missing not at Random (MNAR). Although these mechanisms do not offer a causal explanation for the missingness, they describe generic relations between data and missing elements, which are important to understand when choosing the appropriate method to handle the missing values (Enders, 2022).

Let $X = (x_{ij})$ denote an $(N \times g)$ dataset without missing values, i.e., a complete dataset, where N is the number of samples or observations, and g is the number of features. x_{ij} is the value of variable X_j for observation i. If X contains missing data then the missingness indicator matrix $M = (m_{ij})$ is defined, such that $m_{ij} = 1$ if x_{ij} is missing and $m_{ij} = 0$ otherwise.

According to Little and Rubin (2019), the formal description of the missingness mechanisms relies on the conditional distribution of m_i given x_i, hereby represented as $f_{M|X}(m_i|x_i, \phi)$, where ϕ denotes unknown parameters.

In the MCAR mechanism, the missingness is completely unrelated to the data values, missing or observed. This mechanism verifies the equality

$$f_{M|X}(m_i|x_i, \phi) = f_{M|X}(m_i|x_i^*, \phi) \quad (1)$$

for all i and any distinct values (x_i, x_i^*) in the sample space of X. Accordingly, Eq. 1 acknowledges that the probability of missingness on a variable X_j is not dependent on other measured variables nor on the values of X_j itself.

In the MAR mechanism, the missingness is related to the observed values of the data, but not the missing ones. Let $x_{(0)i}$ and $x_{(1)i}$ denote the observed and missing elements of x_i, respectively. This mechanism verifies the equality

$$f_{M|X}(m_i|x_{(0)i}, x_{(1)i}, \phi) = f_{M|X}\left(m_i|x_{(0)i}, x^*_{(1)i}, \phi\right) \quad (2)$$

for all i and any distinct values $\left(x_{(1)i}, x^*_{(1)i}\right)$ in the sample space of $X_{(1)}$. In conformity with Eq. 2, the probability of missingness on a variable X_j depends solely on the values of another measured variable (or variables) but not on the values of X_j itself.

Finally, the MNAR mechanism's missingness is related to the unobserved data. According to Little and Rubin (2019), the distribution of m_i depends on the missing elements of x_i, i.e. Eq. 2 does not apply for some sample i and some values $\left(x_{(1)i}, x^*_{(1)i}\right)$. This is the only mechanism that permits an association between the probability of missingness on a variable X_j and the values of X_j itself. Also, the missingness can depend on the observed values of the data as long as it still relates to the missing ones.

3.2. Correlation

Correlation is a measure of association between two variables, i.e., an indicator of how much a change in the magnitude of one variable is related to a change in the magnitude of another variable (Schober et al., 2018). Although knowing the values of a specific variable allows a better prediction of the values of a correlated variable, correlation does not necessarily assure causality.

Correlation coefficients are statistical measures of the degree of correlation between variables. There are several coefficients, each suitable for specific types of variables and with distinct underlying assumptions. Below, the coefficients used in this paper will be presented.

Pearson's (product-moment correlation) coefficient is one of the most used correlation measurements in medical research. Commonly denoted by r, this coefficient measures the strength of a linear relationship between two numeric, random variables X and Y, calculated by the following equation:

$$r = \frac{\text{cov}_{XY}}{\sigma_X \sigma_Y} \quad (3)$$

where cov_{XY} is the covariance value between X and Y, and σ_X and σ_Y are the standard deviations of variables X and Y, respectively.

Pearson's coefficient is a normalised covariance, scaled so that it ranges from -1 to +1, indicating a perfect negative and positive linear correlation, respectively. Pearson's correlation is only suitable for random numeric variables that follow a bivariate normal distribution (Schober et al., 2018; Akoglu, 2018).

The Phi coefficient, denoted by ϕ, is the equivalent of Pearson's coefficient, which measures the linear correlation between two binary variables X and Y. It can be calculated through Pearson's chi-square goodness-of-fit statistic for the 2×2 contingency table of the two variables:

$$\phi = \sqrt{\frac{\chi^2}{N}} \quad (4)$$

where χ^2 is the chi-square statistic and $N$is the number of observations. The Phi coefficient also ranges from -1 to +1.

The point biserial correlation coefficient, represented by r_{pbi}, measures the strength of association between a binary nominal variable Y and a numeric variable X:

$$r_{\text{pbi}} = \frac{\overline{X}_1 - \overline{X}_0}{\sigma_X} \sqrt{p_1 p_0} \quad (5)$$

where p_1 and $p_0 = 1 - p_1$ are respectively the proportion of samples with $Y = 1$ and $Y = 0$, $\overline{X}_1$ and $\overline{X}_0$ are respectively the means of X given $Y = 1$ and $Y = 0$, and σ_X is the standard deviation of X. This measurement also ranges from -1 to +1.

The chosen coefficients are all based on Pearson's correlation. Therefore, the correlation values obtained for relationships between different variable types are comparable.

4. Novel correlation-based imputation techniques

This paper proposes three novel techniques which leverage correlation when performing missing value imputation. These techniques aim to tackle some of the drawbacks found in the methods from the literature, such as the need for a complete subset and the evaluation on a single missingness mechanism. Furthermore, they exploit the concept of correlation, a promising but still not widely adopted approach in medical research. Rather than just resorting to the correlation between values, these methods investigate the potential benefits of considering the correlation between values and missingness patterns. If this correlation is strong, then the missing elements on the incomplete variable are confined to a particular segment in the distribution of the other variable's values. Such association may be helpful when computing estimates for the missing elements.

For this purpose, it is key to establish the procedure by which the matrices denoting the correlation between values and missingness patterns are obtained. Consider any dataset and let g denote the total number of features and $g_{\text{miss}} \leqslant g$ the number of features with missing values. g_j represents the jth feature, where $j \in \{1, 2, \ldots, g\}$, and $g_{\text{miss},i}$ represents the ith incomplete feature, where $i \in \{1, 2, \ldots, g_{\text{miss}}\}$. Furthermore, consider that C_{vm} is a $(g_{\text{miss}} \times g)$ matrix. For every pair $\{i, j\}$, with $g_{\text{miss},i} \neq g_j$, $C_{\text{vm}}[i, j]$ stores the correlation between the values of g_j and the missingness pattern of $g_{\text{miss},i}$, i.e., its binary missingness indicator.

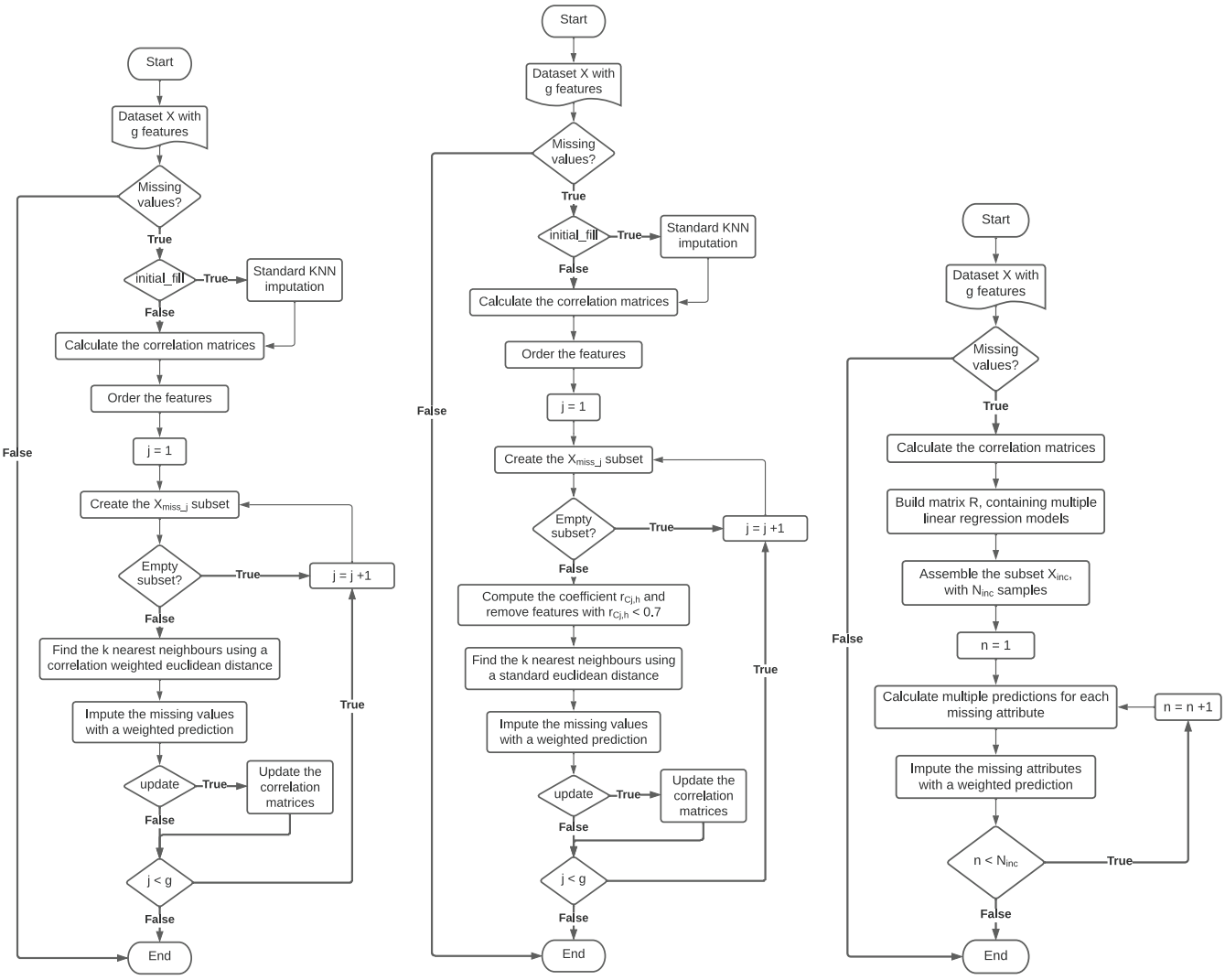
Additionally, let C_{vv} denote the standard $(g \times g)$ correlation matrix. This matrix is computed through a pairwise deletion strategy, i.e., the correlation is calculated between the available values within each pair of attributes on an analysis-by-analysis basis.

4.1. Correlation weighted K-nearest neighbours imputation

The Correlation Weighted K-Nearest Neighbours Imputation (CWKNNI) method was inspired by the imputation technique CoHiKNN, proposed by Liu et al. (2019). However, instead of uniquely considering the correlation between values of different attributes, the weights in CWKNNI are obtained through a weighted average of the correlation between values and the correlation between values and missingness patterns. Additionally, CWKNNI overcomes the limitation of CoHiKNN in terms of its dependency on a complete data subset to impute the missing values, as it computes the correlation matrix through a pairwise deletion approach instead of performing listwise deletion.

A simple flowchart of CWKNNI is shown in Fig. 1a and the following step-by-step explanation provides the outline for this method:

1. Consider a dataset X, with g features and N instances.
2. Compute the $(g \times g)$ correlation matrix, denoted as C_{vv}, and the $(g_{\text{miss}} \times g)$ matrix, hereby denoted as C_{vm}, with the correlations between values and missingness patterns. This approach considers the absolute value of these correlations, thus accounting for the strength of the association, not its direction.
3. Order the features from lowest to highest missing rate. Imputation will be performed in this sequence, in a phased manner.



(a) Flowchart of CWKNNI. (b) Flowchart of KNNSCI. (c) Flowchart of CWRI.

Fig. 1. Flowcharts of the proposed correlation-based methods.

4. Let g_j denote the attribute being imputed. Create a subset $X_{\text{miss\_}j}$ consisting of samples where g_j is missing. Furthermore, create a subset $X_{\text{obs\_}j}$ with samples where g_j is observed. If $X_{\text{miss\_}j}$ is empty, skip to the next attribute.
5. For each instance in $X_{\text{miss\_}j}$ find the k nearest neighbours within $X_{\text{obs\_}j}$. A weighted euclidean distance which accounts for the presence of missing values is used as a distance measure:

$$d_{vt} = \sqrt{w_{\text{D}} \times \sum_{i \in O_g} (1 - w_{\text{Ci}}) \times (v_i - t_i)^2} \tag{6}$$

where d_{vt} is the distance between samples v and t, v is an instance of $X_{\text{miss\_}j}$, t is a sample from $X_{\text{obs\_}j}$. v_i and t_i are the observed values from the ith attribute of v and t respectively. O_g is the set of indexes of the variables that are not missing in v nor in t. w_{D} is the quotient between the total number of features g and the dimension of O_g. As for w_{Ci}, it is a weighted average of the correlation between g_j and g_i, and the correlation between the values of g_i and the missingness pattern of g_j:

$$w_{\text{Ci}} = p \times C_{\text{vv}}[i,j] + (1 - p) \times C_{\text{vm}}[i,j], p \in [0,1] \tag{7}$$

Note that $0 \leqslant w_{\text{Ci}} \leqslant 1$.

6. Replace the missing value of g_j in each instance of $X_{\text{miss\_}j}$ with a weighted prediction, in which the weights are the inverse of the computed distances d_{vt}: a closer neighbour has higher importance (i.e., weight) in the final prediction. The mean value is used for numeric variables, whereas the mode is used to impute binary attributes.
7. Repeat Steps 4–6 until all missing values from all features have been imputed.

The number of neighbours k and the percentage p, i.e., the weight placed on the correlation between values in comparison to the correlation between values and missingness pattern, are two parameters of CWKNNI.

This method also accepts a boolean parameter, **initial_fill**, which determines if an initial and temporary imputation with the standard KNN method is carried out. The values from this imputed dataset will be used for the calculation of both C_{vv} and C_{vm}. Furthermore, the subset $X_{\text{obs\_}j}$ will be formed by these KNN imputed samples. The number of samples within each $X_{\text{obs\_}j}$ remains the same; the only difference is that no instance has missing attributes. As for the sample being imputed, g_j will be its only missing feature.

Additionally, the boolean argument **update** controls if the imputed values will be used in subsequent phases of imputation instead of just considering the original values of each sample. Both matrices C_{vv} and C_{vm} are updated at the beginning of each phase, thus accounting for the newly imputed values. Note that the missingness patterns used to calculate C_{vm} do not change.

4.2. K-nearest neighbours selected by correlation imputation

As with most KNN-based approaches, CWKNNI exhibits an increased computational cost for high dimensional data. The K-Nearest Neighbours Selected by Correlation Imputation (KNNSCI) method was created to overcome this drawback. For each variable to be imputed, this technique performs a pre-selection of features based on correlation, which reduces the dimensionality and facilitates its application on large datasets. Although both CWKNNI and KNNSCI are KNN-based approaches, KNNSCI is presented as a separate method from the former, since it leverages correlation in a distinct manner: while CWKNNI integrates correlation in a weighted distance function, KNNSCI does not because it resorts to a standard euclidean distance; furthermore, KNNSCI uses correlation to carry out a pre-selection of features that is not performed by CWKNNI.

Fig. 1b depicts a simple flowchart of KNNSCI and the outline for this method is given below:

1–4. Equal to Steps 1–4 from CWKNNI.
5. Using the correlation matrices, compute the coefficient $r_{Cj,h}$ between attribute g_j and the hth attribute of X (apart from g_j). This coefficient is equal for all samples within the subset X_{miss_j} and is obtained through the following equation:

$$r_{Cj,h} = p \times C_{vv}[h,j] + (1-p) \times C_{vm}[h,j], p \in [0,1] \quad (8)$$

Note that $0 \leqslant r_{Cj,h} \leqslant 1$.
6. Create a subset X_{obs_j} with samples where g_j is observed. Find the features where $r_{Cj,h} < 0.7$ and remove their columns from both X_{miss_j} and X_{obs_j}. The value of 0.7 was chosen based on the work of Schober et al. (2018), which stated that a coefficient above 0.7 indicates a strong correlation.
7. For each instance in X_{miss_j} find the k nearest neighbours within X_{obs_j} (after removing the columns with $r_{Cj,h} < 0.7$). A standard euclidean distance is used as a distance measure.
8. Replace the missing value on the attribute g_j in each instance of X_{miss_j} with a weighted prediction, in which the weights are the inverse of the computed euclidean distances: a closer neighbour has a higher importance in the final prediction. The mean value is used for numeric variables, whereas the mode is used to impute binary attributes.
9. Repeat Steps 4–8 until all missing values from all features have been imputed.

In addition to the number of neighbours k and the percentage p, KNNSCI also has the parameters **initial_fill** and **update**, which serve a similar purpose as in CWKNNI.

4.3. Correlation weighted regression imputation

The Correlation Weighted Regression Imputation (CWRI) approach is a regression-based method in which the missing values are imputed with predictions drawn from distinct linear regression models. This technique finds several estimates for each missing value and combines them taking into account the correlational importance of the predictor variables. This correlational importance is obtained through a weighted average of the correlation between values and the correlation between values and missingness patterns.

Fig. 1c displays a simple flowchart of CWRI and, as before, a step-by-step outline of this method is presented:

1. Consider a dataset X, with g features and N instances. Additionally, let g_{miss} denote the number of attributes with missing values.
2. Compute the $(g \times g)$ correlation matrix, denoted as C_{vv}, and the $(g_{miss} \times g)$ matrix, whereby denoted as C_{vm}, with the correlations between values and missingness patterns. This approach leverages the absolute value of these correlations.
3. Build a $g \times g$ matrix, denoted as R, containing multiple linear regression models. Let g_i and g_j be two different attributes of X, with indexes i and j, respectively. For every pair $\{i,j\}$, where $i \neq j, R[i,j]$ stores a linear regression model in which g_i is the independent variable and g_j is the predictor.
4. Assemble the subset X_{inc} including all incomplete samples, i.e., instances with at least one missing value.
5. Consider any instance of X_{inc}, and let g_k denote its kth missing attribute. Using every non-missing feature g_t in this sample, obtain the corresponding predicted value for g_k through the regression model $R[k,t]$. If g_k is a numeric variable, the final imputation will be a weighted average of all predicted values. If g_k is binary, a weighted mode is applied. To obtain the weight given to each variable g_t, denoted as $w_{k,t}$, first calculate its correlational importance $w_{Ck,t}$ through the following equation:

$$w_{Ck,t} = p \times C_{vv}[t,k] + (1-p) \times C_{vm}[t,k], p \in [0,1] \quad (9)$$

Note that $0 \leqslant w_{Ck,t} \leqslant 1$. After computing $w_{Ck,t}$ for every non-missing feature, $w_{k,t}$ is a simple normalisation:

$$w_{k,t} = \frac{1}{\sum_t w_{Ck,t}} \times w_{Ck,t} \quad (10)$$

6. Repeat Step 5 until all incomplete samples have been imputed.

The percentage p is the only parameter of CWRI.

5. Experiments

5.1. Datasets

Three complete and publicly available datasets were selected from the UCI Machine Learning Repository: **Wine Data Set**, only comprising numeric variables; **SPECT Heart Data Set**, which solely includes binary attributes; **Statlog (Heart) Data Set**, a mixed-type dataset. The selection process was primarily based on the type of variables of each dataset, as it is essential to ensure that these methods perform a suitable and efficient imputation regardless of the attribute's type.

Within this paper, multiclass nominal variables were one-hot encoded, and ordinal encoding was applied to ordinal attributes. Hence, assessing the imputation precision is only relevant for numeric and binary variables.

Synthetic missing values were injected into these three datasets, under all three missingness mechanisms, with three different missing rates (10%, 30%, and 50%) defined for every attribute. In the MCAR mechanism, all features were incomplete and shared the same missing rate, whereas in the remaining two mechanisms, only 50% of the features were incomplete but also had the same missing rate. A total of nine synthetic datasets were generated per UCI Machine Learning Repository dataset. For this end, the R package missMethods (Rockel, 2022) was used, as it supplies

functions for injecting missing data. This manipulation enabled a comprehensive study that included different missing rates and missingness mechanisms.

Then, since the main focus of this paper is to study missing value imputation in clinical contexts, two real-world medical datasets were chosen: the Osteoporosis Dataset and the Cardiothoracic Surgery Dataset.

The Osteoporosis Dataset assembles publicly available data from the 2013–2014 cycle of the NHANES (National Health and Nutrition Examination Survey Data, 2022). The dataset consists of 37 variables and 1643 subjects classified into three conditions: normal, osteopenia, and osteoporosis. The osteopenia and osteoporosis classes were combined into a single one, thereby transforming this case study into a binary classification problem. As for the collected data, Table 1 gives a brief characterisation of each feature group within the Osteoporosis Dataset, including the number and type of variables, and the average missing rate.

After categorical encoding, the final working dataset was left with 43 variables, 36 numerical and 7 binary. The dataset has a proportion of 73.2% incomplete samples, i.e., subjects with at least one missing value. Furthermore, 38.8% of the individuals were classified as normal (healthy), and the remaining 61.2% were diagnosed with either osteoporosis or osteopenia. The average age of all participants was 58 ± 12 years, with those classified as normal having a mean of 58 ± 10 years and the remaining, considered not healthy, an average age of 62 ± 12 years.

As for the Cardiothoracic Surgery Dataset, it contains clinical and demographic information retrieved by the Cardiothoracic Surgery Service of Hospital de Santa Marta, Portugal, from 2011 to 2019. For each subject, the collection started during the pre-surgery period and extended up to one year after the surgical procedure. The dataset contains records from 8122 patients and was used to predict the occurrence of complications within three months after hospital discharge, in a binary classification problem.

Table 1
Brief characterisation of the Osteoporosis Dataset.

Feature Group	Attributes	Average missing rate (%)
Demographics	1 nominal, 2 ordinal, 1 numeric	$(1.1 \pm 1.9)\%$
Nutrition	6 numeric	$(7.6 \pm 0.2)\%$
Blood pressure	2 numeric	$(3.3 \pm 0.3)\%$
Anthropometrics	2 numeric	$(0.6 \pm 0.1)\%$
Physical fitness	1 numeric	10.0%
Blood lipids	4 numeric	$(27.8 \pm 25.4)\%$
Hormones	3 numeric	$(8.9 \pm 4.5)\%$
Biochemistry	2 numeric	$(3.1 \pm 0.7)\%$
Physical activity	11 numeric	$(10.0 \pm 0.1)\%$
Lifestyle	2 numeric	$(9.8 \pm 0.0)\%$

Data was collected on 106 medically relevant variables, which can be grouped into the categories shown in Table 2.

An initial pre-processing was performed, and the working dataset was left with 5625 subjects, 92.7% of which are negative cases. Those negative cases have an average age of 65 ± 13 years, whereas the positive cases have an average age of 68 ± 12 years. After applying categorical encoding, the resulting number of features is 119, of which 31 are numeric, and 88 are binary. Finally, 94.1% of the samples are incomplete.

In order to perform a time-based analysis, patients whose information was collected in 2019 were assembled in a separate test subset. This group contains 667 subjects, from which 92.7% are negative cases. Therefore, the distribution of class labels is maintained in this test set.

5.2. Experimental setup

Prior to imputation, a grouped stratified 5-fold strategy was applied to each dataset. Note that the UCI Machine Learning Repository datasets were divided after injecting synthetic missing values.

Table 3
Tested hyperparameter values for every classifier.

Classifier	Hyperparameter Values
UCI Machine Learning Repository datasets	
RF	**max_depth** $\in \{3, 5, 7, 9\}$, **n_estimators** $\in \{5, 10, 25, 50, 100, 200\}$, **criterion** $\in$ {'entropy', 'gini'}, **min_samples_split** $\in \{10, 20, 50\}$, **min_samples_leaf** $\in \{5, 10, 25\}$
SVM	**random_state** = 42, **class_weight** = 'balanced', **C** $\in \{0.0001, 0.001, 0.01, 0.1, 1, 10, 100\}$, **gamma** $\in \{0.0001, 0.001, 0.01, 0.1, 1, 10, 100\}$
NB	**var_smoothing** $\in \{1.0 \times 10^{-4}, 2.8 \times 10^{-5}, 7.7 \times 10^{-6}, 2.2 \times 10^{-6}, 6.0 \times 10^{-7}, 1.7 \times 10^{-7}, 4.6 \times 10^{-8}, 1.3 \times 10^{-8}, 3.6 \times 10^{-9}, 1.0 \times 10^{-9}\}$
Real-world medical datasets	
RF	**max_depth** $\in \{3, 6, 9\}$, **n_estimators** $\in \{10, 25, 100\}$, **criterion** $\in$ {'entropy', 'gini'}, **min_samples_split** $\in \{20, 50\}$, **min_samples_leaf** $\in \{10, 25\}$
SVM	**random_state** = 42, **class_weight** = 'balanced', **C** $\in \{0.0001, 0.001, 0.01, 0.1, 1, 10, 100\}$, **gamma** $\in \{0.0001, 0.001, 0.01, 0.1, 1, 10, 100\}$
NB	**var_smoothing** $\in \{1.0 \times 10^{-4}, 2.8 \times 10^{-5}, 7.7 \times 10^{-6}, 2.2 \times 10^{-6}, 6.0 \times 10^{-7}, 1.7 \times 10^{-7}, 4.6 \times 10^{-8}, 1.3 \times 10^{-8}, 3.6 \times 10^{-9}, 1.0 \times 10^{-9}\}$

Table 2
Brief characterisation of the Cardiothoracic Surgery Dataset.

Feature Group	Attributes	Average missing rate (%)
Hospitalisation	4 dates	$(23.7 \pm 47.2)\%$
Cardiac history	4 ordinal, 1 binary	$(16.1 \pm 35.3)\%$
Previous interventions	2 dates, 1 ordinal, 1 nominal	$(44.9 \pm 51.6)\%$
Pre-operative risk factors	4 numeric, 3 ordinal, 4 nominal, 4 binary	$(0.9 \pm 0.2)\%$
Pre-operative haemodynamics and catheterisation	1 date, 5 numeric, 2 ordinal, 1 nominal, 1 binary	$(49.3 \pm 42.8)\%$
Pre-operative status and support	4 binary	$(0.6 \pm 0.2)\%$
Operation	1 text, 2 ordinal, 4 nominal	$(27.2 \pm 46.1)\%$
Coronary surgery	2 numeric, 2 nominal	$(61.4 \pm 0.7)\%$
Valve surgery	1 code, 8 numeric, 3 ordinal, 10 nominal, 7 binary	$(34.8 \pm 30.2)\%$
Cardiac Surgery Morbidity Scale	1 numeric, 8 nominal	$(7.6 \pm 22.1)\%$
Discharge details	2 nominal, 1 binary	$(32.4 \pm 55.6)\%$
Patient demographics/ autocalculations	1 code, 10 numeric, 1 nominal, 1 binary	$(7.8 \pm 26.4)\%$

The performance of the proposed imputation techniques was evaluated through a comparative study. Overall, seven other imputation methods were selected to be part of this study: Mean/ Mode, Regression, KNN, CMIM (Sefidian and Daneshpour, 2020), CoHiKNN (Liu et al., 2019), NMVI (Bhagat and Singh, 2022), and MICE (Van Buuren and Groothuis-Oudshoorn, 2011). CMIM are a compilation of ten distinct imputation techniques; for this work only the fifth one was implemented and tested. Additionally, the results produced by listwise deletion were also included in the performed analysis. This selection sought to encompass both standard and modern methods with diverse baseline strategies. These methods served as benchmarks, enabling a more informative evaluation of the proposed techniques.

Two types of performance evaluation were carried out: imputation precision and classification evaluation. The first type assessed the quality of the imputation procedure by comparing the imputed values with the original ones, i.e., ground truth, resorting to the mean absolute error (MAE). This evaluation could only be carried out on the incomplete datasets generated from the three UCI Machine Learning Repository datasets since it is impossible to trace back the real values of the missing elements in the remaining datasets. The classification evaluation studied the impact of each imputation procedure on an ML model's performance, namely on RF, SVM, and NB classifiers. A 5-fold cross-validation strategy was adopted, and the average AUROC was used to compare the various imputation techniques.

The performed comparative study aimed to be as rigorous and fair as possible. To this end, hyperparameter tuning constituted an essential step to guarantee that each imputation technique was being evaluated under the best possible conditions, minimising the likelihood of any factors external to the imputation procedure corrupting the results. A grid search technique was applied to each stratified fold of the original UCI Machine Learning Repository databases, instead of the imputed datasets. The procedure followed for the real-world datasets was slightly different. As there is no baseline dataset (ground truth), it was considered necessary to compute a grid search with 5-fold cross-validation for each stratified fold of each imputed dataset to ensure a fair comparison between the imputation techniques. In order to soften the computational cost, the number of tested values within some hyperpa-

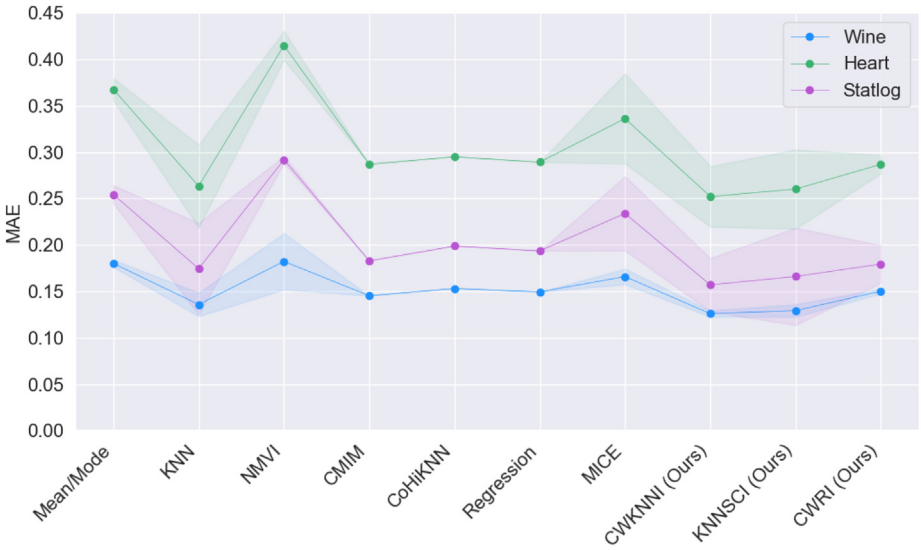


Fig. 2. MAE for all synthetically generated datasets under the MCAR missingness mechanism.

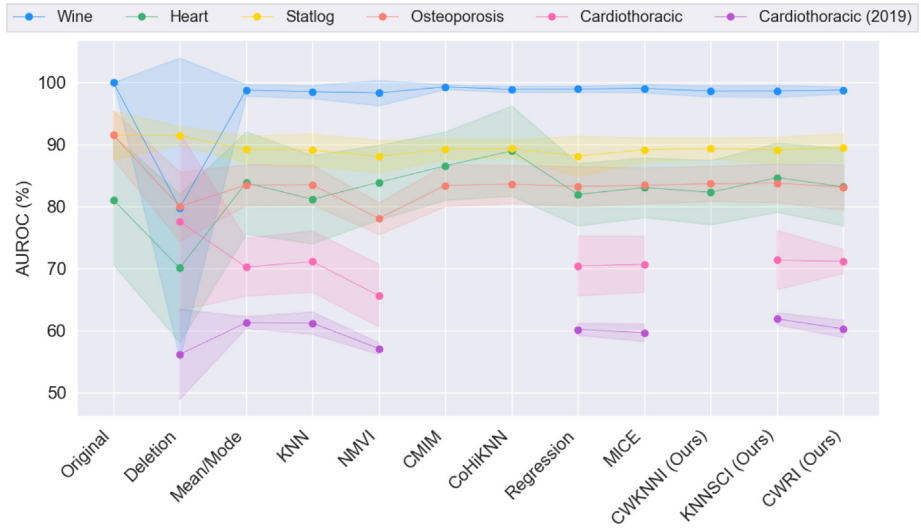


Fig. 3. AUROC for every imputation method performed on the working datasets, obtained through an RF classifier.

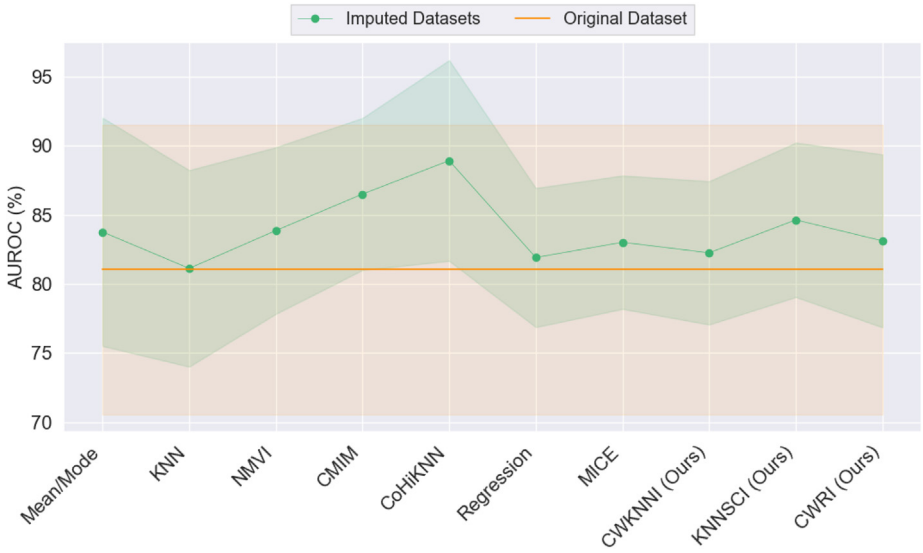


Fig. 4. AUROC for every imputation method performed on the SPECT Heart Data Set, obtained through an RF classifier. These results are compared against the AUROC of the RF trained upon the original dataset (orange line).

rameters of the RF classifier was lowered, as shown in Table 3. For each classifier, the optimal hyperparameters are the ones from the model with the highest average AUROC. The AUROC was therefore selected to assess the classifiers' performance. No threshold optimisation was performed, and thus the default value of 0.5 was used to interpret probabilities to class labels.

6. Results and discussion

The imputation quality of the proposed methods will be first discussed. The average value of the three distinct MAEs corresponding to the three chosen missing rates was computed for each imputation method when applied to every synthetic dataset, as displayed in Table 4. Techniques unable to perform imputation for higher missing rates are marked, as the calculated errors do not fully represent their efficacy. Since the MAE frequently grows with the missing rate, because the amount of useful information decreases, the quality of the imputation performed by those techniques is likely worse than what the displayed MAEs indicate. For this reason, although the marked techniques sometimes have the lowest MAEs, they cannot be considered the best in terms of imputation quality.

The proposed correlation-based imputation methods yield consistently good results in all three missingness mechanisms. Notably, compared to its competitors, CWKNNI is the most precise technique in the MCAR mechanism. Furthermore, the ranking concerning the quality of the MCAR imputation procedure remains nearly unchanged independently of the dataset, i.e., techniques that are superior (inferior) in the Wine Data Set are also superior (inferior) in the SPECT Heart Data Set and the Statlog (Heart) Data Set. Fig. 2 provides a visual representation of this observation. Hence, for MCAR data of both numeric and binary types, it is inferred that CWKNNI will produce estimates that are closer to the real values if they had been observed.

As for the MAR mechanism, the proposed methods exhibit an imputation quality close to their competitors, although not clearly better. The injection of MAR values was based on the pairing of highly correlated features, where one of the features determines the missing elements in the other. Hence it was expected that the proposed methods would yield the best results out of all techniques since they account for the correlation between values in the imputation process. Although their imputation quality was among

Table 4
Average MAE for all missingness mechanisms. The highlighted values are the lower MAEs in each assessment.

Imputation Method	Wine	SPECT Heart	Statlog (Heart)
MCAR mechanism			
Mean/ Mode	0.18 ± 0.00	0.37 ± 0.01	0.25 ± 0.01
KNN	0.14 ± 0.01	0.26 ± 0.04	0.17 ± 0.05
NMVI	0.18 ± 0.03	0.42 ± 0.02	0.29 ± 0.00
CMIM	0.15[a]	0.29[a]	0.18[a]
CoHiKNN	0.15[a]	0.29[a]	0.20[a]
Regression	0.15[a]	0.29[a]	0.19[a]
MICE	0.17 ± 0.01	0.34 ± 0.05	0.23 ± 0.04
CWKNNI*	**0.13 ± 0.01**	**0.25 ± 0.03**	**0.16 ± 0.03**
KNNSCI*	**0.13 ± 0.01**	0.26 ± 0.04	**0.17 ± 0.03**
CWRI*	0.15 ± 0.00	0.29 ± 0.01	0.18 ± 0.02
MAR mechanism			
Mean/ Mode	0.23 ± 0.01	0.46 ± 0.05	0.35 ± 0.08
KNN	**0.13 ± 0.03**	**0.19 ± 0.09**	0.21 ± 0.11
NMVI	0.14 ± 0.02	0.43 ± 0.06	0.32 ± 0.12
CMIM	0.19 ± 0.04	0.23 ± 0.03[b]	0.21 ± 0.08
CoHiKNN	0.19 ± 0.02	0.16 ± 0.01[b]	0.28 ± 0.10
Regression	0.19 ± 0.05	0.13 ± 0.01[b]	19.02 ± 20.38
MICE	0.15 ± 0.03	0.29 ± 0.05	**0.18 ± 0.03**
CWKNNI*	**0.13 ± 0.03**	0.22 ± 0.10	0.24 ± 0.11
KNNSCI*	**0.13 ± 0.03**	0.20 ± 0.09	0.23 ± 0.11
CWRI*	0.18 ± 0.02	0.25 ± 0.13	0.27 ± 0.07
MNAR mechanism			
Mean/ Mode	0.35 ± 0.00	0.62 ± 0.11	0.41 ± 0.13
KNN	0.24 ± 0.05	0.38 ± 0.17	0.30 ± 0.11
NMVI	**0.16 ± 0.04**	0.49 ± 0.08	0.33 ± 0.05
CMIM	0.27 ± 0.06	0.41 ± 0.18	0.18 ± 0.02[b]
CoHiKNN	0.29 ± 0.02	0.41 ± 0.27	0.30 ± 0.04[b]
Regression	0.27 ± 0.07	0.24 ± 0.04[b]	0.20 ± 0.01[b]
MICE	0.24 ± 0.05	**0.35 ± 0.08**	**0.27 ± 0.03**
CWKNNI*	0.24 ± 0.05	0.39 ± 0.20	0.31 ± 0.12
KNNSCI*	0.24 ± 0.05	0.39 ± 0.16	0.30 ± 0.11
CWRI*	0.31 ± 0.01	0.45 ± 0.22	0.36 ± 0.07

[a] No average was computed and only the MAE for a missing rate of 10% is shown because the technique was unable to perform imputation for the missing rates of 30% and 50%.
[b] The average was only computed with the MAEs for the missing rates of 10% and 30% because the technique was unable to perform imputation for a missing rate of 50%.
* Algorithms developed in this work.

Table 5
Average AUROC(%) obtained for each classifier. The highlighted values are the higher scores in each assessment.

	Wine	SPECT Heart	Statlog (Heart)	Osteoporosis	Cardiothoracic	Cardiothoracic (2019)
RF classifier						
Original Dataset	100.00±0.00	81.04±10.47	91.47±3.90	N/A	N/A	N/A
Listwise Deletion	79.70±24.26[a]	70.04±11.94[a]	91.38±1.49[b]	79.96±5.59	77.55±14.35[d]	56.13±7.31
Mean/ Mode	98.71±0.94	83.77±8.26	89.18±2.22	83.43±3.31	70.18±4.66	61.24±0.97
KNN	98.45±1.07	81.12±7.12	89.05±2.62	83.45±3.20	71.08±4.94	61.17±1.81
NMVI	98.27±2.05	83.86±6.03	88.06±2.59	78.04±2.59	65.61±5.04	57.07±0.94
CMIM	99.24±0.39[a]	86.50±5.50[a]	89.23±1.68[b]	83.36±3.36	[c]	[c]
CoHiKNN	98.85±0.51[a]	88.93±7.27[a]	89.30±1.44[b]	83.57±3.12	[c]	[c]
Regression	98.90±0.50[a]	81.91±5.03[a]	88.11±3.22[b]	83.21±3.20	70.36±4.81	60.16±0.97
MICE	**98.99 ± 0.69**	83.02±4.83	89.12±1.96	83.37±2.92	70.63±4.51	59.60±1.41
CWKNNI*	98.56±0.91	82.25±5.19	89.33±1.71	83.64±2.85	[c]	[c]
KNNSCI*	98.57±0.98	**84.63 ± 5.58**	89.09±2.07	**83.73 ± 3.13**	**71.32 ± 4.78**	**61.86 ± 1.00**
CWRI*	98.72±0.56	83.12±6.26	**89.50 ± 2.26**	83.07±3.66	71.10±1.92	60.23±1.39
SVM classifier						
Original Dataset	100.00±0.00	83.39±9.43	91.17±3.17	N/A	N/A	N/A
Listwise Deletion	91.97±15.99[a]	58.39±24.53[a]	89.55±2.9[b]	82.58±4.33	46.37±18.97	49.00±8.66
Mean/ Mode	99.27±0.60	75.71±24.30	88.37±1.95	82.78±2.40	68.58±2.89	55.95±2.55
KNN	99.27±0.62	79.44±13.16	88.27±2.32	82.75±2.35	68.88±2.70	55.97±2.44
NMVI	98.78±1.72	**84.72 ± 4.36**	88.31±2.09	80.39±3.19	68.54±2.92	55.61±2.56
CMIM	99.54±0.34[a]	85.42±3.88[a]	88.81±1.35[b]	82.57±2.52	[c]	[c]
CoHiKNN	99.50±0.36[a]	85.47±5.84[a]	89.26±1.09[b]	82.74±2.32	[c]	[c]
Regression	99.41±0.49[a]	83.85±3.68[a]	87.39±3.90[b]	82.29±2.25	63.72±4.18	**56.67 ± 1.64**
MICE	99.35±0.54	80.77±14.89	88.49±2.00	82.62±2.59	**68.96 ± 2.54**	55.56±2.89
CWKNNI*	99.35±0.50	80.91±9.92	88.54±1.46	82.91±2.70	[c]	[c]
KNNSCI*	99.23±0.59	83.35±6.00	88.35±2.45	**83.02 ± 2.58**	68.91±2.68	56.14±2.63
CWRI*	**99.40 ± 0.44**	84.32±5.76	**88.91 ± 1.84**	82.78±2.41	68.58±1.85	55.93±3.24
NB classifier						
Original Dataset	99.82±0.25	74.86±4.86	86.50±4.33	N/A	N/A	N/A
Listwise Deletion	82.30±20.3[a]	68.97±9.65[a]	82.12±6.69[b]	77.28±5.35	50.85±14.71	49.58±3.16
Mean/ Mode	97.79±1.90	74.63±5.95	85.50±1.55	77.54±2.82	66.58±4.57	58.00±0.88
KNN	98.21±1.13	75.92±4.04	85.93±1.99	77.80±2.93	67.05±4.27	57.86±0.78
NMVI	97.57±2.21	74.32±3.55	86.07±1.81	77.54±2.62	66.90±4.36	58.00±0.85
CMIM	98.07±1.48[a]	77.97±5.04[a]	85.32±1.42[b]	77.47±2.60	[c]	[c]
CoHiKNN	98.58±0.81[a]	78.01±5.85[a]	85.84±0.61[b]	77.58±2.93	[c]	[c]
Regression	98.00±1.50[a]	74.45±3.16[a]	85.69±2.37[b]	76.51±2.56	50.80±7.18	49.56±3.97
MICE	**98.45 ± 1.04**	75.44±2.72	**87.38 ± 1.66**	77.59±2.69	**67.45 ± 4.07**	**58.04 ± 1.15**
CWKNNI*	98.28±1.08	**77.36 ± 5.70**	86.63±0.91	**77.90 ± 2.89**	[c]	[c]
KNNSCI*	98.30±1.15	76.60±3.74	86.19±1.46	77.89±2.87	67.06±4.26	57.94±0.79
CWRI*	98.12±1.40	75.11±5.13	86.28±1.57	77.55±2.80	65.07±3.50	57.56±0.68

[c] The technique was unable to perform imputation due to high computational costs.
[d] Only 7.3% of the samples from the Cardiothoracic Surgery Dataset were used in this complete-case analysis, which is not at all representative.
[b] The technique was unable to perform under the MCAR mechanism with missing rates of 30% and 50%, and under the MNAR mechanism with a missing rate of 50%.
[a] The technique was unable to perform under the MCAR mechanism with missing rates of 30% and 50%.
* Algorithms developed in this work.

the best, this superiority was not observed. The impact of missingness on the correlation between values, which increases as the missing rate grows, may have harmed the precision of the imputation procedure.

The MNAR mechanism generally presents the largest MAEs in all imputation methods, which was expected given that most existing techniques, both standard and state-of-the-art, are MAR-based approaches, and thus provide better results under the MCAR and MAR mechanisms. As for the proposed methods, their MAEs are also higher on the MNAR mechanism and the values are similar to those yielded by the remaining imputation techniques.

Regarding leveraging correlation for the prediction of missing values, the quality of the imputation performed by the proposed techniques CWKNNI and KNNSCI can be considered overall superior to that of competing correlation-based methods, i.e., CMIM and CoHiKNN. Furthermore, unlike these methods, the developed techniques could perform imputations for every missing rate, which again shows their superiority.

Table 5 concerns the classification performance evaluation, in which three different ML classifiers (RF, SVM, and NB) were trained upon the imputed datasets, and their performance was compared in terms of AUROC. This evaluation also includes results concerning the imputation performed on the two real-world medical datasets. As for the synthetic datasets, the AUROCs regarding all missing rates of every mechanism were averaged. As before, techniques unable to perform imputation on certain datasets are marked.

For each classifier, the proposed correlation-based imputation techniques are comparable to their competitors and can even be considered superior in some cases. Particularly, RFs that were trained upon datasets imputed through the proposed KNNSCI method exhibit an overall better performance, presenting the best AUROCs in 4 out of 6 evaluations, including the two real-world datasets, as shown in Fig. 3. Although this result is not statistically significant, combining the KNNSCI method with an RF classifier obtained the best classification performances in the two real-world medical datasets, which is a desired achievement when developing reliable and robust ML-based systems to deploy in clinical contexts.

Moreover, correlation-based imputation yields the best AUROCs in 10 out of 18 evaluations, with the proposed KNNSCI and CWRI techniques being the main contributors to these results. These imputation techniques not only produced higher AUROCs but were also capable of performing imputation under all tested circumstances.

Table 5 also shows that the most prominent differences in AUROC come from the scores produced by Listwise Deletion in comparison to the others. This reinforces that this approach should be used cautiously because simply discarding all incomplete instances may produce a dataset that is not representative of the original problem, particularly for higher missing rates.

Prior to conducting this comparative study, it seemed intuitive that a more accurate imputation entailed a better classification performance. In order to investigate this hypothesis, the ML classifiers were also trained upon the original (and complete) UCI Machine Learning Repository datasets. Since these datasets represent an optimal imputation quality, i.e., a null MAE, comparing the obtained results with those of the classifiers trained upon imputed datasets allows the hypothesis to be tested. Furthermore, recall that the classifiers' hyperparameters were chosen after a grid search was applied to the original dataset, and no hyperparameter tuning was performed on the models trained upon the imputed datasets to optimise computational resources. Hence, the classifiers trained upon the original datasets may have a slight advantage over the remaining, as their performance was optimised.

Even so, in the SPECT Heart Data Set, for example, the vast majority of the RFs trained upon imputed datasets outperform the RF trained upon the original dataset, thus demonstrating that there is not a clear relationship between imputation quality and the performance of an ML model. Fig. 4 depicts this example. This finding raises the question of whether a more suitable imputation method for an ML-based clinical system is one that yields more precise estimates or one by which a better classification performance is achieved. At first glance, a clinical prediction model should have optimal performance, but it may not be acceptable to fully disregard the imputation quality of the chosen method. For instance, consider an imputation technique that produces biased parameter estimates, i.e., distorts the original statistical distribution of the data. In some cases, this permits a greater generalisation ability of the ML model trained upon the imputed data. However, in some other cases, the knowledge that should have been learned from the data may have been corrupted by the imputation procedure, ultimately leading to a facilitated learning task and misleadingly better classification results.

On a final note, distinct algorithms include different data processing approaches, which leads to some techniques being a better fit for specific problems than others. As discussed throughout this work, the proposed imputation methods are suitable for clinical problems, since they exploit the correlations intrinsic to medical datasets and tackle limitations of other techniques, such as the need for a complete subset. However, this may come at the expense of increased computational cost, as is evidenced by the missing results in Tables 4,5 regarding some correlation-based techniques. In order to further understand this drawback, a brief complexity analysis of the three proposed correlation-based methods was performed.

CWKNNI is $O\left(g^2 \times N^2\right)$, in which g is the number of features and N is the number of data samples. KNNSCI aimed to decrease the computational cost of the previous method, and is $O\left(g \times N^2 \times g_{sel}\right)$, where g_{sel} is the number of selected features with respect to the variable being imputed, thus $1 \leqslant g_{sel} < g$. CWRI is $O(N_{miss} \times g_{miss} \times (g-1))$, where g_{miss} is the number of incomplete features, i.e., $1 \leqslant g_{miss} \leqslant g$, and therefore also shows a lower computational complexity when compared to CWKNNI.

Concerning the time efficiency of the three proposed approaches, imputation times were measured for the Osteoporosis Dataset as an example. The results are the following: CWKNNI took 11438.70 s, followed by CWRI with 23.96 s, and KNNSCI with 10.98 s[1]. As expected, the CWKNNI required the most time to perform imputation, which was considerably greater than the ones required by both CWRI and KNNSCI. These runtimes vary with the characteristics of each dataset, which may affect the speed ranking of the three methods. In particular, it may lead to the CWRI method being faster than KNNSCI in some cases. Either way, both techniques overcome the high computational cost demonstrated by CWKNNI on high-dimensional data, as intended.

7. Conclusions

Missing data are ubiquitous in biomedical sciences, posing a recurring predicament in delivering reliable AI-based clinical systems. There has been a growing interest in strategies that address this inevitable challenge, specifically missing value imputation methods.

This paper proposed three novel correlation-based imputation techniques which leverage not only the correlation between values but also the correlation between values and missingness patterns. Their performance was evaluated in a comparative study which included existing methods, both standard and state-of-the-art. This study assessed the imputation quality of the proposed methods under diverse missingness conditions and on distinct variable types. Furthermore, classification performance was assessed on multiple datasets, both synthetic and real-world. Hence, a comprehensive evaluation that is often lacking in literature was ensured. The proposed techniques were in compliance with their competitors, sometimes outperforming them. In fact, the best AUROCs for real-world medical datasets were obtained through an RF trained using data imputed with the proposed KNNSCI method.

One of the proposed imputation methods, CWKNNI, could not be applied to the most complex dataset due to its high computational cost. Even though computational cost should not be a determining factor for dismissing an imputation method, it is a concern that must be considered in future works. Moreover, note that this drawback is shared by some state-of-the-art techniques, although it was overcome by the proposed methods KNNSCI and CWRI.

Lastly, a more accurate imputation did not entail a better classification performance, which may imply that a trade-off between these two properties has to be made when choosing an imputation technique.

In summary, this work confirmed the auspicious role of correlation-based imputation in improving ML-based clinical systems' robustness to missing values while addressing important limitations of current imputation methods.

CRediT authorship contribution statement

Isabel Curioso: Conceptualization, Methodology, Software, Data curation, Writing – original draft. **Ricardo Santos:** Conceptualization, Methodology, Validation, Resources, Data curation, Writing – review & editing, Supervision. **Bruno Ribeiro:** Conceptualization, Methodology, Validation, Writing – review & editing. **André Carreiro:** Writing – review & editing. **Pedro Coelho:** Resources, Supervision. **José Fragata:** Resources, Supervision. **Hugo Gamboa:** Writing – review & editing, Supervision.

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.

[1] Computing environment: Machine with an AMD EPYC 7713P 64-Core Processor @ 2 GHz, 64 GB RAM.

Acknowledgements

This work was done under the project "CardioFollow.AI: An intelligent system to improve patients' safety and remote surveillance in follow-up for cardiothoracic surgery", and supported by national funds through 'FCT – Portuguese Foundation for Science and Technology, I.P.', with reference DSAIPA/AI/0094/2020.

Appendix A

See Tables A.1, A.2, A.3 and A.4.

Table A.1
Tested parameter values for every imputation method within the comparative study.

Imputation Method	Parameter Values
Mean/ Mode	**missing_values** = np.nan
Regression	–
KNN	**missing_values** = np.nan, **n_neighbors** $\in \{5, 10, 15\}$, **weights**='distance', **metric** = 'nan_euclidean'
CMIM	**percentage** $\in \{0.1, 0.5, 0.9\}$, **threshold** $\in \{0.1, 0.5, 0.9\}$
CoHiKNN	**n_neighbors** $\in \{5, 10, 15\}$
NMVI	–
MICE	**m**=3, **maxit**=10, **seed** = 42, **defaultMethod**=c("pmm", "logreg", "polyreg", "polr")
CWKNNI	**n_neighbors** $\in \{5, 10, 15\}$, **percentage** $\in \{0.2, 0.4, 0.6, 0.8\}$, **initial_fill** $\in$ {False, True}, **update** $\in$ {False, True}
KNNSCI	**n_neighbors** $\in \{5, 10, 15\}$, **percentage** $\in \{0.2, 0.4, 0.6, 0.8\}$, **initial_fill** $\in$ {False, True}, **update** $\in$ {False, True}
CWRI	**percentage** $\in \{0.2, 0.4, 0.6, 0.8\}$

Table A.2
Additional performance metrics for the CWKNNI method.

	AUROC(%)	Accuracy(%)	Precision(%)	Recall(%)	Specificity(%)	F1-Score(%)
RF classifier						
Wine	98.56 ± 0.91	91.11 ± 2.97	92.01 ± 2.38	91.70 ± 3.02	95.58 ± 1.58	91.16 ± 2.93
SPECT Heart	82.25 ± 5.19	81.95 ± 2.43	67.15 ± 5.0	71.93 ± 5.9	60.00 ± 12.57	64.62 ± 3.96
Statlog (Heart)	89.33 ± 1.71	81.48 ± 2.22	81.68 ± 2.19	81.35 ± 2.23	82.52 ± 3.45	81.24 ± 2.26
Osteoporosis	83.64 ± 2.85	75.90 ± 2.05	74.85 ± 2.07	75.54 ± 2.02	73.93 ± 3.26	75.03 ± 2.06
Cardiothoracic	[a]	[a]	[a]	[a]	[a]	[a]
Cardiothoracic (2019)	[a]	[a]	[a]	[a]	[a]	[a]
SVM classifier						
Wine	99.35 ± 0.50	93.82 ± 2.96	94.27 ± 2.56	94.31 ± 2.85	96.93 ± 1.51	93.87 ± 2.94
SPECT Heart	80.91 ± 9.92	69.49 ± 6.75	59.91 ± 3.94	73.28 ± 5.87	77.78 ± 7.03	56.09 ± 6.63
Statlog (Heart)	88.54 ± 1.46	81.56 ± 1.74	81.63 ± 1.88	81.30 ± 1.66	83.70 ± 2.52	81.30 ± 1.73
Osteoporosis	82.91 ± 2.70	74.92 ± 2.25	73.97 ± 2.37	74.89 ± 2.62	74.72 ± 4.63	74.16 ± 2.42
Cardiothoracic	[a]	[a]	[a]	[a]	[a]	[a]
Cardiothoracic (2019)	[a]	[a]	[a]	[a]	[a]	[a]
NB classifier						
Wine	98.28 ± 1.08	91.57 ± 3.25	92.62 ± 2.78	92.03 ± 3.17	95.71 ± 1.67	91.69 ± 3.22
SPECT Heart	77.36 ± 5.70	67.12 ± 5.87	57.99 ± 4.32	69.28 ± 5.54	71.85 ± 8.18	53.54 ± 6.01
Statlog (Heart)	86.63 ± 0.91	80.33 ± 1.98	80.64 ± 1.55	80.23 ± 1.81	81.11 ± 3.49	80.06 ± 2.04
Osteoporosis	77.90 ± 2.89	72.98 ± 3.62	71.64 ± 3.97	70.22 ± 4.10	57.94 ± 6.61	70.63 ± 4.09
Cardiothoracic	[a]	[a]	[a]	[a]	[a]	[a]
Cardiothoracic (2019)	[a]	[a]	[a]	[a]	[a]	[a]

[a] The technique was unable to perform imputation due to high computational costs.

Table A.3
Additional performance metrics for the KNNSCI method.

	AUROC(%)	Accuracy(%)	Precision(%)	Recall(%)	Specificity(%)	F1-Score(%)
RF classifier						
Wine	98.57 ± 0.98	91.10 ± 2.48	92.09 ± 2.00	91.49 ± 2.61	95.56 ± 1.31	91.07 ± 2.54
SPECT Heart	84.63 ± 5.58	84.28 ± 3.06	69.30 ± 5.38	71.85 ± 6.93	57.04 ± 11.81	66.38 ± 5.99
Statlog (Heart)	89.09 ± 2.07	81.11 ± 2.03	81.44 ± 1.91	80.97 ± 2.12	82.22 ± 3.17	80.81 ± 2.14
Osteoporosis	83.73 ± 3.13	75.65 ± 2.49	74.65 ± 2.46	75.34 ± 2.27	73.93 ± 2.77	74.80 ± 2.43
Cardiothoracic	71.32 ± 4.78	82.65 ± 2.98	56.89 ± 2.51	63.16 ± 4.07	85.95 ± 3.08	57.89 ± 3.44
Cardiothoracic (2019)	61.86 ± 1.00	84.17 ± 1.87	54.16 ± 0.49	56.32 ± 0.86	88.96 ± 2.31	54.58 ± 0.51
SVM classifier						
Wine	99.23 ± 0.59	93.51 ± 2.89	93.86 ± 2.54	94.10 ± 2.74	96.79 ± 1.47	93.60 ± 2.81
SPECT Heart	83.35 ± 6.00	73.49 ± 7.10	61.83 ± 5.72	74.10 ± 5.58	74.81 ± 6.87	59.52 ± 7.78
Statlog (Heart)	88.35 ± 2.45	80.66 ± 2.56	80.63 ± 2.49	80.56 ± 2.49	81.48 ± 3.28	80.45 ± 2.57
Osteoporosis	83.02 ± 2.58	75.41 ± 2.39	74.43 ± 2.57	75.31 ± 2.90	74.88 ± 5.56	74.62 ± 2.62
Cardiothoracic	68.91 ± 2.68	75.15 ± 2.61	54.85 ± 1.10	63.39 ± 3.21	77.17 ± 2.55	53.86 ± 1.94
Cardiothoracic (2019)	56.14 ± 2.63	78.35 ± 1.05	52.63 ± 0.50	55.81 ± 0.92	82.23 ± 1.14	52.10 ± 0.78
NB classifier						
Wine	98.30 ± 1.15	90.75 ± 3.08	92.04 ± 2.49	91.26 ± 3.15	95.31 ± 1.60	90.86 ± 3.11
SPECT Heart	76.60 ± 3.74	68.93 ± 6.07	58.24 ± 3.94	69.58 ± 3.93	70.37 ± 8.95	54.56 ± 5.39
Statlog (Heart)	86.19 ± 1.46	80.04 ± 1.45	80.30 ± 1.25	79.99 ± 1.33	80.44 ± 2.79	79.80 ± 1.48
Osteoporosis	77.89 ± 2.87	72.92 ± 3.43	71.59 ± 3.76	70.11 ± 3.90	57.63 ± 6.38	70.53 ± 3.89
Cardiothoracic	67.06 ± 4.26	83.60 ± 0.53	56.77 ± 1.30	62.58 ± 2.62	87.16 ± 0.82	57.94 ± 1.61
Cardiothoracic (2019)	57.94 ± 0.79	84.86 ± 0.45	53.21 ± 0.47	54.44 ± 0.54	90.10 ± 0.46	53.55 ± 0.53

Table A.4
Additional performance metrics for the CWRI method.

	AUROC(%)	Accuracy(%)	Precision(%)	Recall(%)	Specificity(%)	F1-Score(%)
RF classifier						
Wine	98.72 ± 0.56	92.40 ± 2.22	93.02 ± 1.95	92.95 ± 2.24	96.22 ± 1.13	92.44 ± 2.28
SPECT Heart	83.12 ± 6.26	83.12 ± 3.31	70.54 ± 10.13	70.54 ± 5.95	55.56 ± 10.30	66.04 ± 7.25
Statlog (Heart)	89.50 ± 2.26	81.40 ± 3.26	81.54 ± 3.27	81.21 ± 3.22	82.93 ± 4.29	81.12 ± 3.30
Osteoporosis	83.07 ± 3.66	75.71 ± 3.18	74.76 ± 3.10	75.50 ± 2.86	74.56 ± 2.22	74.92 ± 3.10
Cardiothoracic	71.10 ± 1.92	84.87 ± 3.67	57.44 ± 0.31	62.20 ± 3.02	88.57 ± 4.18	58.41 ± 0.74
Cardiothoracic (2019)	60.23 ± 1.39	85.51 ± 2.63	53.43 ± 1.22	53.97 ± 1.12	90.94 ± 3.05	53.38 ± 1.23
SVM classifier						
Wine	99.40 ± 0.44	93.00 ± 2.34	93.56 ± 2.08	93.58 ± 2.18	96.51 ± 1.20	93.10 ± 2.27
SPECT Heart	84.32 ± 5.76	72.98 ± 9.31	63.90 ± 10.71	75.68 ± 7.40	78.89 ± 6.62	60.51 ± 11.68
Statlog (Heart)	88.91 ± 1.84	81.84 ± 2.37	81.89 ± 2.48	81.64 ± 2.25	83.44 ± 3.74	81.59 ± 2.36
Osteoporosis	82.78 ± 2.41	74.50 ± 1.35	73.44 ± 1.48	74.22 ± 1.76	72.99 ± 3.91	73.63 ± 1.55
Cardiothoracic	68.58 ± 1.85	75.39 ± 1.80	54.74 ± 0.58	63.20 ± 2.49	77.47 ± 1.61	53.81 ± 0.73
Cardiothoracic (2019)	55.93 ± 3.24	78.36 ± 0.82	52.20 ± 0.59	54.81 ± 1.30	82.42 ± 0.88	51.59 ± 0.82
NB classifier						
Wine	98.12 ± 1.40	91.05 ± 3.25	92.19 ± 2.83	91.62 ± 3.12	95.47 ± 1.67	91.20 ± 3.22
SPECT Heart	75.11 ± 5.13	68.21 ± 5.96	60.15 ± 8.44	68.18 ± 6.49	68.15 ± 9.31	55.09 ± 8.19
Statlog (Heart)	86.28 ± 1.57	79.18 ± 2.15	79.96 ± 1.77	79.11 ± 2.16	79.74 ± 3.77	78.78 ± 2.29
Osteoporosis	77.55 ± 2.80	72.31 ± 3.50	70.95 ± 3.82	69.19 ± 4.04	55.27 ± 6.50	69.63 ± 4.06
Cardiothoracic	65.07 ± 3.50	83.29 ± 0.58	56.06 ± 1.81	61.22 ± 3.29	86.98 ± 0.92	57.02 ± 2.23
Cardiothoracic (2019)	57.56 ± 0.68	85.06 ± 0.19	53.61 ± 0.39	54.98 ± 0.52	90.24 ± 0.15	54.02 ± 0.45

References

Akoglu, H., 2018. User's guide to correlation coefficients. Turkish J. Emergency Med. 18 (3), 91–93.

Ambinder, E.P., 2005. Electronic Health Records. J. Oncol. Practice 1 (2), 57.

Bhagat, H.V., Singh, M., 2022. NMVI: A data-splitting based imputation technique for distinct types of missing data. Chemomet. Intell. Lab. Syst. 223, 104518.

Enders, C.K., 2022. Applied Missing Data Analysis. Guilford Publications.

Iranfar, A., Arza, A., Atienza, D., 2021. ReLearn: A Robust Machine Learning Framework in Presence of Missing Data for Multimodal Stress Detection from Physiological Signals. URL https://arxiv.org/abs/2104.14278.

Kang, M., Tian, J., 2018. Machine Learning: Data Pre-processing. Prognostics and Health Management of Electronics: Fundamentals, Machine Learning, and the Internet of Things, 111–130.

Khan, H., Wang, X., Liu, H., 2022. Handling missing data through deep convolutional neural network. Inf. Sci. 595, 278–293.

Little, R.J., Rubin, D.B., 2019. Statistical Analysis with Missing Data, vol. 793. John Wiley & Sons.

Liu, X., Lai, X., Zhang, L., 2019. A Hierarchical Missing Value Imputation Method by Correlation-Based K-Nearest Neighbors. In: Proceedings of SAI Intelligent Systems Conference. Springer, pp. 486–496.

Mishra, P., Mani, K.D., Johri, P., Arya, D., 2021. FCMI: Feature Correlation based Missing Data Imputation. arXiv preprint arXiv:2107.00100.

National Health and Nutrition Examination Survey Data, 2022. Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). URL https://www.cdc.gov/nchs/nhanes/index.htm.

Rockel, T., 2022. missMethods: Methods for Missing Data. R package version 0.3.0. URL https://CRAN.R-project.org/package=missMethods.

Rubin, D.B., 1976. Inference and Missing Data. Biometrika 63 (3), 581–592.

Schober, P., Boer, C., Schwarte, L.A., 2018. Correlation coefficients: appropriate use and interpretation. Anesthesia Analgesia 126 (5), 1763–1768.

Sefidian, A.M., Daneshpour, N., 2020. Estimating missing data using novel correlation maximization based methods. Appl. Soft Comput. 91, 106249.

Tabarestani, S., Aghili, M., Eslami, M., Cabrerizo, M., Barreto, A., Rishe, N., Curiel, R.E., Loewenstein, D., Duara, R., Adjouadi, M., 2020. A distributed multitask multimodal approach for the prediction of Alzheimer's disease in a longitudinal study. NeuroImage 206, 116317.

Van Buuren, S., Groothuis-Oudshoorn, K., 2011. mice: Multivariate imputation by chained equations in R. J. Stat. Softw. 45, 1–67.

Yoon, J., Zame, W.R., van der Schaar, M., 2019. Estimating missing data in temporal data streams using multi-directional recurrent neural networks. IEEE Trans. Biomed. Eng. 66 (5), 1477–1490.

Zhou, L., Rueda, M., Alkhateeb, A., 2022. Classification of breast cancer nottingham prognostic index using high-dimensional embedding and residual neural network. Cancers 14 (4).