A review on missing values for main challenges and methods

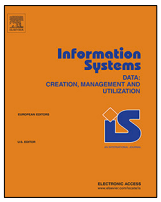
Lijuan Ren, Tao Wang, Aicha Seklouli-Sekhari, Haiqing Zhang, Abdelaziz Bouras

Information Systems 119 (2023) 102268

Contents lists available at ScienceDirect

Information Systems

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



A review on missing values for main challenges and methods

Lijuan Ren [a,*], Tao Wang [b], Aicha Sekhari Seklouli [c], Haiqing Zhang [a], Abdelaziz Bouras [d]

[a] Chengdu University of Information Technology, School of Software Engineering, 610225, Chengdu, China
[b] Univ Lyon, Univ Jean Monnet Saint-Etienne, INSA Lyon, Univ Lyon 2, Université Claude, France
[c] Univ Lyon, Univ Lyon 2, INSA Lyon, Université Claude Bernard Lyon 1, DISP-UR4570, 69676, Bron, France
[d] Qatar University, CSE, College of Engineering, 2713, Doha, Qatar

ARTICLE INFO

Article history:
Received 18 July 2023
Received in revised form 9 August 2023
Accepted 9 August 2023
Available online 11 August 2023
Recommended by Dennis Shasha

Index terms:
Missing values
Imputation
Deletion
Missing mechanism
Machine learning

ABSTRACT

Several recent reviews summarize common missing value analysis methods. However, none of them provide a systematic and in-depth summary of the analytical challenges and solutions for dealing with missing values. For the purpose of guiding the handling of missing values, this review aims to consolidate current developments in novel missing-value research methodologies. In particular, we comprehensively investigated cutting-edge missing value solutions and methodically studied the main challenges associated with missing values analysis (missing mechanisms, missing patterns, and missing rates). Furthermore, we reviewed 63 publications that compare different strategies for deleting and imputing missing values. Then we investigated data characteristics, highlighted three main problems when analyzing missing values, and analyzed the performance of missing value solutions in these studied papers. Moreover, we conducted comprehensive experiments on 9 public datasets using typical missing value processing methods and provided a simple guided decision tree for handling missing values. Finally, we described current Research hotspots and open challenges, which give potential research topics.

Contents

* Corresponding author.
E-mail address: renlijuan@cuit.edu.cn (L. Ren).

https://doi.org/10.1016/j.is.2023.102268

1. Introduction

With the rapid development of information technology, people desire to retrieve hidden but useful information from growing data. However, missing data is an unavoidable problem for data analysis. The common reasons for missing values (MVs) are diverse, including respondents in the household survey may refuse to report income; in industry experiments, some results are missing because of mechanical failures unrelated to the experimental process; in medical experiments, some participants drop out because of drug allergies, deaths or other reasons [1]. To sum up, these reasons can be roughly divided into four types, including (1) human mistakes when processing data, (2) machine error caused by equipment malfunction, (3) respondents' refusal to answer specific questions, (4) drop-out from studies and merging unrelated data [2–4]. Missing data is unavoidable, despite the fact that we are all aware that gathering as much data as possible is the ideal strategy for data analysis. Although it is commonly known that erasing missing information is simple and quick, several studies have come to the conclusion that this approach does not work in all situations [5,6]. For instance, removing will result in the loss of some important data when the missing value is not entirely random [7,8].

Other researchers investigated novel approaches known as missing values imputation, which substitutes plausible values for missing variables [9–11]. This approach can retain more data than deletion, but it takes time to produce reasonable values [12]. To sum up, deletion and imputation are commonly used techniques for handling missing information, although there is a wide range of opinions regarding their performance and application scenarios. Some review papers about missing values have been published. Sinharay et al. [8] clearly introduced the basic knowledge of processing methods that deletion and imputation for missing values, but they mainly paid more attention to imputation methods of single imputation and multiple imputation. In another review, García-Laencina et al. [13] introduced four techniques to deal with missing values in pattern classification including deletion of incomplete cases, imputation, model-based procedures, and machine learning procedures. In addition, Jadhav et al. [14] concentrated on the effectiveness of various imputation techniques. In order to assess the performance of seven imputation methods, including mean imputation, median imputation, K nearest neighbor (KNN) imputation, predictive mean matching, Bayesian Linear Regression (norm), Linear Regression, non-Bayesian (norm. nob), and random sample, they reviewed some papers about performance comparison, but this study only took into account datasets with numerical variables. On the other hand, Lin and Tsai [15] investigated and analyzed 111 journal articles that were released between 2006 and 2017. They outlined a few issues with these studies, including the small size of experimental datasets, and the lack of attention to missing mechanisms. Recently, Emmanuel et al. [2] compiled some literature with a focus on machine learning methods. They tested with the KNN and random forest (RF) imputation techniques at the same time, however, they only employed two tiny datasets, the Iris and ID fan datasets [16].

2. Necessity of missing values review

Overall, these review papers summarize common missing value analysis methods, but none of them provide a systematic and in-depth summary of the analytical challenges and solutions for dealing with missing values. They have the following drawbacks: (1) lack of comparative analysis and review of deletion and imputation performance; (2) lack of analysis in major challenges of missing value analysis and process; (3) lack of analysis of performance indicators on different research tasks; (4) lack of guidance on how to deal with missing values. Consequently, we address these four issues simultaneously in this work. Specifically, we first outline the three primary challenges in missing value analysis, namely missing mechanisms, missing patterns, and missing rates. Then, we explore and evaluate approaches for dealing with missing values. Then, we investigated a large number of papers focusing on the performance comparison of popular missing value processing methods, and we conducted an in-depth comparative analysis of missing value processing methods according to the included papers. Meanwhile, we summarized some rules based on the research results to help readers choose missing value processing methods. Overall, our study addresses four drawbacks with missing value review papers while providing four novel contributions: (1) We reviewed and analyzed the experimental results of numerous studies to verify that the imputation method generally outperforms the deletion method; (2) We then thoroughly studied the experimental results of the included studies in order to analyze the situations in which the missing value deletion method was appropriate and provide some useful rules to guide readers in choosing missing value processing

Table 1
The introduction of missing mechanism.

Missing mechanism	Description	Condition expression	Example
MCAR	The probability of missing variables is independent of the variable itself and any other external influences	$p\,\|M\|\xi\,\|$	A blood value is missing because the blood sample is broken by accident or a questionnaire is accidentally lost
MAR	The likelihood of a missing value in MAR is traceable or predictable from the observable data.	$p\,\|M\|X_o, \xi\,\|$	Women tend not to report age and weight in questionnaires. Thus, the missingness of variable 'weight' depends on the variable 'sex'.
NMAR	The pattern of data missingness is non-random and depends on the missing variable	$p\,\|M\|X_o, X_m, \xi\,\|$	In the income survey, low-income people do not respond. Thus, the missingness of variable 'income' depends on itself.

methods. (3) We provided a simple but guided decision tree by conducting comprehensive experiments in 9 public datasets. (4) In order to provide potential research topics for future studies, we analyzed and summarized the existing research hotspots and open challenges of missing values.

The rest of the paper is organized as follows: Section 3 lists the three main challenges of missing value analysis. The typical missing value processing methods are described in Section 4. Performance comparisons of popular missing value processing techniques are presented in Section 5. The 6 Section provides a detailed introduction to the experimental design and outcome analysis. We analyze current research hotspots and open challenges in Section 7. The paper is concluded in Section 8.

3. Main challenges of missing value analysis

In real-world research, the investigation of missing values is essential because it is unavoidable and makes typical data mining techniques challenging to use. Three factors need to be considered when analyzing missing values: the reason for missing data (missing mechanism), the location of missing data (missing pattern), and the amount of missing data (missing rate). When missing values are observed in research, these three factors pose the main challenges for missing value analysis. The analysis of the causes of missing data is the most challenging of these since, in practice, these causes are complicated and influenced by plenty of external factors. Choosing the appropriate processing methods for various missing patterns and rates in missing value analysis is also challenging despite the fact that the location of missing values (missing pattern) and the number of missing values (missing rate) can be easily expressed formally. For instance, the processing methods employed may change depending on whether all the missing values are concentrated in one column or are dispersed across several columns. At the same time, it is difficult to determine when missing values can be dropped directly. In summary, the three basic issues in missing value analysis – missing mechanisms, missing patterns, and missing rates – are crucial in determining how missing values should be treated. Next, an in-depth introduction to the three main factors of missing value analysis will be presented.

3.1. Missing mechanism

In most cases, it is critical to identify the type of missing data in order to choose the appropriate missing data approach. Actually, Little and Rubin [1] distinguish three different categories of missing data mechanisms: Missing Completely At Random (MCAR), Missing At Random (MAR), and Not Missing At Random (NMAR). Let X be the matrix representing the full dataset, where X_o and X_m stand for the observed and missing data, respectively. Let M represent a missing value matrix, where M has a value of 0 if X is observed and 1 otherwise. Let ξ stand for a vector of values indicating the relationship between the missingness in M and the dataset X. We provided details about various missing types, including descriptions, conditional expressions, and examples. Here, the condition expression is defined by the probability of whether a value is observed or missing. Table 1 illustrates the introduction of the missing mechanism.

In contrast to MCAR and MAR, NMAR is frequently regarded as the worst missing mechanism since it easily produces biased results [17]. Several studies advised recovering as much missing data as possible [18]. Unfortunately, we frequently encounter the problem of missing data when collecting experimental data, and it is challenging to recover them. As a result, it is an effective strategy to make an effort to identify the missing mechanisms and choose the best treatment methods for dealing with them. Even if it is not yet able to confirm whether missing is caused by MAR or NMAR, the Chi-square [19] test can assist to distinguish the MCAR missing mechanism [1]. For instance, the chi-square test can reveal that women have a higher percentage of missing data than men on the weight variable if women are truly less likely than men to report their weight.

3.2. Missing pattern

Missing patterns can be used to describe missing or observed values in a dataset and to illustrate the relationship between missingness and variable values in a data matrix [1,2]. The literature does not, however, provide a common description of missing patterns for missing values. We outlined the three currently popular categories. As shown in Table 2 and Fig. 1, the first category uses six missing patterns that concentrate on the causes of missing values.

Then, Emmanuel and Tlamelo [2] described the three types of missing data patterns that are most common in the literature for the second category: univariate, monotone, and non-monotone. The terms "Univariate" and "Monotonic" have the same definitions as "Univariate Nonresponse" and "Monotonic", respectively, in Table 2. "Non-Monotone" denotes that the absence of one variable has no bearing on the absence of any other variables. Specifically, examples of the second category of missing patterns are shown in Fig. 2.

For the third category, some researchers choose to simulate datasets with missing values of various structures to conduct experiments in order to analyze the effectiveness of various processing algorithms for missing values [20–25]. According to the complexity of the missing data, four missing patterns were identified for this classification. It includes simple, medium, complex, and blend patterns. The third category of missing patterns is shown in Table 3.

According to the description of the third category of missing pattern, the missing rate must be taken into account in the datasets with missing values for this classification. Let us first assume that dataset X consists of 20 rows and 5 columns, and then we use the 10% missing rate for the entire dataset as an example. Thus, each pattern included 10 missing values (row $\times$

Table 2
The first category of missing patterns.

Missing pattern	Description	Example	Legend
Univariate Nonresponse	Missingness is confined to a single variable	In the context of agricultural trials, the units are experimental plots	Fig. 1(a)
Unit and Item Nonresponse	All observed or missing on the same set of units	A subset of sampled individuals does not complete the questionnaire because of noncontact, refusal.	Fig. 1(b)
Monotone	Units drop out prior to the end of the study and do not return	In a clinical trial, some units may drop out for unknown reasons	Fig. 1(c)
Haphazard	Missing values distribute haphazard on some items	Some questions in the questionnaire were not responded	Fig. 1(d)
Variables Never Jointly Observed	Some variables are never observed together	The File-Matching Problem	Fig. 1(e)
Patterns with Latent Variables	Latent variables are completely missing	Factor Analysis	Fig. 1(f)

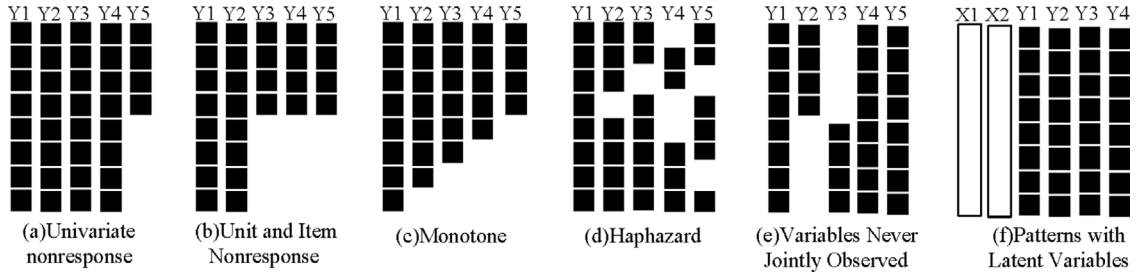


Fig. 1. Examples of the first category for missing patterns. There are six patterns where rows represent instances and columns represent variables of data. White is a missing value, while black represents the observed value.

Table 3
The third category of missing patterns s.

Missing pattern	Description	Legend
Simple Pattern	Each record at most has one missing value	Fig. 3(a)
Middle Pattern	A record can have missing values between 2 and 50% of the number of the variables	Fig. 3(b)
Complex Pattern	A record can have missing values between 50% and 80% of the number of the variables	Fig. 3(e)
Blend Pattern	It is a mixture of other three patterns, where 25% records are simple pattern, 50% records are medium pattern and 25% records are complex pattern.	Fig. 3(d)

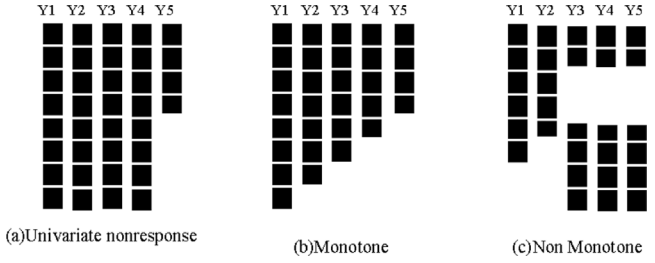


Fig. 2. Examples of the second category for missing patterns.

column × missing rate = 20 × 5 × 0.1 = 10). The examples of the third category of missing patterns as shown in Fig. 3. It is worth noting that when there are many columns (variables) in the dataset, it is impossible to create simple patterns with a high missing rate.

3.3. Missing rate

The missing rate is one of the crucial metrics to measure the number of missing values in a dataset. The pattern of missing data and the proportion of missing data, particularly when the percentage of missing data surpasses 40%, have a considerable negative impact on the accuracy of prediction (or imputation), according to Song et al. [26]. When choosing approaches for processing missing data, Diane et al. [27] stressed that it is important to consider the percentage of cases with missing data. Additionally, two studies have demonstrated that the effectiveness of imputation gradually declines as the missing rate rises [25,28]. These studies demonstrate the significance of missing ratios in missing value analysis. The total missing rate, which displays the missing situation across the entire dataset, is a popular way to depict the number of missing values [29–31]. In addition, the missing situation of rows and columns was also described in various works using the row missing rate [32–34] and the column missing rate [30,35,36]. The percentage of rows in a dataset with missing values is known as the row missing rate. The number of missing values for each column having missing values is indicated by the column missing rate.

For clarity of definitions, we assume that dataset X includes n rows (instances) and k columns (attributes). Let M represent a missing value matrix, where m_{ij} has a value of 0 if any value $x_{ij}(i \leq n, j \leq k)$ in X is observed and 1 otherwise. The total missing rate MR can be represented as

$$MR = \frac{\sum_{i=1}^{n} \sum_{j=1}^{k} m_{ij}}{m \times n} \tag{1}$$

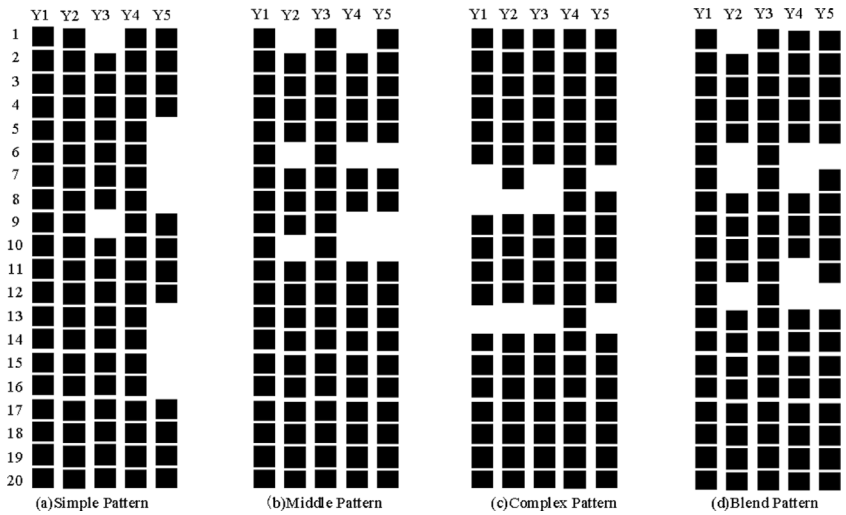


Fig. 3. Examples of the third category for missing patterns.

The row missing rate of data can be calculated by

$$MR_r = \frac{\sum_{i=1}^{n} \tau, \tau = \begin{cases} 0, \forall m_{ij} \neq 0 \\ 1, \exists m_{ij} = 1, \end{cases} \quad j \epsilon [1, k]}{n} \tag{2}$$

The column missing rate of data can be computed by

$$MR_c = \frac{\sum_{j=1}^{k} \tau, \tau = \begin{cases} 0, \forall m_{ij} \neq 0 \\ 1, \exists m_{ij} = 1, \end{cases} \quad i \epsilon [1, n]}{k} \tag{3}$$

In addition, the missing rate of each row and column is also important. The missing rate of lth row (denoted by r_l) can be computed by

$$MR_{r_l} = \frac{\sum_{j=1}^{k} m_{lj}}{k} \tag{4}$$

The missing rate of lth column (denoted by c_l) can be computed by

$$MR_{c_l} = \frac{\sum_{i=1}^{n} m_{lj}}{n} \tag{5}$$

In reality, in order to achieve various experimental objectives, it is important to investigate various missing rate aspects. For instance, in the statistical analysis of medical data, records (patients) with missing rates are typically given more attention; in feature selection trials, the missing rate of features (columns) is typically the emphasis, and some characteristics with a high missing rate may be disregarded. Even if the equations for calculating the missing rate are quite simple, research on missing values should specify the approach employed, which helps other researchers avoid misinterpretation.

4. Processing methods for missing values

It is a known reality that choosing an appropriate approach to deal with the missing values is challenging. Researchers studying Management Information Systems (MIS) [37] have discovered that researchers rarely explicitly discuss the existence and treatment of missing values. As a result, when missing data are present, they frequently employ the listwise and pairwise deletion techniques. There have been numerous ways of handling missing values, but the classification methods for handling missing values in published works varied slightly. For example, Little et al. [1] categorize the processing techniques into 4 categories: procedures based on fully recorded units (i.e., complete case analysis), weighing procedures, imputation, and model-based techniques. Weighting processes and model-based methods are less frequently employed than deletion and imputation methods, hence some research has classified processing methods for missing values into these two categories: deletion and imputation [2,12,14,15,38–40]. Similarly, since they are the most widely used approaches and are typically applied without any limitations, deletion and imputation procedures for missing values are the focus of our work.

4.1. Deletion methods

Missing value deletion, also known as disregarding missing values, is the process of explicitly deleting instances or variables that contain missing data items to solve the problem of missing data [2]. Although a test pattern with missing values cannot be classified since the deletion procedure would ignore it, deletion methods have the advantage of allowing the normal pattern classification methods to be used directly for complete data [14]. For ignoring missing data, there are two general strategies [1,2,14]. First, Listwise Deletion (LD), also known as case-wise deletion, or case-removal, is a technique for removing instances (rows, cases) with missing data. This technique is also known as complete case analysis because it only keeps complete cases for analysis (CCA). The analysis is then restricted to those observations for which all values are observed, which frequently leads to biased estimates and loss of precision [17] because this method excludes all cases with missing values for any variable of interest. The second technique is known as Pairwise Deletion (PD) or Available Case Analysis (ACA), also referred to as variable deletion, and it is used to delete variables (columns) with missing data. This method analyses all situations in which the variables of interest are present, using as much data from each case as is feasible rather than excluding the entire case. Even though some of its variables have missing values, it can nevertheless maintain the most amount of data possible for analysis since it uses distinct sample sizes for each variable [17]. As a result, the ACA approach has a larger sample size than the CCA method. Additionally, based on studies from [37], we contrasted these two deletion techniques, as shown in Table 4.

4.2. Imputation methods

There are numerous imputation strategies for missing values in contrast to deleting them. As an illustration, Alireza Farhangfar

Table 4
Comparison of two deletion methods.

Name	Description	Scene	Advantage	Disadvantage
Listwise Deletion	Eliminate all cases with predictor variables or standard variables with any number of missing data	The number of missing cases is small relative to the sample size, and missing mechanism is MCAR	Simple operation, no calculation	When the pattern of missing values is more random, the number of items eliminated is higher
Pairwise Deletion	When calculating different parameter estimates, the model based on missing values uses different sample sizes *	The proportion of missing cases for each variable is small relative to the sample size, and missing mechanism is MCAR	It can retain cases with missing values	Reduce the sample size, the correlation matrix generated by the result data is difficult to interpret

Note*: Inconsistent correlations or non-positive definite covariance matrices can occasionally result from using different sample sizes [37]. To estimate the components of the cross-correlation matrix, PD thus only uses variables with non-missing items, although the statistical parameters are based on various datasets and sample sizes, which typically result in various standard errors.

Table 5
The comparison of single and multiple imputation.

Method name	Common methods	Advantage	Disadvantage
Single imputation	Imputation with constant, Mean Imputation, Regression Imputation, KNN Imputation	Can use predictive distribution for imputation. Completes dataset by filling in missing values.	Underestimates variance, leads to under coverage of confidence intervals
Multiple imputation	Multivariate imputation by chained equations (MICE)	Considering the uncertain of missing values	Requires additional steps and high complexity

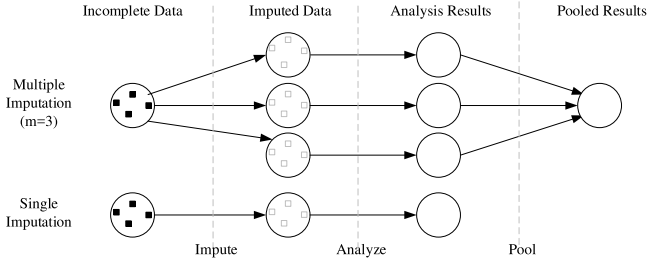


Fig. 4. The imputation process of single and multiple imputation. Black squares represent missing values. The $\boldsymbol{m}$ imputed data ($\boldsymbol{m} = 3$) be generated in multiple imputation.

et al. [12] separated the imputation into two categories: (1) model-based, and (2) quasi-randomization inference-based. Generally, there are two common categories for imputation methods. The first category separates imputation methods into single and multiple imputation [6,14] methods, while the second category divides imputation methods into imputation based on statistics and machine learning [13,15,38,41].

4.2.1. Single and multiple imputation

The first categorization is determined by whether the approach considers the uncertainty of missing values. Single imputation (SI) refers to one value for a missing data element that is filled in without defining an explicit model for the partially missing data. Single imputation techniques include mean value, hot and cold deck, and even regression imputation. Obviously, the single imputation approach ignores the uncertainty of missing data. As a result, the multiple imputation approach (MI), which incorporates the method's uncertainty into the estimated value, was proposed. For a dataset containing missing values, the MI process produces m imputed results, and the final imputed result is produced by combining the m results. Numerous researchers have chosen the MI approach, which is frequently utilized in a variety of domains [30,42–45]. Fig. 4 depicts the single and multiple imputation processes.

Furthermore, SI and MI have different advantages and disadvantages. The common methods of the two methods and their advantages and disadvantages are presented in Table 5 (was updated based on [18]).

4.2.2. Imputation methods based on statistics and machine learning

The second category of methods is classified according to whether the model is inspired by machine learning. Specifically, early approaches for imputing missing data were specifically motivated by traditional statistical models and estimate processes, which are referred to as imputation methods based on statistics. These techniques are designed to model the information included in the non-missing parts of the dataset in order to as correctly estimate the missing values as possible [41]. Researchers initially substituted missing values with the mean, median, mode, and zero values. The disadvantage is that when there are numerous missing data, a significant portion of the data is replaced by the same value (i.e., mean, median, mode, zero), which can easily lead to serious deviation. The mean imputation approach should not be used, according to certain recent research that has demonstrated its shortcomings [44,46]. The in-depth study on missing values has been accompanied by the proposal of a number of innovative techniques. For instance, the LS (Least Squares) imputation approach is based on the least squares principle to estimate missing values, whereas the hot-deck imputation method predicts missing values by seeking the nearest neighbor using non-missing information [47].

However, machine learning-based imputation approaches are complex processes that often include building a predictive model to estimate values that will substitute those missing [13]. The machine learning-based imputation method often involves building a predictive model to predict the values for missing data. Many machine learning-based imputation methods have been proposed recently, and these methods frequently produce good imputation results. Examples of these methods include imputation methods based on decision trees (DT) [48,49], imputation using multilayer perceptrons [50], imputation using artificial neural networks (ANNs) [51], and imputation using self-organizing maps (SOMs) [52]. Fig. 5 depicts the processing flow of machine learning-based imputation approaches.

Obviously, there is a range of imputation techniques based on statistical analysis and machine learning. According to a review on missing value imputation from 2006 to 2017 [15], expectation management (EM), linear regression (LR), least squares (LS), and mean/mode are the top four most often employed statistical techniques in imputation technologies, while Clustering, Decision Tree (DT), KNN, and RF are the top four machine learning approach used in imputation technologies.

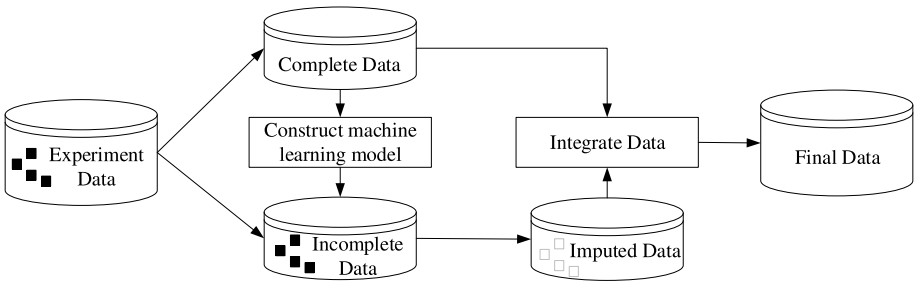


Fig. 5. The processing flow of imputation methods based on machine learning. Black squares in the diagram stand in for missing values. The experiment data is separated into two categories in the processing flow: complete data without any missing values and incomplete data with missing values. Finally, the complete data and the imputed data are combined to create the final data. First, the appropriate machine learning technique is used to build a model using the complete data. Next, the generated model is used to predict the missing values for the incomplete dataset.

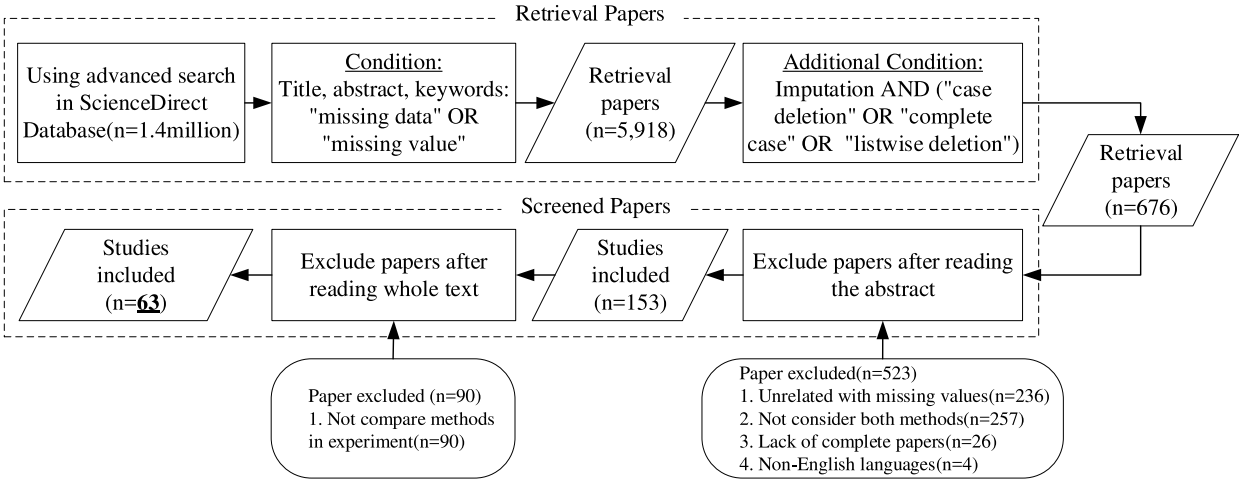


Fig. 6. The flowchart of the literature screening process.

5. Comparison of common processing methods

In the previous section, we introduced the two well-liked techniques of deletion and imputation for handling missing values in the previous section. Furthermore, we discovered that there are numerous research comparing various approaches for deleting and imputing missing data. In order to assist choose a more suitable approach to deal with missing values, these studies compare the effects of two missing value processing methods on the research results. The applicability of the two study methodologies was investigated in order to better explore the state of the research on missing value processing techniques. After conducting an exhaustive literature search in ScienceDirect, we eventually examined and analyzed 63 works from 1994 to 2021 with a focus on the contrast between deletion and imputation of missing values. Fig. 6 depicts the flowchart of the literature screening process.

The publication years of the papers included in the study were first counted in order to analyze the interest of researchers in this topic year by year, as shown in Fig. 7.

We observe that studies comparing deletion and imputation of missing values have become more and more concentrated since 2005. In addition, since 2005, academics have continued to emphasize comparing two widely used approaches, indicating that further study on this topic is still essential.

5.1. Overview of studied research

We need to evaluate from many perspectives in order to improve statistics and analyze the research on missing value processing techniques in these studies. We specifically investigate from six perspectives, including the type of experimental data, mechanism of missing values, missing rate, missing pattern, comparative indicators, and experimental results. The next step is to classify these six dimensions.

Category of dimension 1: the type of experiment datasets contains real, semi-simulated (Semi), and simulated (Sim). Real data means that the data was gathered from the outside world and that any missing values were produced naturally. Semi-simulated data is data that is gathered from the real world but has artificially manufactured missing values. Simulation data refers to datasets and missing values are created artificially. Some researchers employed datasets with missing values to conduct experiments specifically to examine the effectiveness of deletion and imputation of missing values, although it is challenging to know the causes and the true values for missing values. To conduct the comprehensive analysis, they simulate experimental data with various missing mechanisms. Fig. 8 illustrates the experimental procedure of the three data types in the studied papers.

Category of dimension 2: the type of missing mechanisms includes MCAR, MAR, NMAR, and Other. As mentioned before, missing processes can be categorized into three groups: MCAR, MAR, and NMAR. While several studies employed real datasets, their missing mechanisms could not be identified or analyzed, hence we will use "other" in those cases. However, we will carry out a more detailed analysis in the following sections to provide a clearer understanding of the missing mechanisms in these investigations.

Category of dimension 3: the type of missing rates includes row, column, and total.

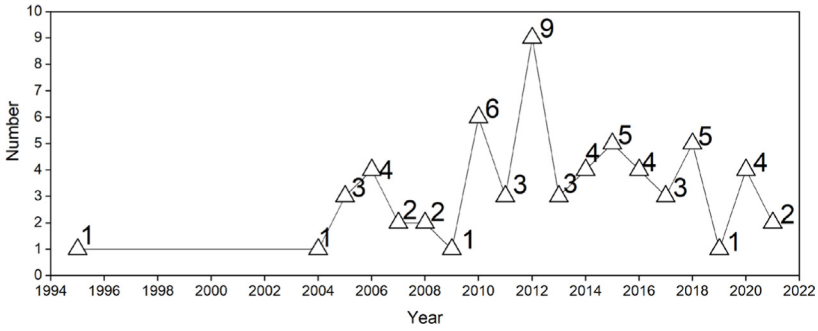


Fig. 7. The publication years of the syudied papers.

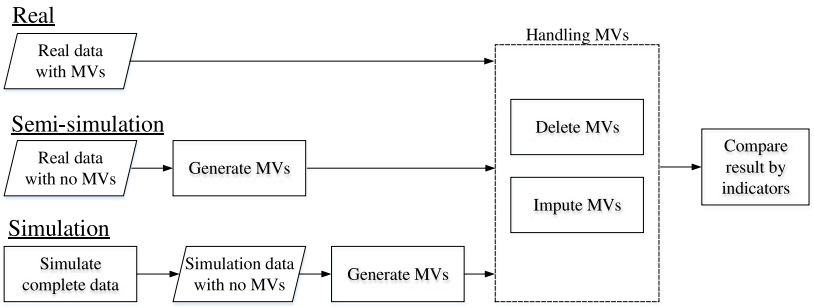


Fig. 8. The experimental process of three data types.

Table 6
The type of experiment results.

Category	Classification	Description
First	Imputation better	Experimental results show that imputation methods have better performance than deleting missing values
Second	Similar Performance	Experimental results show that deletion and imputation have similar performance
Third	Discussion	Experimental results show that the deletion and imputation methods for missing values are suitable for different scenarios

Category of dimension 4: the type of missing patterns includes univariate (Univ), monotone (Mon), and non-monotone (Non-Mon). As we analyzed earlier, missing pattern classifications are diverse. Additionally, as the research paper only provided a rough analysis of the missing patterns, we adopted a simple classification that proposed by Emmanuel and Tlamelo [2] as our survey categorization.

Category of dimension 5: the type of comparison matrix includes statistics and learning. We also pay attention to the research's comparison indicators, which are classified into statistical and machine learning indicators. The statistical indicators refer to the employment of various approaches to deal with missing values in the research, followed by the comparison and analysis of the processed data using widely used statistical indicators (i.e., mean, error). Learning indicators refer to the use of different methods to deal with missing values in the research, followed by utilizing machine learning techniques such as prediction and clustering to further mine the processed data, and finally employ various indicators like prediction accuracy, AUC for comparison.

Category of dimension 6: the type of experiment results includes imputation better (First), similar performance (Second), and discussion (Third). As indicated in Table 6, we divided the papers into three categories based on the experimental results and summaries in order to compare the performance differences between the deletion and imputation approaches in the research papers. We do not categorize deletion approaches as being superior because there is no research demonstrating that they can perform better across all experiments. Certainly, in those studies that belong to the category of "Similar Performance", deletion approaches and imputation approaches perform similarly. Additionally, in some experiments from those papers that appear under the "Discussion" category, deletion methods may perform better than imputation methods.

The articles included in the study were then reviewed and analyzed in accordance with the classification of the six introduced dimensions, the year that the literature was published, and whether or not new approaches were suggested in the study. Ultimately, an overview Table 7 was obtained.

5.2. Analysis of experimental datasets

5.2.1. Dataset domains

First, we conduct statistics and analysis on the data domains of the studied papers in order to identify which research fields pay more consideration to the influence of various approaches for handling missing values. Particularly, 74.6% of the papers had a medical theme, including epidemiology, medical trials, and economic analyses. The primary reason is that medical trials are prone to missing values, such as patients leaving the experiment early because of poor compliance, unfavorable occurrences, or inadequate efficacy. Two of these publications [90,95] are related

Table 7
The overview of studied research on six dimensions.

Literature	Year	Data type			Missing mechanisms				Missing rates			Missing patterns			Comparison indicator		New method	Experiment result		
		Real	Semi	Sim	MCAR	MAR	NMAR	Other	Column	Row	Total	Univ	Mon	Non-Mon	Statistic	Learning		First	Second	Third
[32]	2021	✔		✔	✔	✔	✔			✔			✔		✔			✔		
[33]	2015	✔				✔	✔		✔	✔		✔			✔		✔	✔		
[35]	2016	✔		✔		✔			✔				✔		✔		✔	✔		
[53]	2010	✔		✔	✔	✔	✔	✔	✔	✔		✔	✔		✔		✔	✔		
[54]	2018		✔		✔				✔	✔		✔			✔		✔	✔		
[34]	2017	✔				✔				✔			✔		✔			✔		
[55]	2014			✔		✔			✔				✔		✔		✔	✔		
[56]	2006	✔						✔	✔	✔			✔		✔	✔		✔		
[57]	2016	✔		✔		✔			✔	✔		✔			✔		✔	✔		
[58]	2016	✔	✔		✔	✔	✔	✔	✔	✔			✔		✔			✔		
[59]	2011			✔	✔	✔	✔		✔	✔		✔			✔		✔	✔		
[60]	2005	✔						✔	✔	✔		✔			✔			✔		
[61]	2007	✔				✔				✔			✔		✔				✔	
[29]	2006		✔		✔	✔			✔		✔		✔		✔			✔		
[31]	2015		✔		✔	✔	✔				✔		✔		✔		✔			✔
[44]	2014			✔	✔	✔	✔			✔	✔		✔		✔			✔		
[62]	2018	✔						✔		✔			✔		✔			✔		
[63]	2007		✔		✔					✔			✔		✔					✔
[42]	2009	✔				✔			✔	✔		✔			✔			✔		
[64]	2019		✔			✔					✔		✔		✔			✔		
[65]	2010		✔		✔	✔			✔	✔		✔			✔					✔
[66]	2018	✔				✔			✔	✔			✔		✔					✔
[67]	2012	✔						✔		✔			✔		✔			✔		
[68]	2010	✔				✔			✔		✔		✔		✔	✔		✔		
[69]	2017	✔		✔	✔	✔	✔	✔	✔	✔		✔	✔		✔					✔
[70]	2015		✔		✔	✔			✔	✔		✔			✔			✔		
[71]	2018	✔				✔			✔	✔			✔		✔			✔		
[72]	2015	✔						✔	✔		✔		✔			✔	✔			✔
[73]	2021	✔		✔		✔		✔	✔	✔		✔			✔		✔	✔		
[74]	2010	✔				✔				✔				✔	✔		✔			✔
[75]	2018		✔		✔	✔					✔		✔		✔		✔			✔
[76]	2011	✔				✔				✔			✔		✔			✔		
[77]	2020	✔				✔			✔	✔			✔		✔			✔		
[78]	2008			✔		✔			✔	✔		✔			✔					✔
[79]	2014	✔		✔	✔	✔	✔	✔	✔		✔		✔		✔			✔		
[80]	2012	✔				✔			✔	✔	✔		✔		✔					✔
[36]	2015	✔						✔	✔				✔			✔				✔
[81]	2014		✔		✔	✔				✔			✔		✔				✔	
[82]	2020			✔	✔	✔	✔			✔			✔		✔					✔
[83]	2013		✔		✔		✔			✔				✔	✔			✔		
[84]	2012	✔						✔		✔			✔		✔			✔		
[85]	2013		✔		✔	✔			✔	✔		✔			✔					✔
[86]	2010		✔			✔			✔				✔		✔	✔		✔		
[87]	2020		✔				✔			✔	✔		✔			✔		✔		
[88]	2017	✔			✔	✔	✔		✔	✔		✔			✔			✔		
[27]	2010		✔					✔	✔	✔		✔			✔					✔
[89]	2012		✔				✔			✔				✔	✔			✔		
[90]	2006		✔		✔	✔	✔				✔		✔		✔			✔		
[91]	2004		✔		✔	✔					✔		✔		✔					✔
[92]	2020	✔				✔			✔	✔			✔		✔			✔		
[93]	2012		✔			✔			✔	✔			✔		✔		✔	✔		
[94]	2012			✔	✔	✔				✔			✔			✔				✔
[95]	2016		✔		✔	✔	✔			✔			✔			✔		✔		
[96]	2008		✔		✔	✔	✔			✔			✔		✔	✔		✔		
[97]	2012		✔		✔						✔		✔		✔					✔
[98]	2012			✔		✔				✔			✔		✔			✔		
[99]	2005			✔	✔						✔		✔		✔		✔	✔		
[100]	1995			✔	✔						✔		✔		✔					✔
[101]	2013	✔						✔	✔	✔		✔			✔			✔		
[102]	2006		✔		✔	✔	✔		✔	✔		✔			✔					✔
[103]	2005			✔		✔			✔				✔		✔					✔
[104]	2012		✔					✔	✔	✔		✔			✔					✔
[105]	2011	✔						✔	✔				✔		✔					✔

to software development in addition to the medical field. Since missing values can make it more difficult to use the data to create an accurate prediction system for software development workload, these researchers investigated the effectiveness of strategies of two strategies for handling missing values in this context. In addition, two research on social networks [94,96] explored missing values in the network structure and discovered that the handling is taken into account as well as the amount and kind of missing data that can affect the depicted block model structure. In particular, a study [36] discusses the case of encountering missing values in audio data in detail. Dataset type

In order to further analyze the usage frequency and patterns of various dataset types in the study, we utilized histograms to compare the number of quantities and combinations of various dataset types, as shown in Fig. 9.

We noticed that researchers tended to compare the effectiveness of removing and imputing missing values using real datasets more frequently (43.66%). Second, we can observe from

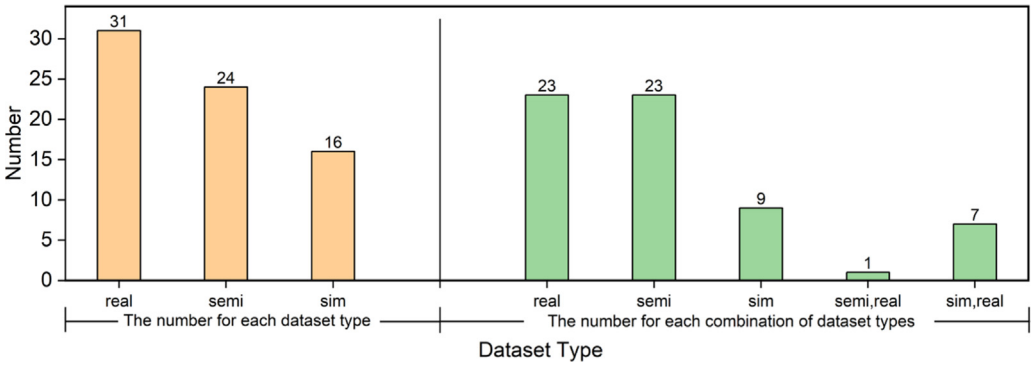


Fig. 9. The quantity and combinations of various data types. "semi" and "sim" in the figure stand for simulation and semi-simulation, respectively. The total number of various dataset types is shown on the left. The number of combinations of data types used in the studies is shown on the right.

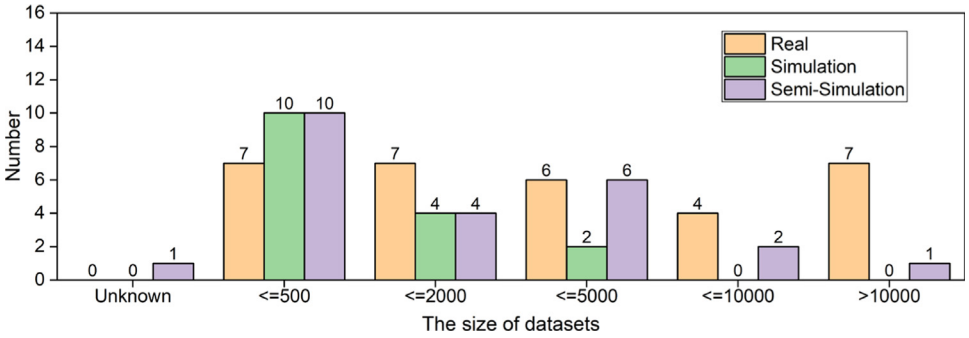


Fig. 10. Sizes of different experimental dataset types.

the right side of Fig. 10 that only a few studies (12.70%) employed different types of datasets for experiments. Since missing mechanisms and missing patterns may be clearly exhibited and results can be quantitatively compared, the researchers in this case used semi-simulation or simulation datasets first, followed by real datasets. In these studies papers, some simulated data were created depending on the traits of the real data, such as the literature [58].

5.2.2. Dataset size

The size of the dataset must also be taken into account while processing missing values because it may influence the researcher's choice of missing value processing techniques. For instance, the hot-deck imputation approach of missing values typically performs well but is challenging to use with big datasets. The greatest data sizes in the various types of sets utilized in each publication were counted and are displayed in Fig. 10. We only record the dataset with the largest data size when a study uses datasets in different sizes.

The real dataset has the largest data, with 22.58% of its total exceeding 10,000 according to Fig. 10. The label "Unknown" in the figure denotes the fact that papers have not stated the size of the dataset. Figs. 9 and 10 show that researchers employed one or two different types of datasets for study in order to assess the efficacy of deleting and imputing missing values. In general, different types of data have varied sizes. On the other hand, the size of the simulated or semi-simulated datasets employed by researchers is typically smaller than that of the real datasets. One of the causes is that when employing simulated or semi-simulated datasets for experiments, researchers establish various missing mechanisms, missing patterns, and missing rates. Large datasets would present significant space and computing challenges if they were employed in studies. The data sizes of two studies [94,96] about non-response in social networks are also relatively small, and the approaches taken to handle missing values differ from those used in other studies.

5.3. Main challenges of missing value analysis

5.3.1. Missing mechanism

One of the primary issues with missing value analysis is the missing mechanism, which may assist researchers in identifying the causes of missing values and the appropriate approaches for processing them. Since the missing mechanism for simulation and semi-simulation data types in this study is artificial, statistics for these two types were conducted concurrently. The missing mechanism from the real dataset is more complex and deserves its own discussion. When the missing value was simulated, "unknown" meant that the researcher had not explicitly stated the cause of the missing value.

From Fig. 11, it is obvious that researchers frequently take into consideration the three missing mechanisms simultaneously when employing both simulated and semi-simulated data, with the semi-simulation accounting for 43.75% and the simulation accounting for 25%, respectively. In addition, experiments in the semi-simulated research frequently use the MCAR and MAR combination (29.17%) Because they attempted to simulate the missing scenario of the real dataset, some researchers took into consideration the single missing mechanism, such as not MCAR, MAR. Furthermore, it might be challenging to identify the causes of missing values in real datasets due to their complexity. Fortunately, the majority of the studies included in the discussion investigated the missing mechanism of real datasets, with only a small number of studies failing to report it. In particular, we reviewed and categorized the missing mechanisms provided in the study, and then, as shown in Table 8, we categorize the missing mechanisms of real data into 7 groups.

From Table 8, the majority of researchers assumed that 54.84% of their real data was missing at random. Although the datasets we studied concentrated on handling missing values, approximately 16.13% of the papers did not examine or report missing

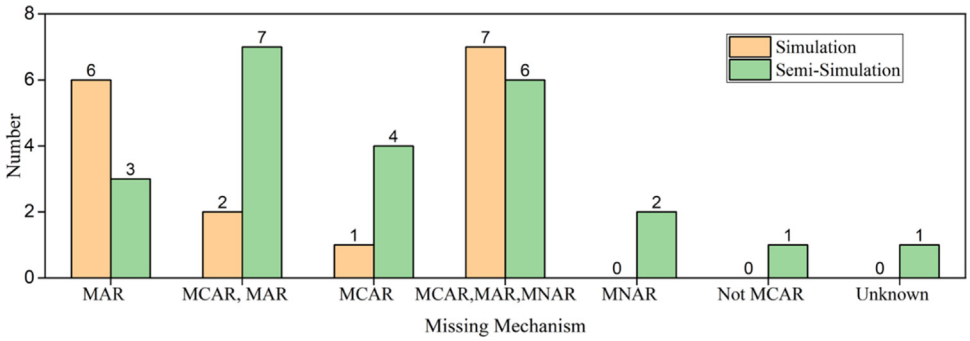


Fig. 11. The number of missing mechanism combinations used in experiments with simulated and semi-simulated data types.

Table 8
Missing mechanism of real datasets in studied papers.

Missing mechanism	Description	Literature list
MCAR, MAR, NMAR	Missing values are assumed to be multiple missing mechanisms for analysis	[88]
Not MCAR	Missing values are assumed to not be missing completely at random through analysis	[53], [56], [60], [101], [105]
MAR	Missing values are assumed to be missing at random through analysis	[32], [35], [34], [57], [61], [42], [66], [67], [68], [71], [74], [76], [77], [80], [92], [103], [73]
MAR, MNAR	Missing values are assumed to be missing at random through analysis and missing not at random	[33]
MCAR or MAR	Missing values are assumed to be missing completely at random or missing at random through analysis	[84]
Part MAR	A part of missing values is assumed to be missing at random or missing at random through analysis	[62]
Unknown	Missing mechanism is not reported or cannot be confirmed after analysis in the paper	[58], [69], [72], [79], [36]

processes. Researchers are encouraged to analyze the missing mechanism of missing values in datasets even though the current technology makes it challenging to report missing data accurately because the missing mechanism can have an impact on the results of the researcher's report [86,89].

5.3.2. Missing pattern

As we introduced before, we adopt the types of missing patterns proposed by Emmanuel Tlamelo [2]. We conduct statistical analysis on the missing patterns of datasets with various data types to investigate the missing patterns of data in different types of datasets, as illustrated in Fig. 12.

Fig. 12 reveals that the majority (74.19%) of the datasets of real data types consist of complex missing patterns with numerous missing values. In-depth descriptions of missing patterns were published by researchers in the literature [56,58,61,68,71,92]. Researchers should conduct a thorough analysis of the many missing value patterns since in real life, they show diversity. In addition, different missing patterns for the same missing rate produce varied relevant information in the data. A dataset with 10 rows and 10 columns and a 10% missing rate serves as an extreme example. If there is only one missing value in each row of data, the missing pattern is univariate, and there is no complete instance. In this case, the missing value processing method complete case analysis cannot be used.

5.3.3. Missing rate

We counted the three missing rates in three types of datasets in order to analyze the report situation. We specifically counted the row, column, and total missing rates in datasets when research introduced them. Additionally, the biggest missing rate in the publication is utilized for statistics in order to analyze the missing rate employed in the research papers, as shown in Fig. 13.

The category of missing rate that concerns those researchers, and the amount of the missing can be investigated in Fig. 13. Meanwhile, a paper may provide many types of missing rate information. Additionally, the percentage of column, row, and total missing rates for real datasets is all under 80%. On the other hand, the total missing rates in the real datasets are all under 30%, indicating that the real data available in the real world cannot possibly miss too much information. Moreover, researchers are more likely to report missing values for columns and rows in real datasets (93.75%) than in simulation datasets (80%), according to a comparison of the number of real-type datasets. Researchers estimate that the total missing rate is higher (35%) only in the semi-simulation datasets.

5.4. Analysis of experimental performance

5.4.1. Comparison matrix

Following the previous definition, there are primarily two types of comparison indicators in the comparative investigation

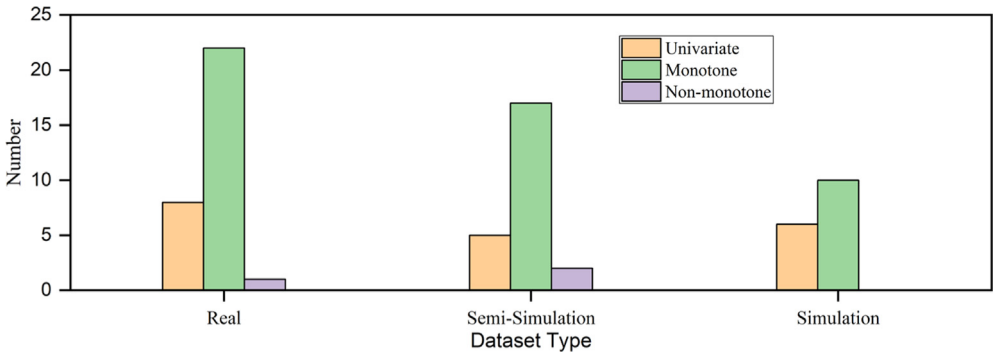


Fig. 12. The missing pattern in different dataset types.

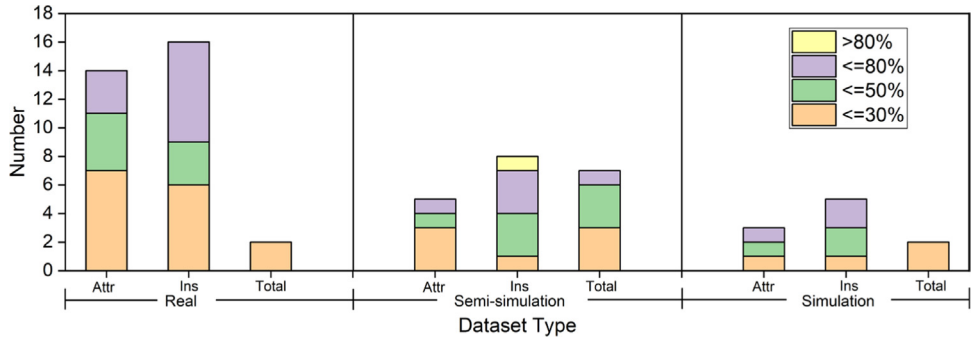


Fig. 13. The missing rate distribution in different dataset types. We analyze the column missing rate (Column), row missing rate (Row), and total missing rate (Total) of various data types in the studied articles in the figure.

Table 9
Two types of indicators sin studied papers.

Category	Name	Equation	Reference
Statistics	Mean Squared Errors (MSE)	$\frac{1}{n}\sum_{i=1}^{n}(x_i-\widehat{x_i})^2$	[55], [57], [68], [73], [77], [79], [90], [94], [102]
	Root Mean Square Error (RMSE)	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\widehat{x_i})^2}$	[54], [59], [44], [65], [69], [98]
Learning	Area Under Curve (AUC)	The area enclosed by the coordinate axis under the ROC (Receiver Operating Characteristic) curve.	[34] [56] [68]
	The percentage of correct predictions (PCP)	$\frac{Number\ \ of\ \ correct\ \ predictions}{Total\ \ number\ \ of\ \ predictions}$	[72,99]

of various missing value processing methods. We performed a statistical analysis of the two types of indicators in accordance with Table 9 in order to investigate the frequency of use of the two types of indicators. Only 15.87% of the research employed the learning indicators; 92.06% of the studies used statistical indicators; and four studies took into account both statistical and learned indicators. In addition, we introduce the two types of indicators that are utilized the most frequently in the studied paper.

Where x_i is the original value, $\widehat{x_i}$ is the imputed value and n is the total number of missing values. Lower MSE and RMSE values indicate better estimates of missing values. Conversely, higher AUC and PCP represent better prediction performance. Statistical indicators are primarily used to compare the difference between two values, as can be seen from the definition of indicators. The learning metric is to concentrate on data processing, then learn from the data to create some sophisticated models. In addition to these four variables, statistical variables like mean [33,59] and deviation [29,58] are also frequently utilized. They can be used to compare two datasets that have been processed by various missing value processing methods, such as an experimental group and control group in the medical field [32], or a dataset based on various hypotheses on missing mechanisms [85,92].

5.4.2. Processing method

Furthermore, a variety of imputation techniques are utilized in research publications to process missing values. We performed statistics on the deletion and imputation techniques used in the studies under consideration, and Fig. 14 shows those techniques that were applied more than three times. We utilize the colors green and orange to distinguish between deletion and imputation approaches, respectively.

We observe that the CCA approach was applied in each paper. The CCA and MI approach was employed by the majority of researchers (74.6%) for comparison. Three articles [32,75,86] compare data using both the CCA and ACA deletion procedures. Medical investigations, such as clinical trials, typically employed

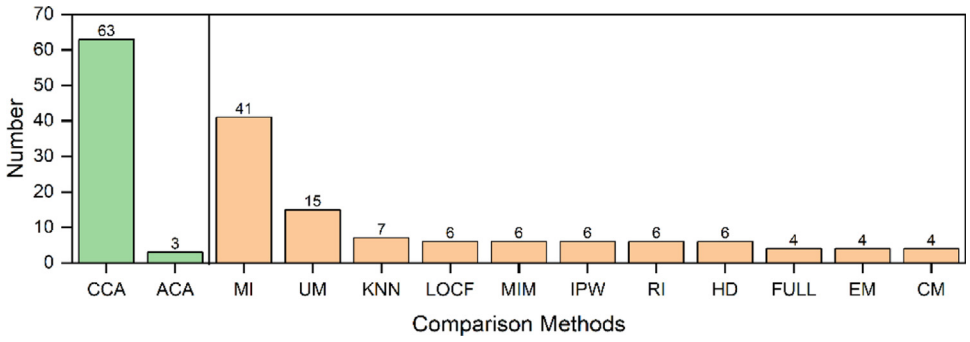


Fig. 14. The comparison techniques were employed in the studied papers more than three times. The abbreviations of models are Unconditional Mean (UM), Last Observation Carried Forward (LOCF), Missing Indicator Method (MIM), Inverse Probability Weighting (IPW), Regression Imputation (RI), Hot-deck (HD), The standard logistic regression with full data (FULL), Expectation–Maximization imputation (EM), Conditional Mean (CM).

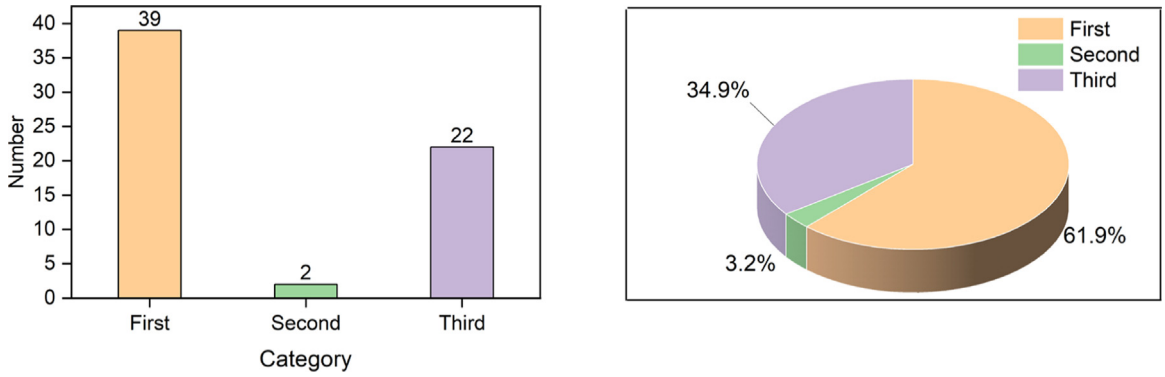


Fig. 15. The number of studied papers for each category.

LOCF (Last Observation Carried Forward), WCI (Worst Case Imputation), BCS (Best-Case Scenario), and BOCF (Baseline Observation Carried Forward) [82]. Only the three model-based imputation approaches illustrated in Fig. 14 are KNN, RI, and EM. In the publications analyzed, multiple imputation techniques are in fact more widespread than model-based techniques for imputation, and some researchers have used multi-class imputation techniques simultaneously [63,64,68]. It is important to keep in mind that, as we introduced earlier, the multiple imputation (MI) approach, which has a high model complexity, requires more time to estimate *m* full datasets.

5.4.3. Experiment results and analysis

Finally, we separated the studied articles into three groups based on the results of the experiments. As shown in Fig. 15, we statistically determined the number of publications that were studied for each category in accordance with Table 9. For ease of display, we employ the sequence to depict each category in turn.

According to the pie chart, the majority of the studies showed that the experimental performance of imputation methods is superior to that of deletion methods (61.9%), with some studies using traditional imputation techniques like MI and KNN and the remainder using more advanced techniques. Additionally, about 34.92% of researchers noticed differences in performance between the deletion and imputation strategies. The performance of these methods is influenced by a number of factors, including data size, missing mechanism, missing rate, and missing pattern. We observe that imputation technique research receives greater attention than deletion method research. The experimental findings of the publications that were considered also suggest that the imputation method is a superior approach to handling missing variables because in most cases, it can perform better than deletion.

Specific of articles in this work that fall under category 3 "Discussion" demonstrate deletion strategies also exhibit better results in some situations. we analyzed those papers and provided them in Table 10 along with some helpful information concerning deletion techniques.

Table 10 provides some helpful guidelines that can be used to determine the most suitable approach to handle missing information, (1) deletion of missing values can decrease statistical power; (2) when the missing rate is less than 5% or the missing mechanism is MCAR, unbiased analysis results can be obtained by using the deletion method; (3) if more than 10% of observations had missing values, deletion of missing values introduces bias, regardless of the missing data mechanism.

6. Experiment and analysis on processing missing values

In order to evaluate and compare the effectiveness of the traditional missing value processing methods, we conducted comprehensive experiments and analysis using the CCA deletion method and Mean&Mode imputation, KNN imputation, Multiple imputation (MI), and MissForest imputation [106] methods. In particular, we examined 9 complete and open-source data sets with various sizes and manually introduced 8 specific missing ratios of overall missing values, including 1%, 3%, 5%, 10%, 20%, 30%, 40%, and 50%. We then simulated four missing patterns according to the third category of missing patterns [20–25], including simple, middle, complex, and blend patterns. Therefore, for each dataset, 32 combinations are generated in the experiments. In order to reduce the effect of contingency, we created 10 incomplete datasets for each combination. Since the KNN imputation approach is sensitive to feature scale, we normalize all datasets first. Specifically, to assess and compare different methods for processing missing values, we conducted two types of experiments, including imputation error

Table 10
The useful information of deletion methods.

Information description	Literature list
Statistical power is reduced when using deletion methods of missing values	[63,65,71,105]
When the missing rate is less than 5%, the deletion will show better performance	[31], [63], [104]
The deletion method is only used in the MCAR mechanism	[61], [65], [82], [91]
When missing values more than 10%, deletion of missing values introduces bias.	[27], [58]

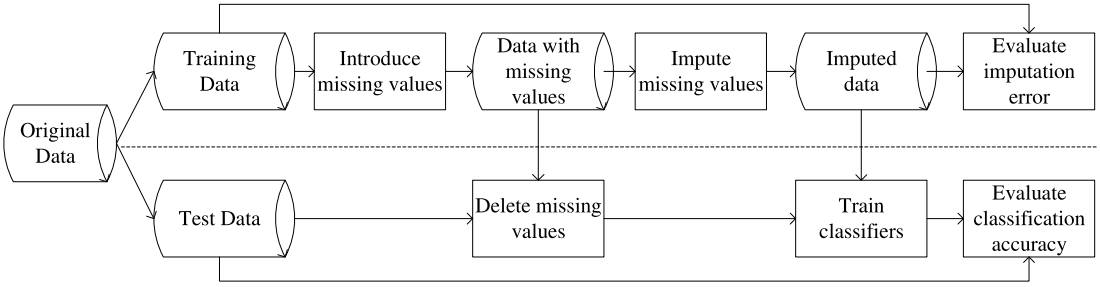


Fig. 16. The flow of two types of experiments.

Table 11
The information of experiment datasets.

Dataset	Total samples	Total variables	Categorical	Numerical
BMI	152	6	2	4
Heart Failure	299	13	6	7
GBSG2	686	10	4	6
Yeast	1484	9	1	8
CMC	1,473	10	8	2
Obesity	2,110	17	9	8
Cardiovascular	10,000	12	7	5
Nursery	12,960	8	8	0
NHANES	24,434	8	8	0

and classification performance. Imputation error is the difference between the imputed values by the missing value imputation methods and the actual values in the original data. In addition, we also pay attention to the effects of various processing techniques on baseline tasks (such as classification tasks) in practice. For instance, when performing classification tasks, we handle missing values in the data, use processed data to train classifiers, and then assess the effectiveness of the classifiers. Fig. 16 illustrates the experiment's flow.

6.1. Experimental datasets

In order to evaluate the impact of missing values on datasets with different sizes, we utilized 9 public datasets of various sizes, and the information of the data is displayed in Table 11. Specifically, three data sets with less than 1000 instances, among which German Breast Cancer Study Group 2 (GBSG2) and Body Mass Index (BMI) can be found in the R package, and Heart Failure comes from the UCI machine learning repository. At the same time, we used three datasets with more than 1000 but less than 5000 instances, all from the UCI machine learning repository. Further, we included three data sets with more than 5000 instances, among which the Nursery data set is also from the UCI machine learning repository, the Cardiovascular data set is obtained from the Kaggle platform (we randomly sampled 10,000 records from the original data for experiments), NHANES From the study by Fernando López-Martínez et al. [107].

6.2. Performance measures

Since our data contains both numerical and categorical attributes, we use the falsely imputed categories (PFC) and the mean of squared imputation errors (MSIE) in the imputation error analysis to assess the performance of different imputation methods on categorical and numerical attributes. The smaller values of PFC and MSIE mean the better the performance of the method. PFC can be computed by

$$PFC = \frac{1}{N}\sum_{i}^{n}\sum_{j}^{m} I(x_{ij} \neq x'_{ij}) \tag{6}$$

where $I(\cdot)$ is an indicator function, which is 1 when the predicted value and the true value are the same. In addition, we employed the mean of squared imputation errors ($MSIE$) for the numerical attributes and it can be calculated by

$$MSIE = \frac{1}{N}\sum_{i}^{n}\sum_{j}^{m}(x_{ij} - x'_{ij})^2 \tag{7}$$

where N is the number of numerical missing values, x_{ij} is the true value in the complete data matrix, and x'_{ij} is the corresponding imputed value. Furthermore, we use classification accuracy as a performance metric to assess how different missing value methods affect classification tasks and it can be computed by

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

where TP is true positive, TN is true negative, FP is false positive and FN is false negative.

6.3. Missing patten analysis

As previously introduced, this experiment considered four different missing patterns, which means the distribution of missing values in the data set is different. Therefore, in order to observe the distribution of missing values in each missing pattern, we took a 1% missing rate as an example to show the distribution of four patterns of the BMI dataset, as shown in Fig. 17.

Fig. 17 exhibits observed values as black and missing values as white. We observed Figs. 3 and 12 show a similar distribution of missing values, proving that Table 3 is satisfied by our method for generating missing values.

6.4. Experiment and analysis on imputation error

In the imputation error experiment, our goal is to assess the imputation effectiveness of traditional imputation methods for

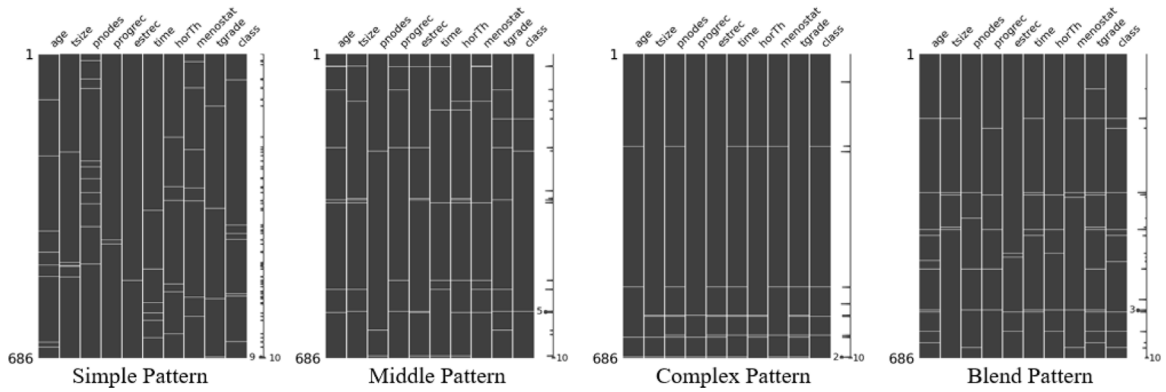


Fig. 17. Examples of missing patterns for the BMI dataset at a missing rate of 0.1.

processing missing values. Specifically, all methods are implemented using Python 3.8 and use the default parameters. All methods are implemented using Python 3.8. In the experiment, we first employed these four methods to fill in the missing values that we have artificially introduced and then compared the differences between actual values and imputed values. Fig. 18 displays the experiment's outcomes.

First of all, it is worth noting that all datasets cannot generate simple patterns under a 50% missing rate, because the total number of missing values is already greater than the number of rows in the data. Second, as seen in Fig. 18, the multiple imputation (MI) method consistently has the minimum imputation error when the dataset is small, followed by MissForest and Mean&Mode imputation methods, while the KNN imputation method performs the worst. The imputation performance of MissForest improves when the data set is larger than 1000, especially for categorical variables. In all datasets, the KNN imputation method performs poorly, whereas the Mean&Mode imputation method performs somewhat better. Since the MI method accounts for the uncertainty of missing values, it generally outperforms other numerical variables. In addition, MissForest always performs best on categorical variables in datasets larger than 1000, because when the data set is small, there is insufficient data to construct a classification or regression model to estimate missing values. At the same time, the MissForest imputation method performs poorly in numerical types, but it is also better than the KNN imputation method. Furthermore, we compared the average execution time of all combinations of the four methods, as shown in Fig. 19.

In Fig. 19, MissForest has the largest execution time, except in the NHANES dataset which is lower than the KNN imputation method. The time of the KNN imputation method is similar to Mean&Mode and MI imputation methods when the data set is less than 1000, but when the data set size expands, the time of KNN imputation grows rapidly. Therefore, the KNN imputation method is not suitable for large-scale datasets.

6.5. Experiment and analysis on classification task

Additionally, we used the traditional random forest (RF) model as the base classifier in order to investigate the effects of various methods for handling missing values on the baseline classification task. As shown in Fig. 16, we first divided the data into 70% training data and 30% test data, and then introduced a specific proportion of missing values in the training data set, and use different methods to process them. Finally, we input the data processed by different methods into the RF classifier, and then 30% of the complete data set is used for performance evaluation. In this experiment, we included Mean&Mode, multiple imputation (MI), KNN, MissForest imputation approaches, and the CCA deletion method for comparative analysis. The experimental results are shown in Fig. 20.

Fig. 20 indicates that, with the exception of the smallest BMI dataset, the performance of all missing value processing methods exhibits a negative trend as the missing ratio rises. Meanwhile, the classification performance of all approaches is comparable when the missing rate is less than 0.1, but when it is greater than 0.1, the results of the methods are diverse. In addition, the Mean&Mode method performs the worst, with the exception of the BMI data set, because all missing values are replaced by mean or mode values which can easily destroy the data's distribution. Further, when the missing ratio is larger than 0.1 and the data size is more than 1000, the MissForest and CCA methods perform well. However, since a substantial portion of the available data is discarded and only a tiny portion of the data can be utilized to train the classifier, CCA performs badly when the missing rate is more than 0.1 and the dataset is small. On the other hand, in different missing patterns, these approaches only exhibit a significant performance difference in the middle pattern. In addition, the MissForest and CCA methods perform slightly better in all other patterns. Although the MI method performs a lower imputation error, it performs slightly lower on the classification accuracy than the MissForest and CCA methods. According to the experimental results, CCA tends to be more effective when the dataset is large and when researchers simply evaluate the classification performance, but it frequently removes a sizable percentage of instances.

Finally, in order to offer readers more information and guidance to handle missing values, we manually completed a decision tree for missing value processing according to our experimental results, as shown in Fig. 21.

As shown in Fig. 21, the missing rate is the main factor when processing missing values. When the missing rate is less than 0.1 and researchers need to retain all the data, it is recommended to use the MissForest method to impute missing values for data sets with more categorical variables, and the MI method can be used when the data includes many numerical variables. On the other hand, the CCA approach is more appropriate when the missing rate is less than 0.1 and the researchers intend on using a simple process because it performs well in this case. However, when the missing rate is more than 0.1, various processing methods produce varying consequences, therefore we have to take the size of the data set into account. When the dataset is small (e.g. less than 1000), it does not recommend the CCA method because it slightly reduces the classification accuracy. In addition, when the research uses a medium-sized data set (e.g. from 1000 to 5,000), it recommends the MissForest method. Moreover, when using a large data set, when the researchers need to retain all missing

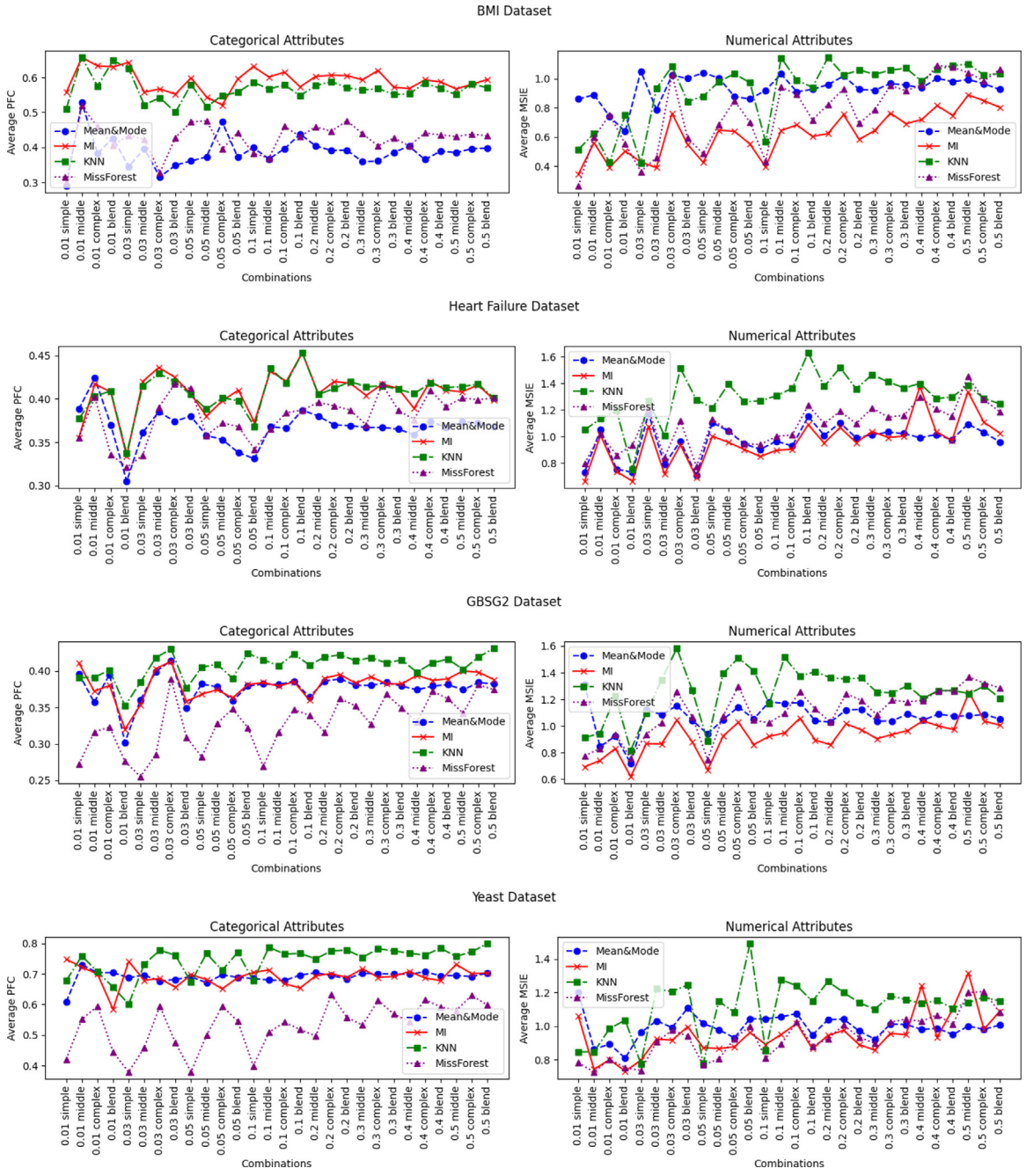


Fig. 18. Average of performance measure for all 32 combinations.

values, the performance of the MissForest method is relatively high. Additionally, the CCA approach performs better for larger data sets when the missing ratio is higher (such as 0.5), but more data rows would be lost. Therefore, we advise utilizing a mixed processing approach when the missing rate is greater than 0.5 since the performance of the imputation method is consistently worse than that of the CCA method in the middle missing pattern. Specifically, by analyzing the distribution of missing values in the data, rows or columns with large missing values might be eliminated first and then processed using the missing value imputation method. Although the Mean&Mode approach performs well in some data sets in imputation error experiments, we do not suggest it because of its tendency to skew the data's distribution. Obviously, our decision tree has a limitation due to we only considered 9 datasets and 5 missing values processing methods, but it can offer the researchers fast instructions.

7. Discussion

We undertake in-depth research and analysis on the main challenges of missing value analysis, missing value processing approaches, and a comparison of common processing methods based on the previous research. Next, we need to discuss two crucial topics: (1) what are the current research hotspots for missing values, and (2) what open challenges in missing value analysis still need to be addressed.

7.1. Current research hotspots

7.1.1. Processing methods selection for missing value

Missing values are receiving more and more attention, and people are becoming increasingly concerned about how to handle

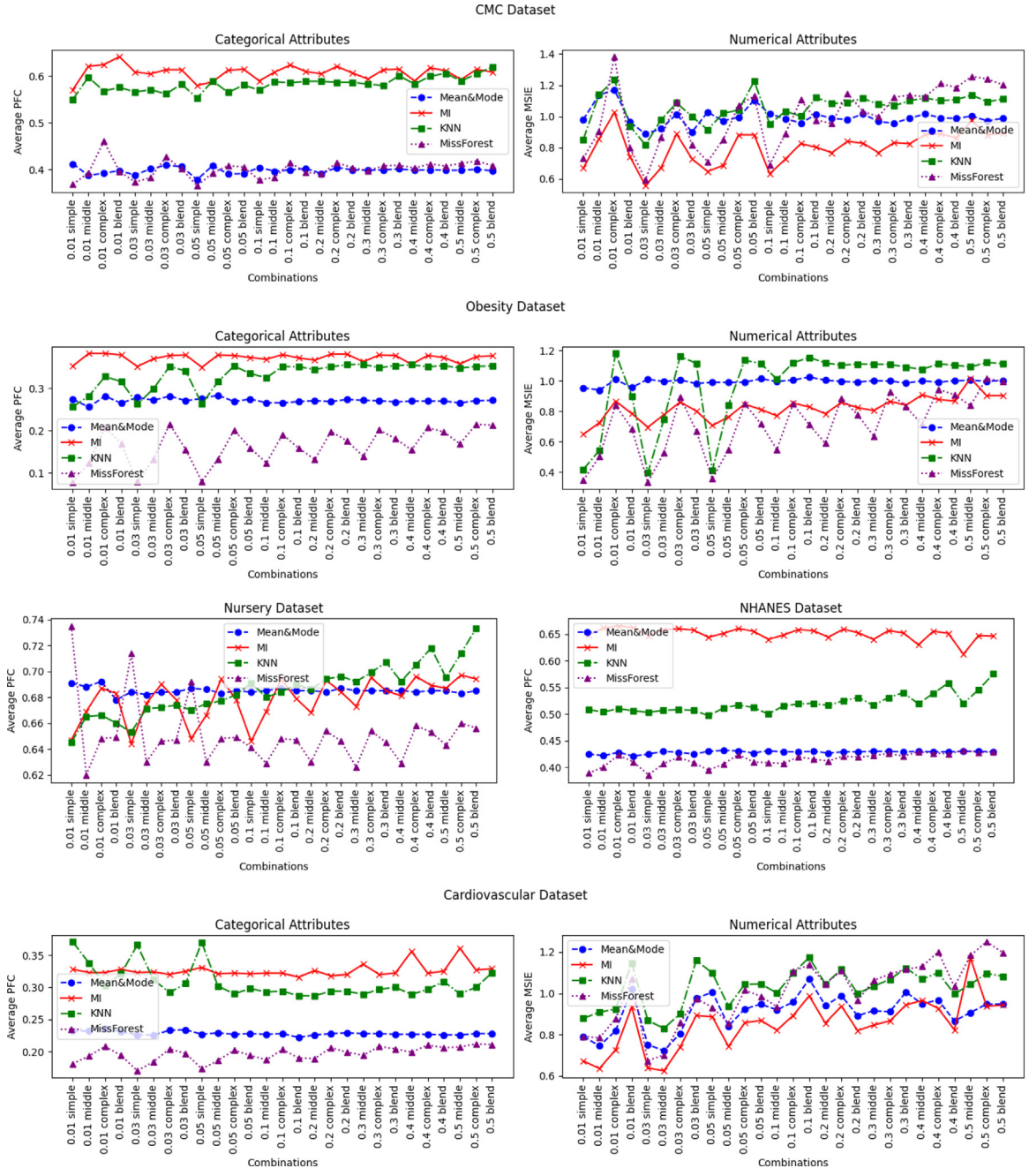


Fig. 18. (*continued*).

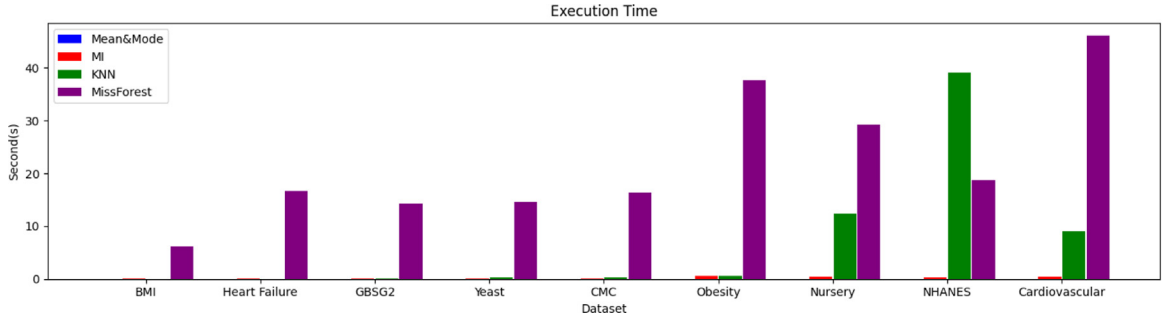


Fig. 19. Execution time of imputation methods in nine datasets.

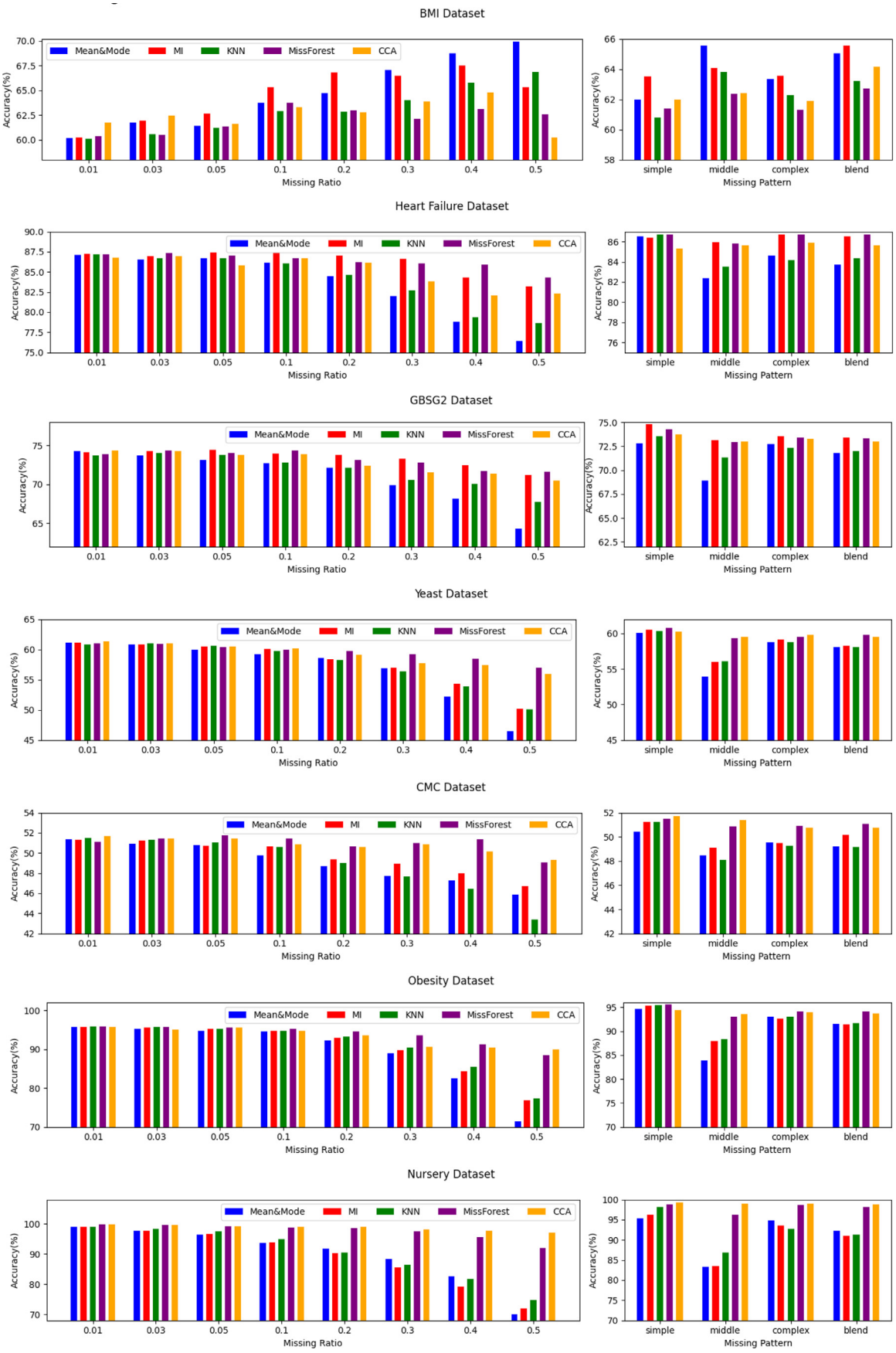


Fig. 20. Comparative analysis of classification performance of different processing methods.

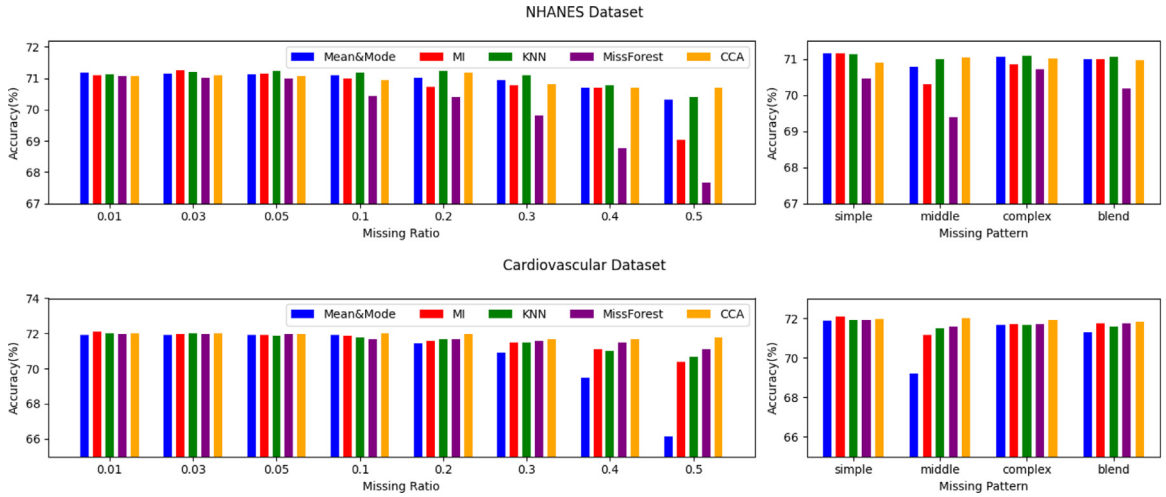


Fig. 20. (*continued*).

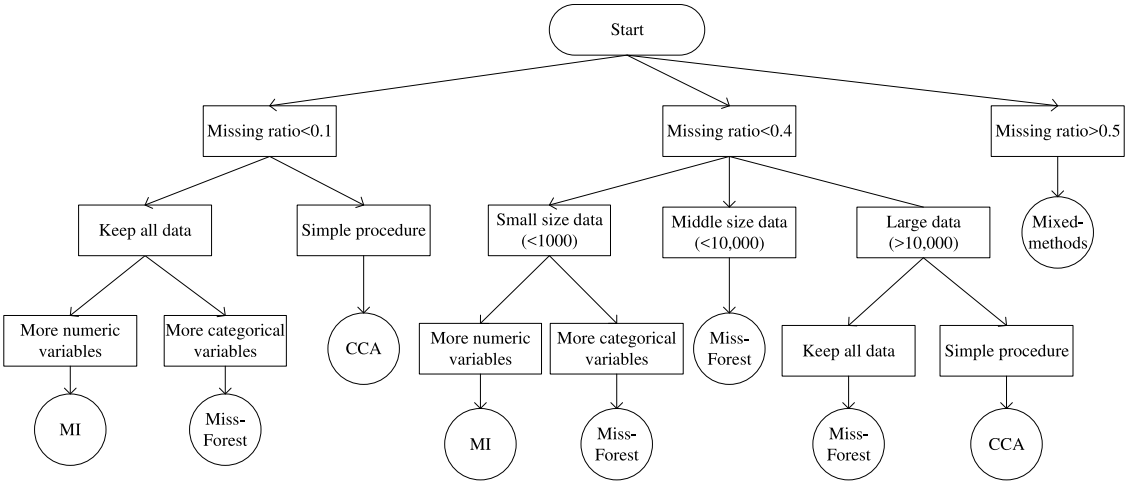


Fig. 21. Guided decision tree for missing value handling.

them effectively. Other than simply deleting them, there are numerous imputation techniques, therefore researchers must make a reasonable decision. Researchers should therefore concentrate on analyzing three main characteristics of missing values when they come across them: the missing mechanism, the missing pattern, and the three missing rates. Techniques like chi-square can be utilized to help with the study of the missing mechanism. Additionally, the results of multiple techniques can be compared by assuming different missing mechanisms when the existence of the missing mechanism cannot be confirmed [83,85]. For instance, a medical study [62] proposed a method that incorporates a complete case analysis (CCA) in the beginning, followed by sensitivity studies that generate progressively more exact predictions about the outcomes of patients with missing data.

7.1.2. Comparison of missing mechanisms

From Table 9, some researchers prefer to compare the effects of missing values according to different mechanisms [32,59,69,78,91]. For example, one paper sets different scenarios for the MAR mechanism [65]. Additionally, only 9.52% of the studied papers did not analyze the missing mechanism [36,58,69,72,79,104]. And a large number of studies have simulated data with different missing mechanisms for experimental analysis [32,33,53], which indicates that the missing mechanism is one of the hotspots of current research. Additionally, for real datasets with missing values, sensitivity analysis [62,85,88,92,97] is also used to make various assumptions about missing mechanisms in the study, which can more comprehensively and scientifically assess the impact of missing values on the research's conclusions.

7.1.3. Missing value imputation

From the papers included in the study, 22.22% of the articles included in the analysis proposed sophisticated missing value imputation techniques. For instance, Cheng [73] investigates an inverse probability weighting arrangement using an IS imputation approach and its modified algorithm for quantile regression with missing covariates. On the other hand, Fig. 15 shows that in 61.9% of the experiments, the missing value imputation strategy performs better than deletion. Meanwhile, Fig. 14 shows that the majority of researchers prefer the multiple imputation method. Obviously, the selection of a missing value imputation method is a hot research topic because there are numerous strategies for achieving this goal. Mean imputation was discovered to be a terrible strategy for imputation of missing values in some of the studies we included [33,44] and therefore to be avoided in investigations. Since of the outstanding performance of missing value imputation methods, we believe that the study of missing value imputation will remain a significant topic in the field of missing value research in the future.

7.2. Open challenges

7.2.1. Analysis of missing patterns

The missing mechanism and the missing rate have received a lot of scientific attention lately, according to the articles we included in the study. For instance, even though there is still a lack of effective missing value processing techniques for the NMAR missing mechanism [17], many researchers have employed simulation approaches to simulate this missing mechanism in order to conduct experimental analysis. For the majority of investigations, their missing rate is also reported, and some studies have come to the experimental conclusion that deletion approaches are also applicable at 5% missing rates [33,63,104]. The investigation and analysis of the missing patterns, however, lack sufficient detail. The research we cited indicated that the study of missing patterns was restricted to three categories – univariate, monotone, and non-monotone – and that a more comprehensive analysis of the complex non-monotonic situation was not carried out. Analyzing the missing pattern in practice is a complex process, which needs to be analyzed together with the background of the study, the missing mechanism and the missing rate. To evaluate the missing mechanism and missing pattern of the data, researchers might attempt using several visualization tools, such as the "missingno" toolkit in Python.

7.2.2. Complexity of imputation methods

The imputation method has been favored by most researchers and exhibits good performance, similar to our earlier analysis. However, the difficulty of various imputation techniques has not been sufficiently analyzed in the current research. Only one study [73] evaluated the time complexity of the imputation approach, even though 74.6% of the studied papers used the multiple imputation method. In general, the complexity of the imputation approach can be ignored when dealing with little amounts of research data, but it must be taken into account when dealing with large amounts of data. Therefore, in the future, the complexity of imputation methods is also one of the open challenges of missing value research. Researchers must maintain a balance between the performance that imputation techniques bring and the resulting model complexity.

7.2.3. Missing value analysis on data mining

Increasingly data are being collected and processed as information systems keep growing. According to the papers we reviewed, people generally give less attention to how different missing value processing methods affect data mining tasks and instead concentrate more on the statistical distinctions between various approaches in comparative studies of missing value processing methods. Specifically, compared to statistical analysis, the performance of the construction model, such as prediction accuracy, is given more consideration in data mining tasks. There currently are few researches on the effects of various missing value processing techniques on data mining tasks, and missing values make it challenging to apply standard data mining techniques. Therefore, future research on missing values in data mining will be another open challenge.

7.2.4. Mixed missing value handling method

At present, the imputation of missing values has received a lot of attention in recent studies, but few studies have focused on how to combine the imputation and deletion methods of missing values. In this study, we discovered that the missing value deletion approach can perform well on benchmark tasks and benefits from simplicity. However, the deletion method loses a large amount of useful data, so how to balance the data loss and performance improvement caused by deleting missing values in complex missing patterns is also worth studying in the future.

8. Conclusion

Missing values are a challenge that arises frequently during data analysis. The presence of missing values can easily influence the results of data analysis or data mining. There are numerous approaches to handling missing values, with deletion and imputation being the two most widely used approaches. The dataset can be rapidly and easily completed by deleting the missing values, and imputation completes the data by imputing the most reasonable values in place of the missing values. By reviewing studies that compare deletion and imputation, this study intends to enhance understanding of the main challenges of missing value analysis and common missing value processing approaches and their progression.

In order to analyze missing values in the data, we first analyzed the three major difficulties with missing value analysis, including missing mechanism, missing pattern, and missing rate. Three missing mechanisms were thoroughly introduced, and three examples were created to assist in understanding. Furthermore, we are the first to provide a thorough analysis of the classification of missing patterns, and we also include example figures to help with comprehension. We then separated the missing rates into three groups based on the row, column, and total dimensions of the data. Meanwhile, we offered them three equations to show their calculation methods while introducing them.

Second, we investigated the procedures for handling missing data and provided a thorough introduction to the categorization and characteristics of deletion and imputation missing approaches. There are several techniques for handling missing values, including deletion, imputation, and weight approach, where deletion and imputation are two typical techniques. Deletion methods reduce the amount of data and lose some information. In contrast, imputation approaches replace missing values with plausible values in order to retain more data, however, the values used in these methods are not real and may be cheated. Now, a variety of missing value imputation techniques have been proposed, each with unique benefits and drawbacks, thus choosing a suitable approach to handle missing values is so challenging.

We investigated 63 papers comparing deletion and imputation, and we analyzed experimental datasets, the main challenges of missing value analysis, and experimental performance to evaluate the performance differences between deletion and imputation of missing values. In order to highlight the key distinctions between these two approaches, we thoroughly reviewed and analyzed these studies. On the basis of the experiments' findings and conclusions, we categorized these publications into three groups and present a detailed comparison of deletion and imputation strategies. In this work, we notice that missing value imputing appears to be increasingly common. The main cause can be that, despite the fact that deleting missing values is quick and easy, some useful information is ignored. However, in some situations, deletion techniques might provide faster performance while handling missing values with simplicity. We provide a helpful information table for deletion methods as a result.

We then conducted comprehensive experiments on comparing different methods for handling missing values in terms of imputation error and the effect of classification tasks. Specifically, we included 9 public datasets and 4 missing value imputation methods, and 1 missing value deletion method. According to the experimental results, we provided a simple guided decision tree to help researchers deal with missing values.

Finally, we provide potential study topics for the future by summarizing the current research hotspots and open challenges for missing values. The best strategy to reduce bias is generally for researchers to purposefully avoid missing values in their studies.

But when missing values inevitably occur, we should analyze the missing mechanism, missing pattern, and missing rate of missing values before deciding on the most appropriate action to take to decrease the bias brought on by missing values. We address drawbacks in previous review papers and offer some practical guidelines for processing methods of missing value in this work. In addition, we demonstrate a variety of techniques for assessing missing data before dealing with them, such as visualization techniques that are beneficial for studies that come across missing values.

Declaration of competing interest

All authors declare that they have no conflicts of interest.

Data availability

No data was used for the research described in the article.

Acknowledgments

This research is supported by the Sichuan Science and Technology Program of China (No. 2021YFH0107 and 2022NSFSC0571), and the Science and Technology Innovation Capability Improvement Program of Chengdu University of Information Technology, China (No. KYQN202223). All the authors are thankful to their respective universities for their support: Chengdu University of Information Technology, University of Lyon and Qatar University.

References

[1] Roderick J.A. Little, Donald B. Rubin, Statistical Analysis with Missing Data, in: Wiley Series in Probability and Statistics, 2014.
[2] Tlamelo Emmanuel, A Survey on Missing Data in Machine Learning, Research Square, 2021.
[3] Rima Houari, Ahcène Bounceur, A. Kamel Tari, M. Tahar Kecha, Handling missing data problems with sampling methods, in: Proceedings - 2014 International Conference on Advanced Networking Distributed Systems and Applications, INDS 2014, 2014, pp. 99–104.
[4] Bhavisha Suthar, Hemant Patel, Ankur Goswami, A survey: Classification of imputation methods in data mining, Int. J. Emerg. Technol. Adv. Eng. 2 (1) (2012) 309–312.
[5] Deepak Adhikari, Wei Jiang, Jinyu Zhan, Imputation using information fusion technique for sensor generated incomplete data with high missing gap, Microprocess. Microsyst. (2021) 103636.
[6] Shinichi Nakagawa, Robert P. Freckleton, Missing inaction: the dangers of ignoring missing data, Trends Ecol. Evol. 23 (11) (2008) 592–596.
[7] Judith Godin, Janice Keefe, Melissa K. Andrew, Handling missing mini-mental state examination (MMSE) values: Results from a cross-sectional long-term-care study, J. Epidemiol. 27 (4) (2017) 163–171.
[8] Sandip Sinharay, Hal S. Stern, Daniel Russell, The use of multiple imputation for the analysis of missing data, Psychol. Methods 6 (3) (2001) 317–329.
[9] Md Geaur Rahman, Md Zahidul Islam, Missing value imputation using a fuzzy clustering-based EM approach, Knowl. Inf. Syst. 46 (2) (2016) 389–422.
[10] Patrick Royston, Ian R. White, Journal of statistical software multiple imputation by chained equations (MICE): Implementation in stata, J. Stat. Softw. 45 (4) (2011) 1–20.
[11] Budhaditya Saha, Sunil Gupta, Dinh Phung, Svetha Venkatesh, Effective sparse imputation of patient conditions in electronic medical records for emergency risk predictions, 2017.
[12] Alireza Farhangfar, Lukasz Kurgan, Jennifer Dy, Impact of imputation of missing values on classification error for discrete data, Pattern Recognit. 41 (12) (2008) 3692–3705.
[13] Pedro J. García Laencina, José Luis Sancho-Gómez, Aníbal R. Figueiras-Vidal, Pattern classification with missing data: A review, Neural Comput. Appl. 19 (2) (2010) 263–282.
[14] Anil Jadhav, Dhanya Pramod, Krishnan Ramanathan, Comparison of performance of data imputation methods for numeric dataset, Appl. Artif. Intell. 33 (10) (2019) 913–933.
[15] Wei Chao Lin, Chih Fong Tsai, Missing value imputation: a review and analysis of the literature (2006–2017), Artif. Intell. Rev. 53 (2) (2020) 1487–1509.
[16] Ronald A Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugen. 7 (2) (1936) 179–188.
[17] Schafer Joseph L., Graham John W., Missing data: Our view of the state of the art, Psychol. Methods 7 (2) (2002) 147–177.
[18] Sherene E. Sharath, Nader Zamani, Panos Kougias, Soeun Kim, Missing data in surgical datasets: a review of pertinent issues and solutions, J. Surg. Res. 232 (2018) 240–246.
[19] X. Karl Pearson, On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, Lond. Edinb. Dublin Philos. Mag. J. Sci. 50 (302) (1900) 157–175.
[20] Heikki Junninen, Harri Niska, Kari Tuppurainen, Juhani Ruuskanen, Mikko Kolehmainen, Methods for imputation of missing values in air quality datasets, Atmos. Environ. 38 (18) (2004) 2895–2907.
[21] Sanaz Nikfalazar, Chung Hsing Yeh, Susan Bedingfield, Hadi A. Khorshidi, Missing data imputation using decision trees and fuzzy clustering with iterative learning, Knowl. Inf. Syst. 62 (6) (2020) 2419–2437.
[22] Md Geaur Rahman, Md Zahidul Islam, Terry Bossomaier, Junbin Gao, CAIRAD: A co-appearance based analysis for incorrect records and attribute-values detection, in: Proceedings of the International Joint Conference on Neural Networks, 2012.
[23] Md Geaur Rahman, Md Zahidul Islam, Missing value imputation using decision trees and decision forests by splitting and merging records: Two novel techniques, Knowl.-Based Syst. 53 (51–65) (2013).
[24] Md Geaur Rahman, Md Zahidul Islam, KDMI: A Novel Method for Missing Values Imputation using Two Levels of Horizontal Partitioning in a Dataset, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 8347 LNAI (PART 2), 2013, pp. 250–263.
[25] Md Geaur Rahman, Md Zahidul Islam, Data quality improvement by imputation of missing values, in: International Conference on Computer Science and Information Technology, 2013, pp. 82–88, Authors Suppressed Due to Excessive Length.
[26] Qinbao Song, Martin Shepperd, Xiangru Chen, Jun Liu, Can KNN imputation improve the performance of c4.5 with small software project datasets? a comparative evaluation, J. Syst. Softw. 81 (12) (2008) 2361–2370, Best papers from the 2007 Australian Software Engineering Conference (ASWEC 2007), Melbourne, Australia, April (2007) 10-13..
[27] Diane L. Langkamp, Amy Lehman, Stanley Lemeshow, Techniques for handling missing data in secondary analyses of large surveys, Acad. Pediatr. 10 (3) (2010) 205–210.
[28] Kiki Aristiawati, Titin Siswantining, Devvi Sarwinda, Saskya Mary Soemartojo, Missing values imputation based on fuzzy C-means algorithm for classification of chronic obstructive pulmonary disease (COPD), in: AIP Conference Proceedings, 2192(December), 2019.
[29] Karel G.M. Moons, Rogier A.R.T. Donders, Theo Stijnen, Frank E. Harrell Jr., Using the outcome for imputation of missing predictor values was preferred, J. Clin. Epidemiol. 59 (10) (2006) 1092–1101.
[30] Alexander D. Stead, Phill Wheat, The case for the use of multiple imputation missing data methods in stochastic frontier analysis with illustration using english local highway data, European J. Oper. Res. 280 (1) (2020) 59–77.
[31] W.L. Junger, A. Ponce de Leon, Imputation of missing data in time series for air pollutants, Atmos. Environ. 102 (2015) 96–104.
[32] Andrea Gabrio, Rachael Hunter, Alexina J. Mason, Gianluca Baio, Joint longitudinal models for dealing with missing at random data in trial-based economic evaluations, Value Health 24 (5) (2021) 699–706.
[33] Danielle Sullivan, Rebecca Andridge, A hot deck imputation procedure for multiply imputing nonignorable missing data: The proxy pattern-mixture hot deck, Comput. Statist. Data Anal. 82 (2015) 173–185.
[34] Md Nazmul Karim, Christopher M. Reid, Lavinia Tran, Andrew Cochrane, Baki Billah, Missing value imputation improves mortality risk prediction following cardiac surgery: An investigation of an Australian patient cohort, Heart Lung Circ. 26 (3) (2017) 301–308.
[35] Fang Fang, Jun Shao, Iterated imputation estimation for generalized linear models with missing response and covariate values, Comput. Statist. Data Anal. 103 (2016) 111–123.
[36] E. Vaiciukynas, A. Verikas, A. Gelzinis, M. Bacauskiene, J. Minelga, M. Hallander, E. Padervinskis, V. Uloza, Fusing voice and query data for non-invasive detection of laryngeal disorders, Expert Syst. Appl. 42 (22) (2015) 8445–8453.
[37] Erastus Karanja, Jigish Zaveri, Ashraf Ahmed, How do mis researchers handle missing data in survey-based research: A content analysis approach, Int. J. Inf. Manage. 33 (5) (2013) 734–751.
[38] Jane Y. Nancy, Nehemiah H. Khanna, Kannan Arputharaj, Imputing missing values in unevenly spaced clinical time series data to build an effective temporal classification framework, Comput. Statist. Data Anal. 112 (2017) 63–79.

[39] Qiujun Lan, Xuqing Xu, Haojie Ma, Gang Li, Multivariable data imputation for the analysis of incomplete credit data, Expert Syst. Appl. 141 (2020) 112926.
[40] Madan Lal Yadav, Basav Roychoudhury, Handling missing values: A study of popular imputation packages in r, Knowl.-Based Syst. 160 (2018) 104–118.
[41] Tero Aittokallio, Dealing with missing values in large-scale studies: microarray data imputation and beyond, Brief. Bioinform. 11 (2) (2009) 253–264.
[42] Zuber D. Mulla, Byungtae Seo, Ramaswami Kalamegham, Bahij S. Nuwayhid, Multiple imputation for missing laboratory data: An example from infectious disease epidemiology, Ann. Epidemiol. 19 (12) (2009) 908–914.
[43] J.A. Delaney, R.L. McClelland, E. Brown, D.A. Bluemke, Multiple imputation for missing with cardiac magnetic resonance imaging data: results from the multi-ethnic study of atherosclerosis (mesa), Can. J. Cardiol. 25 (2009) 07.
[44] Iris Eekhout, Henrica C.W. de Vet, Jos W.R. Twisk, Jaap P.L. Brand, Michiel R. de Boer, Martijn W. Heymans, Missing data in a multi-item instrument were best handled by multiple imputation at the item score level, J. Clin. Epidemiol. 67 (3) (2014) 335–342.
[45] Young-Saver Dashiell, Gornbein Jeffrey, Starkman Sidney, Saver Jeffrey, Handling of missing outcome data in acute stroke trials: Advantages of multiple imputation using baseline and postbaseline variables, J. Stroke Cerebrovasc. Dis. 27 (2018) 10.
[46] Jason Van Hulse, Taghi M. Khoshgoftaar, A comprehensive empirical evaluation of missing value imputation in noisy software measurement data, J. Syst. Softw. 81 (5) (2008) 691–708, Software Process and Product Measurement.
[47] Trond Hellem Bø, Bjarte Dysvik, Inge Jonassen, Lsimpute: accurate estimation of missing values in microarray data with least squares methods, Nucleic Acids Res. 32 (3) (2004) e34.
[48] Tobias Rockel, Dieter William Joenssen, Udo Bankhofer, Decision trees for the imputation of categorical data, Kit Sci. Publ. 2 (1) (2017) 1–15.
[49] Xianping Du, Hongyi Xu, Feng Zhu, A data mining method for structure design with uncertainty in design variables, Comput. Struct. 244 (2021) 106457.
[50] Kancherla Jonah Nishanth, Vadlamani Ravi, Narravula Ankaiah, Indranil Bose, Soft computing based imputation and hybrid data and text mining: The case of predicting the severity of phishing alerts, Expert Syst. Appl. 39 (12) (2012) 10583–10589.
[51] Bahareh Fallah, Kelvin Tsun Wai Ng, Hoang Lan Vu, Farshid Torabi, Application of a multi- stage neural network approach for time-series landfill gas modeling with missing data imputation, Waste Manag. 116 (2020) 66–78.
[52] T. Vatanen, M. Osmala, T. Raiko, K. Lagus, M. Sysi-Aho, M. Orešič, T. Honkela, H. Lähdesmäki, Self-organization and missing values in SOM and GTM, Neurocomputing 147 (2015) 60–70, Advances in Self-Organizing Maps Subtitle of the special issue: Selected Papers from the Workshop on Self-Organizing Maps 2012 (WSOM 2012).
[53] Mulugeta Gebregziabher, Stacia M. DeSantis, Latent class based multiple imputation approach for missing categorical data, J. Statist. Plann. Inference 140 (11) (2010) 3252–3262.
[54] Dashiell F. Young-Saver, Jeffrey Gornbein, Sidney Starkman, Jeffrey L. Saver, Handling of missing outcome data in acute stroke trials: Advantages of multiple imputation using baseline and postbaseline variables, J. Stroke Cerebrovasc. Dis. 27 (12) (2018) 3662–3669.
[55] Michael Schomaker, Christian Heumann, Model selection and model averaging after multiple imputation, Comput. Statist. Data Anal. 71 (2014) 758–770.
[56] Geert J.M.G. van der Heijden, A. Rogier T. Donders, Theo Stijnen, Karel G.M. Moons, Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: A clinical example, J. Clin. Epidemiol. 59 (10) (2006) 1102–1109.
[57] Qixuan Chen, Myunghee Cho Paik, Minjin Kim, Cuiling Wang, Using link-preserving imputation for logistic partially linear models with missing covariates, Comput. Statist. Data Anal. 101 (2016) 174–185.
[58] Ian R. Dohoo, Christel R. Nielsen, Ulf Emanuelson, Multiple imputation in veterinary epidemiological studies: a case study and simulation, Prevent. Vet. Med. 129 (2016) 35–47.
[59] Matthias Templ, Alexander Kowarik, Peter Filzmoser, Iterative stepwise regression imputation using standard and robust methods, Comput. Statist. Data Anal. 55 (10) (2011) 2793–2806.
[60] Nicola Sartori, Alberto Salvan, Karl Thomaseth, Multiple imputation of missing values in a cancer mortality analysis with estimated exposure dose, Comput. Statist. Data Anal. 49 (3) (2005) 937–953.
[61] L. Christine Bono, Douglas Ried, Carole Kimberlin, Bruce Vogel, Missing data on the center for epidemiologic studies depression scale: A comparison of 4 imputation techniques, Res. Soc. Adm. Pharm. 3 (1) (2007) 1–27.
[62] Nathaniel T. Ondeck, Michael C. Fu, Laura A. Skrip, Ryan P. McLynn, Edwin P. Su, Jonathan N. Grauer, Treatments of missing values in large national data affect conclusions: The impact of multiple imputation on arthroplasty research, J. Arthrop. 33 (3) (2018) 661–667.
[63] N.-M. Shara, J.-G. Umans, W. Wang, B.-V. Howard, H.-E. Resnick, Assessing the impact of different imputation methods on serial measures of renal function: The strong heart study, Kidney Int. 71 (7) (2007) 701–705.
[64] María Elisa Quinteros, Siyao Lu, Carola Blazquez, Juan Pablo Cárdenas-R, Ximena Ossa, Juana-María Delgado-Saborit, Roy M. Harrison, Pablo Ruiz-Rudolph, Use of data imputation tools to reconstruct incomplete air quality datasets: A case-study in temuco, Chile, Atmos. Environ. 200 (2019) 40–49.
[65] Mirjam J. Knol, Kristel J.M. Janssen, A. Rogier T. Donders, Antoine C.G. Egberts, E. Rob Heerdink, Diederick E. Grobbee, Karel G.M. Moons, Mirjam I. Geerlings, Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example, J. Clin. Epidemiol. 63 (7) (2010) 728–736.
[66] Nathaniel T. Ondeck, Michael C. Fu, Laura A. Skrip, Ryan P. McLynn, Jonathan J. Cui, Bryce A. Basques, Todd J. Albert, Jonathan N. Grauer, Missing data treatments matter: an analysis of multiple imputation for anterior cervical discectomy and fusion procedures, Spine J. 18 (11) (2018) 2009–2017.
[67] David Vergouw, Martijn W. Heymans, Daniëlle A.W.M. van der Windt, Nadine E. Foster, Kate M. Dunn, Henriette E. van der Horst, Henrica C.W. de Vet, Missing data and imputation: A practical illustration in a prognostic study on low back pain, J. Manipulative Physiol. Ther. 35 (6) (2012) 464–471.
[68] José M. Jerez, Ignacio Molina, Pedro J. García-Laencina, Emilio Alba, Nuria Ribelles, Miguel Martín, Leonardo Franco, Missing data imputation using statistical and machine learning methods in a real breast cancer problem, Artif. Intell. Med. 50 (2) (2010) 105–115, 22 Authors Suppressed Due to Excessive Length.
[69] Ian K. McDonough, Daniel L. Millimet, Missing data, imputation, and endogeneity, J. Econometrics 199 (2) (2017) 141–155, The Creative Mind in Econometrics: Studies in Celebration of Robert Basmann's 90th Year on Causation, Identification and Structural Equation Estimation.
[70] Alicia S. Chua, Svetlana Egorova, Mark C. Anderson, Mariann Polgar-Turcsanyi, Tanuja Chitnis, Howard L. Weiner, Charles R.G. Guttmann, Rohit Bakshi, Brian C. Healy, Using multiple imputation to efficiently correct cerebral MRI whole brain lesion and atrophy data in patients with multiple sclerosis, NeuroImage 119 (2015) 81–88.
[71] Sherene E. Sharath, Nader Zamani, Panos Kougias, Soeun Kim, Missing data in surgical datasets: A review of pertinent issues and solutions, J. Surg. Res. 232 (2018) 240–246.
[72] Archana Purwar, Sandeep Kumar Singh, Hybrid prediction model with missing value imputation for medical data, Expert Syst. Appl. 42 (13) (2015) 5621–5631.
[73] Hao Cheng, Importance sampling imputation algorithms in quantile regression with their application in CGSS data, Math. Comput. Simulation 188 (2021) 498–508.
[74] Ian R. White, Rhian Daniel, Patrick Royston, Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables, Comput. Statist. Data Anal. 54 (10) (2010) 2267–2275.
[75] Frans E.S. Tan, Shahab Jolani, Hilde Verbeek, Guidelines for multiple imputations in repeated measurements with time-dependent covariates: a case study, J. Clin. Epidemiol. 102 (2018) 107–114.
[76] Richard A. Burns, Peter Butterworth, Kim M. Kiely, Allison A.M. Bielak, Mary A. Luszcz, Paul Mitchell, Helen Christensen, Chwee Von Sanden, Kaarin J. Anstey, Multiple imputation was an efficient method for harmonizing the mini-mental state examination with missing item-level data, J. Clin. Epidemiol. 64 (7) (2011) 787–793.
[77] Rupert Weaver, Cattram D. Nguyen, Jocelyn Chan, Keoudomphone Vilivong, Jana Y.R. Lai, Ruth Lim, Catherine Satzke, Malisa Vongsakid, Paul N. Newton, Kim Mulholland, Amy Gray, Audrey Dubot-Pérès, David A.B. Dance, Fiona M. Russell, The effectiveness of the 13-valent pneumo-coccal conjugate vaccine against hypoxic pneumonia in children in lao people's democratic republic: An observational hospital-based test-negative study, Lancet Reg. Health - West. Pac. 2 (2020) 100014.
[78] Agus Salim, Andrew Mackinnon, Helen Christensen, Kathleen Griffiths, Comparison of data analysis strategies for intent-to-treat analysis in pre-test–post-test designs with substantial dropout rates, Psychiatry Res. 160 (3) (2008) 335–345.
[79] A. Hapfelmeier, K. Ulm, Variable selection by random forests using data with missing values, Comput. Statist. Data Anal. 80 (2014) 129–139.
[80] Karen A. Ertel, Ken Kleinman, Lenie van Rossem, Sharon Sagiv, Henning Tiemeier, Albert Hofman, Vincent W.V. Jaddoe, Hein Raat, Maternal perinatal depression is not independently associated with child body

mass index in the generation r study: methods and missing data matter, J. Clin. Epidemiol. 65 (12) (2012) 1300–1309.
[81] Apostolos Papageorgiou, André Miede, Stefan Schulte, Dieter Schuller, Ralf Steinmetz, Decision support for web service adaptation, Pervasive Mob. Comput. 12 (2014) 197–213.
[82] Yulia Sidi, Ofer Harel, Incomplete data analysis of non-inferiority clinical trials: Difference between binomial proportions case, Contemp. Clin. Trials Commun. 18 (2020) 100567.
[83] Marijka J. Batterham, Linda C. Tapsell, Karen E. Charlton, Analyzing weight loss intervention studies with missing data: Which methods should be used? Nutrition 29 (7) (2013) 1024–1029.
[84] Maren K. Olsen, Karen M. Stechuchak, Jack D. Edinger, Christi S. Ulmer, Robert F. Woolson, Move over LOCF: Principled methods for handling missing data in sleep disorder trials, Sleep Med. 13 (2) (2012) 123–132.
[85] Antonia J. Henry, Nathanael D. Hevelone, Stuart Lipsitz, Louis L. Nguyen, Comparative methods for handling missing data in large databases, J. Vasc. Surg. 58 (5) (2013) 1353–1359, e6.
[86] Kristel J.M. Janssen, A. Rogier T. Donders, Frank E. Harrell, Yvonne Vergouwe, Qingxia Chen, Diederick E. Grobbee, Karel G.M. Moons, Missing covariate data in medical research: To imputeis better than to ignore, J. Clin. Epidemiol. 63 (7) (2010) 721–727.
[87] Rosemary Tawn, Jethro Browell, Iain Dinwoodie, Missing data in wind farm time series: Properties and effect on forecasts, Electr. Power Syst. Res. 189 (2020) 106640.
[88] Ji ping Tan, Nan Li, Xiao yang Lan, Shi ming Zhang, Bo Cui, Li xin Liu, Xin He, Lin Zeng, Li yuan Tau, Hua Zhang, Xiao xiao Wang, Lu ning Wang, Yi ming Zhao, The impact of methods to handle missing data on the estimated prevalence of dementia and mild cognitive impairment in a cross-sectional study including non-responders, Arch. Gerontol. Geriatr. 73 (2017) 43–49.
[89] I.C. Olsen, T.K. Kvien, T. Uhlig, Consequences of handling missing data for treatment response in osteoarthritis: a simulation study, Osteoarthr. Cartil. 20 (8) (2012) 822–828.
[90] Panagiotis Sentas, Lefteris Angelis, Categorical missing data imputation for software cost estimation by multinomial logistic regression, J. Syst. Softw. 79 (3) (2006) 404–414.
[91] Lawrence Joseph, Patrick Bélisle, Hala Tamim, John S. Sampalis, Selection bias found in interpreting analyses with missing data for the prehospital index for trauma, J. Clin. Epidemiol. 57 (2) (2004) 147–153.
[92] Chang Wook Jeong, Samuel L. Washington, Annika Herlemann, Scarlett L. Gomez, Peter R. Carroll, Matthew R. Cooperberg, And end results prostate with watchful waiting database: Opportunities and limitations, Eur. Urol. 78 (3) (2020) 335–344.
[93] An Creemers, Marc Aerts, Niel Hens, Geert Molenberghs, A nonparametric approach to weighted estimating equations for regression analysis with missing covariates, Comput. Statist. Data Anal. 56 (1) (2012) 100–113.
[94] Anja Žnidaršič, Anuška Ferligoj, Patrick Doreian, Non-response in social networks: The impact of different non-response treatments on the stability of block models, Social Networks 34 (4) (2012) 438–450.
[95] Ali Idri, Ibtissam Abnane, Alain Abran, Missing data techniques in analogy-based software development effort estimation, J. Syst. Softw. 117 (2016) 595–611.
[96] Mark Huisman, Christian Steglich, Treatment of non-response in longitudinal network studies, Social Networks 30 (4) (2008) 297–308.
[97] Shin-Feng Chen, Shuyi Wang, Chen-Yuan Chen, A simulation study using EFA and CFA programs based the impact of missing data on test dimensionality, Expert Syst. Appl. 39 (4) (2012) 4026–4031.
[98] Shin-Soo Kang, Michael D. Larsen, Tests of independence in incomplete multi-way tables using likelihood functions, J. Korean Stat. Soc. 41 (2) (2012) 189–198.
[99] K. Pelckmans, J. De Brabanter, J.A.K. Suykens, B. De Moor, Handling missing values in support vector machine classifiers, Neural Netw. 18 (5) (2005) 684–692, IJCNN 2005.
[100] Philip L. Roth, Fred S. Switzer, A monte Carlo analysis of missing data techniques in a HRM setting, J. Manag. 21 (5) (1995) 1003–1023.
[101] Waqas R. Shaikh, Martin A. Weinstock, Allan C. Halpern, Susan A. Oliveria, Alan C. Geller, Stephen W. Dusza, The characterization and potential impact of melanoma cases with unknown thickness in the united states' surveillance, epidemiology, and end results program, 1989–2008, Cancer Epidemiol. 37 (1) (2013) 64–70.
[102] Marc H. Gorelick, Bias arising from missing data in predictive models, J. Clin. Epidemiol. 59 (10) (2006) 1115–1123.
[103] Peter C. Austin, Michael D. Escobar, Bayesian modeling of missing data in clinical research, Comput. Statist. Data Anal. 49 (3) (2005) 821–836.
[104] Doh-Soon Kwak, Kwang-Jae Kim, A data mining approach considering missing values for the optimization of semiconductor-manufacturing processes, Expert Syst. Appl. 39 (3) (2012) 2590–2596.
[105] Helen M. Parsons, William G. Henderson, Jeanette Y. Ziegenfuss, Michael Davern, Waddah B. Al-Refaie, Missing data and interpretation of cancer surgery outcomes at the American college of surgeons national surgical quality improvement program, J. the American College of Surgeons 213 (3) (2011) 379–391.
[106] D.J. Stekhoven, P. Bühlmann, MissForest—non-parametric missing value imputation for mixed-type data, Bioinformatics 28 (1) (2012) 112–118.
[107] F. López-Martínez, E.R. Núñez Valdez, R.G. Crespo, et al., An artificial neural network approach for predicting hypertension using NHANES data, Sci. Rep. 10 (1) (2020) 10620.