Thailand Statistician
April 2026; 24(2): 321-338
http://statassoc.or.th
Contributed paper

A Comparative Study of Imputation Techniques for Handling Multivariate Missing Completely at Random in Numeric Datasets

Ratchaneewan Paisanwarakiat [a], Anamai Na-udom*[a] and Jaratsri Rungrattanaubol [b]

[a] Department of Mathematics, Faculty of Science, Naresuan University, Phitsanulok, Thailand

[b] Department of Computer Science and information Technology, Faculty of Science, Naresuan University, Phitsanulok, Thailand

*Corresponding author; e-mail: anamain@nu.ac.th

Received: 1 April 2025
Revised: 14 October 2025
Accepted: 23 October 2025

Abstract

Effective handling of missing data is essential in Big Data analytics, as missing values, particularly those occurring randomly across input variables, can significantly affect the reliability and accuracy of results. This research compares and evaluates the effectiveness of eleven imputation techniques: Mean Imputation (MI), Median Imputation (MEI), Deterministic Linear Regression (DLR), Stochastic Linear Regression (SLR), Bayesian Linear Regression (BLR), Bootstrap Linear Regression (BTLR), Predictive Mean Matching (PMM), Expectation-Maximization (EM), K-Nearest Neighbors with Median (KNNM), K-Nearest Neighbors with Weighted Average (KNNW), and Random Forest (RF). The study utilized nine numeric datasets of varying sizes: three small, three medium, and three large, to assess these techniques. Multivariate missing data were simulated using the Missing Completely at Random (MCAR) mechanism, with missing rates ranging from 10% to 50%. The effectiveness of imputation techniques is evaluated using NRMSE, while their performance consistency is tested with Kendalls W test. The results indicated that MI outperformed MEI. Among the linear regression techniques, DLR excelled compared to the other methods, including SLR, BLR, and BTLR. Additionally, KNNW demonstrated better performance than KNNM. In terms of overall dataset performance, RF, KNNW, and EM were the top performers. For recommending imputation techniques, EM is most suitable for small datasets, KNNW or EM are effective for medium datasets, and RF shows the best performance for large datasets. However, both RF and KNNW demand considerably longer processing times, particularly with large datasets. These insights provide practical guidance for selecting the most appropriate imputation method based on the characteristics of the dataset.

Keywords: Linear regression imputation, predictive mean matching, expectation-maximization imputation, K-nearest neighbors imputation, random forest imputation, normalized root mean squared error

1. Introduction

In today's digital era, data is crucial across various sectors, driving business decision-making, marketing and trends analysis, and the advancement of new technologies. The term "Big Data" has gained prominence due to the rapid growth of digital technology and internet connectivity, enabling

the collection of large amounts of data from diverse sources such as social media, online transactions, healthcare records, and sensor data from IoT (Internet of Things) devices. The collection of data is generally automated. The "5Vs" characteristics of big data are Volume, Velocity, Variety, Veracity and Value. Veracity, which refers to the quality and accuracy of data, including handling missing data, plays the key role in the success of big data analytics.

Although Big Data offers numerous benefits, it also presents many challenges. One of the most significant challenges is missing data, which can occur for several reasons, including incomplete data entry, data loss during collection, or errors in sensors and devices used to gather data. The presence of missing data in large datasets can significantly impact the accuracy and reliability of analyses and decisions that rely on complete and accurate data. Missing data can occur randomly or in a non-random (systematic) manner.

Missing data mechanisms are classified into three types Little and Rubin (2002): Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). The effective handling of missing data is a crucial and important area of focus. There are two primary methods for handling missing data. The first approach involves ignoring missing data by removing records with missing values, a method known as complete case analysis. The second approach involves replacing missing values with estimates derived from other data in the dataset, known as missing data imputation (Han et al., 2012). The main issue with complete case analysis is that it excludes any cases with missing data, which can lead to biased results and reduced statistical power, especially if the data are not missing completely at random. This exclusion can result in a loss of valuable information. Consequently, researchers often prefer imputation methods, which handle missing data more effectively by estimating missing values, thereby preserving the sample size and reducing potential bias. In this research, the focus is on handling missing data through imputation techniques.

Effective management of missing data is crucial for harnessing the full potential of Big Data, leading to more reliable and insightful analyses. Limitations associated with complete case analysis can impede the development of more efficient models (Ezzine and Benhlima, 2018). Many studies have investigated various data imputation techniques. Nookhong and Kaewrattanapat (2015) found that Decision Trees were effective for imputing fully discrete data, while K-Nearest Neighbors (KNN) worked well for continuous data, and Naive Bayes was suitable for ordinal data. Schmitt et al. (2015) demonstrated that Fuzzy K-Means (FKM) imputation outperformed other methods in small datasets, whereas Bayesian Principal Component Analysis (BPCA) excelled in larger datasets. Chhabra and Vashisht (2017), in a comparison of the MICE package in R, identified Bayesian Linear Regression (BLR) as superior, while Le et al. (2018) highlighted Expectation-Maximization (EM) method as the most effective for missing data imputation in healthcare datasets. Jadhav et al. (2019) found that K-Nearest Neighbors with Weighted Average (KNNW) demonstrated strong representational performance across various datasets. Mohammed et al. (2021) indicated that when handling missing observations in a univariate continuous frequency table, MICE is the best method. Paisanwarakiat et al. (2022) found that KNNM performed best in a multivariate missing data context. Thongsri and Samart (2022) proposed a three-component imputation approach that combines KNN, SLR, and RF, which demonstrated superior performance compared to other methods. Suh and Song (2023) reported that KNN and a multiple imputation through XGBoost (mixgb) techniques performed better than RF on both numerical and categorical datasets. Rahman et al. (2023) found that DLR and PMM were superior to other imputation techniques when applied to numeric survey datasets. These findings emphasize that the effectiveness of an imputation method depends on factors such as dataset size, the nature of the missing data, and the chosen imputation technique, indicating that no single method is universally superior.

Table 1 displays various imputation techniques, including simple methods, statistically based estimations, model-based approaches, and those inspired by machine learning. Most comparative studies analyzed data from the UCI Machine Learning Repository, focusing on missing data mechanisms classified as MCAR, as well as multivariate missing patterns, primarily involving numeric data. The techniques that demonstrated the best performance are highlighted in bold, such as KNNW,

KNNM, BLR, EM, and PMM. Some studies evaluated a single dataset, while others assessed multiple datasets. The effectiveness of these techniques was measured by the error rate of the estimations.

However, the reviewed papers did not examine how different dataset sizes impact the effectiveness of imputation, nor did they compare the performance of various imputation techniques. In this paper, we will explore how dataset size and different grouped imputation methods affect the imputation process.

As described before, this study aims to compare the effectiveness of eleven different imputation techniques. These techniques were tested on nine numeric datasets, as shown in Table 2. These datasets varied in size, comprising three small, three medium, and three large datasets, with missing data percentages ranging from 10% to 50%. The analyses were conducted using MCAR mechanisms and a multivariate missing data pattern. The evaluation criterion employed was the normalized root mean square error (NRMSE), calculated over ten iterations. The eleven imputation techniques examined in this study are as follows: Mean Imputation (MI), Median Imputation (MEI), Deterministic Regression (DLR), Stochastic Regression (SLR), Bayesian Regression (BLR), Bootstrap Regression (BTLR), Predictive Mean Matching (PMM), Expectation-Maximization (EM), K-Nearest Neighbors with Median (KNNM), K-Nearest Neighbors with Weighted Average (KNNW), and Random Forest (RF), which are categorized into four groups: Simple, Regression, KNN, and other imputations.

The organization of this paper is structured as follows: Section 2 discusses theoretical frameworks and related work; Section 3 outlines the research methodology; Section 4 presents the results and discussion; and Section 5 provides conclusions and suggestions for further research.

2. Theoretical and Related Work

Table 1 clearly indicates that most of the examined studies involve patterns of multivariate (Multi) missing data, with only a few focusing on univariate (Uni). Most of the data are numeric and follows an MCAR mechanism which has been tested across varying numbers of datasets (No.). The percentage of missing data typically ranges from 10% to 50%. The most commonly used imputation methods include MI, MEI, MICE, KNNW, and LR along with its variants. Among these techniques, KNNW, BLR, EM, PMM, and MICE are identified as the most effective methods for imputation, as highlighted in bold.

Table 1 Summary of review articles on numerical imputations

	Uni	Multi	No.		Mech.
(Schmitt et al. 2015) A Comparison of Six Methods for Missing Data Imputation		✓	4	MI, **FKM**, SVD, **BPCA**, MICE, KNNW	MCAR
(Chhabra et al. 2017) A Comparison of Multiple Imputation Methods for Data with Missing Values		✓	1	PMM, **BLR**, SLR, DLR, RF, CART	MCAR
(Le et al. 2018) Comparison of the Most Influential Missing Data Imputation Algorithms for Healthcare		✓	2	MI, MICE, **EM**, KNNW	MCAR
(Jadhav et al. 2019) Comparison of Performance of Data Imputation Methods for Numeric Dataset		✓	5	MI, MEI, LR, PMM, BLR, SLR, Sample, **KNNW**	MCAR
(Mohammed et al. 2021) Comparison of five imputation methods in handling missing data in a continuous frequency table	✓		4	MI, MEI, Sample, **MICE**, KNNW	MCAR
(aisanwarakiat et al. 2022) Comparing Statistical and K Nearest Neighbor Imputation Techniques in Diabetes Dataset		✓	1	MI, MEI, LR, **KNNM**, KNNW	MCAR
(Suh and Song, 2023) A Comparison of Imputation Methods Using Machine Learning Models		✓	6	MI, MO, **KNN**, MICE, RF **mixgb**	MCAR
(Rahman et al., 2023) Comparison of Imputation Methods on Different Types of Survey Research Data: A Continuous Variable		✓	2	MI, MEI, **DLR**, SLR, **PMM**	-

Note: Mech. is a type of missing data mechanism; SVD is Singular Value Decomposition; BPCA is Bayesian Principal Component Analysis.; mixgb is a multiple imputation through XGBoost.

2.1. Missing data mechanisms

According to Little and Rubin (2002), there are three distinct mechanisms that explain the occurrence of missing data:

Missing Completely at Random (MCAR): The assumption of MCAR is that the probability of the missingness is independent of all measured and unmeasured characteristics of the individuals under study.

$$Pr\left(R = 0|Y_{obs}, Y_{mis}, \psi\right) = Pr\left(R = 0|\psi\right) \tag{1}$$

Missing at Random (MAR): The assumption of MAR is that the probability of the missingness depends on the observed values but does not depend on the missing values.

$$Pr\left(R = 0|Y_{obs}, Y_{mis}, \psi\right) = Pr\left(R = 0|Y_{obs}, \psi\right) \tag{2}$$

Missing Not at Random (MNAR): The assumption of MNAR is that the probability of the missingness also depends on the values of the missing data (Jadhav et al., 2019).

$$Pr\left(R = 0|Y_{obs}, Y_{mis}, \psi\right) \tag{3}$$

where R is a matrix storing the locations of the missing data in Y. The distribution of R may depend on $Y = (Y_{obs}, Y_{mis})$, either by design or by happenstance, and this relation is described by the missing data model. Let ψ contain the parameters of the missing data model Buuren (2018).

2.2. Statistical imputation

Statistical imputation is a method used to fill in missing data points by estimating their values based on statistical principles. It employs simple methods such as mean, median, and mode, as well as more advanced techniques, including LR, PMM, and EM. These methods rely on specific assumptions about data distribution or the relationships between variables to accurately estimate the missing values.

2.2.1. Simple imputation

Simple imputation replaces missing values with existing data attributes (Garca et al., 2009). For quantitative data, the mean or median is used, and for qualitative data, the mode is applied. Due to its simplicity, simple imputation serves as a benchmark technique (Jerez et al., 2010). In this study, we used mean and median imputations for quantitative data.

Mean Imputation (MI) is a widely used method that involves replacing missing values with the mean of the observed values for each variable (Jamshidian and Bentler, 1999).

Median Imputation (MEI) is a method that replaces missing values with the median of the observed values for each variable, as opposed to the mean (Zhang, 2016).

2.2.2. Regression imputation (LR)

Regression Imputation, or conditional mean imputation, is a statistical method for estimating missing values in data. It uses other variables to create a model that predicts the dependent variable with missing values and fills in the gaps based on this model (Song and Shepperd, 2007). The regression model can be formulated as Equation (4). According to the definition of the error term, there are four different types of linear regression models Buuren (2018).

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \epsilon \tag{4}$$

where Y is the dependent variable, X_i is the independent variables, and β_i is the regression coefficients.

Deterministic Regression Imputation (DLR) is a fixed and exact relationship between the variables, where no consideration is given to random error or uncertainty.

Stochastic Regression Imputation (SLR) is an enhancement of regression imputation that aims to mitigate correlation bias by introducing noise to the predictions, where ϵ is randomly sampled from a normal distribution: $\epsilon \sim N\left(0, \sigma^2\right)$.

Bayesian Regression Imputation (BLR) methods draw the parameter directly from their posterior distributions. Where $\epsilon \sim N\left(0, \sigma^2\right), \beta_i$ and σ are random draws from their posterior distributions.

Bootstrap Regression Imputation (BTLR) methods resample the observed data and re-estimate the parameters from the resampled data. Where β_i and σ are least squares estimates calculated from a bootstrap sample taken from the observed data.

2.2.3. Predictive mean matching (PMM)

Predictive Mean Matching is a statistical technique for filling in missing data by using values from similar observations based on a specific set of predictor variables. The PMM process is structured as follows (Little, 1988; Buuren, 2018):

Model Building A predictive regression model is built from the variables of observed data to predict the missing values.

Prediction The predicted values from the regression model are not directly used to replace the missing data. Instead, the model searches for the most similar observed data to them.

Matching The process matches the observed values whose predicted values from the regression model are closest to the predicted value of the missing case.

Imputation The actual observed values corresponding to these closest predicted values are used to replace the missing data.

PMM helps preserve the distribution of the variable with missing values and can handle complex relationships between variables. However, it can be sensitive to outliers in the predictor variables and computationally expensive, especially for large datasets.

2.2.4. Expectation maximization (EM)

Expectation maximization is a method used to find the maximum likelihood estimates of parameters in probabilistic models, particularly when dealing with missing data. The procedure involves an iterative process that alternates between two algorithms: the Expectation (E-step) and the Maximization (M-step) algorithms (Dempster et al, 1977; McLachlan and Krishnan, 1997). During the E-step, expected values are calculated based on all available complete data points. In the subsequent M-step, the missing values are filled in using the values generated during the E-step, and new expected values are then recalculated. This two-step process continues iteratively until the changes in expected values from one iteration to the next become negligible (Hedderley and Wakeling, 1995).

for $t = 1, 2, 3, \ldots, N$

E-Step: $Q\left(\theta|\theta^{(t-1)}\right) = E_{(t-1)}\left(log\ L\left(\theta; X\right)|Y = y\right)$

M-Step: $\theta^{(t)} = \arg\max_{\theta} Q(\theta \mid \theta^{(t-1)})$

end for

restore $\theta^{(N)}$

where X is a vector of the missing data variable, Y is a vector of data variable, θ is a parameter of interest in the joint distribution of X and Y, and $L\left(\theta; X|y\right)$ is the maximum likelihood function of θ.

The stopping criteria for the EM algorithm depend on the specific application. Typically, these criteria include reaching a set number of iterations, achieving algorithm convergence, or attaining a specific accuracy level in imputation. The EM algorithm is efficient for certain models and works with various probabilistic models. However, it may converge slowly, and the choice of initial parameter estimates can impact its convergence.

2.3. Machine learning imputation

Machine learning imputation is a widely used technique for handling missing data in datasets by predicting or estimating missing values based on existing data patterns. It leverages algorithms like

KNN, RF, and Neural Networks to identify complex patterns and relationships, effectively managing non-linear relationships and variable interactions. Unlike statistical imputation, which relies on assumptions about data distributions and is generally easier to interpret, machine learning imputation sometimes offers more flexibility and accuracy in estimating missing values.

2.3.1. K Nearest Neighbor (KNN)

K-Nearest Neighbor imputation is a simple, yet powerful machine learning algorithm based on instance-based learning. It handles missing data in datasets by finding the nearest neighbors to a missing value and using their average or mode for imputation. The process of KNN to impute missing data can be briefly explained as follows (Troyanskaya et al., 2001):

Step 1: Determine k Select the number of nearest neighbors (k) to consider for imputation.

Step 2: Calculate distances Calculate Euclidean distances, as defined in Equation (5), between the missing data point and all other observed data in the dataset.

$$d\left(x_m, x_o\right) = \sqrt{\sum_{i=1}^{n} \left(x_{mi} - x_{oi}\right)^2} \tag{5}$$

Where x_m is a data point with missing value, x_o is a observation data, i is an index of variables, which have no missing value and x_{mi} and x_{oi} are values of variable i on x_m,x_o respectively, and n is the number of variables.

Step 3: Find k nearest neighbors Choose k nearest observed data to the missing value, based on the calculated distances.

Step 4: Impute value For numeric data, calculate the mean or median of the values of the k nearest neighbors and use it to impute the missing value.

KNN imputation with median method (KNNM) uses the median of k neighbors to replace missing value.

KNN imputation with weighted average method (KNNW) uses a weighted average of the values of k neighbors (Torgo, 2016) to replace missing value (Zhang, 2012). The missing value $(\hat{x}_{mis})$ is estimated with the weighted average of its k nearest neighbors as explained in Equation (6).

$$\hat{x}_{mis} = \frac{\sum_{j=1}^{k} w_j x_j}{W}, \tag{6}$$

where w_j is an inverse distance weight, $w_j = \frac{1}{d(x_{mis}, x_j)^2}$, x_j is a k nearest neighbors observation, and W is the sum of distance weights of the k nearest neighbours.

2.3.2. Random Forest (RF)

Random Forest imputation is a technique for handling missing data by utilizing the power of RF, an ensemble learning method that combines multiple decision trees to make predictions (Shah et al., 2014; Tang and Ishwaran, 2017; Breiman, 2001). The process can be summarized as follows:

Step 1: Initially, the missing values in datasets are inputted using simple techniques such as mean or median.

Step 2: A RF model is constructed from the imputed dataset. This model learns the relationships between variables and enabling it to predict missing values based on the available information.

Step 3: The trained RF model is then applied to predict the missing values in the original dataset. For each missing value, the model considers the values of other variables in that row and predicts the most likely value based on the patterns it has learned during training.

Step 4: The imputation process can be iterated multiple times. After the first round, the imputed values are used to retrain the RF model, allowing for more accurate predictions in subsequent rounds. This iterative process continues until the imputed values converge or a maximum number of iterations is reached.

RF can capture non-linear relationships between variables and can identify the most important variables for imputation. However, if the model becomes too complex, it may overfit the training data and perform poorly on new data. Additionally, RF can be computationally intensive, especially for large datasets and numerous iterations.

3. Research Methodology

A review of studies showed that most of the studied data were numerical. Nine datasets were selected from the UCI Machine Learning Repository and categorized into three size groups, as detailed in Table 2. The methodology used the Multivariate MCAR mechanism with missing rates of 10%, 20%, 30%, 40%, and 50%. Four groups of imputation techniques addressed the missing values, and the error rate of the estimations was measured using the NRMSE.

3.1. The dataset

The nine datasets employed in this study are numeric and were obtained from the UCI. All datasets are 2-class classification problems with varying independent variables, as outlined in Table 2. They are categorized into three groups based on the number of records:

Small size category includes datasets with fewer than 1,000 records: which include Breast Cancer Coimbra, BUPA Liver Disorders and Pima.

Medium size category comprises datasets with fewer than 10,000 records: which include Banknote Authentication, svmguide1 and Rice.

Large size category consists of datasets with more than 10,000 records, which include Magic, Codrna.

Table 2 Overview of nine selected numeric datasets

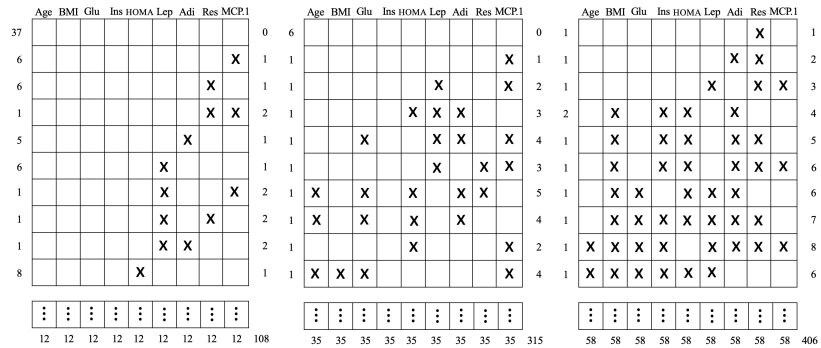
Size	Datasets	No. of Variables	No. of Instances
Small	Breast Cancer Coimbra	9	116
	BUPA Liver Disorders	6	345
	Pima	8	768
Medium	Banknote authentication	4	1,372
	svmguide1	4	3,089
	Rice	7	3,810
Large	Magic	10	19,020
	Codrna	8	59,535
	Skin Segmentation	3	245,057

3.2. Experimental design

To assess the effectiveness of different imputation techniques on datasets of varying sizes, an experiment was conducted using multivariate MCAR with different missing rates. The steps of the experiment, illustrated in Figure 2, are outlined below:

Step 1: Applying multivariate MCAR to the initial datasets with varying missing rates of 10%, 20%, 30%, 40%, and 50%. This process involves randomly removing data points from each variable according to specific rates. For example, considering the Breast Cancer Coimbra dataset, which consists of 9 variables and 116 records. To illustrate the application of multivariate MCAR at a 10% missing rate. In this case, we remove 12 data points from each variable, resulting in a total of 12 $\times$ 9 = 108 missing data points. Consequently, some records may have multiple missing data points, as shown in Figure 1(a). The approach for other missing rates follows a similar pattern, with an increasing number of missing data points. This is depicted in Figures 1(b) and 1(c) for 30% and 50% missing rates, respectively. As the missing rate increases, the density of multiple missing data points within records also rises. This results in a reduction in the number of observation records that contain

no missing data points. This issue becomes particularly evident when the dataset size is small, as illustrated in Figure 1(c), where all records contain missing data.



(a) 10% missing rate **(b)** 30% missing rate **(c)** 50% missing rate

Figure 1 The example pattern of the dataset with 10%, 30%, and 50% missing rate

Step 2: The dataset containing missing data points, obtained from Step 1, was subject to imputation, as shown in Figure 2. The imputation techniques were categorized into four main groups:

1) Simple techniques: MI and MEI

2) Regression techniques: DLR, SLR, BLR, and BTLR

3) KNN Techniques: KNNM and KNNW, with k values of 3, 5, 7, and 9

4) Other techniques: PMM, EM, and RF.

At the end of this step, a dataset with imputed values was obtained.

Step 3: The performance of the imputation techniques was evaluated using the NRMSE criterion. This involved comparing the original dataset with the datasets that included the imputed values from Step 2. The evaluation process was repeated 10 times for each dataset and for each missing rate.

Step 4: The performance of each imputation technique was analyzed by calculating the average NRMSE for each dataset and missing rate.

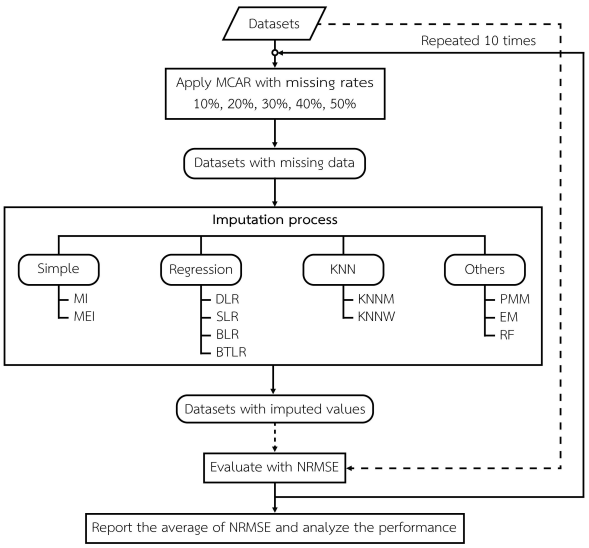


Figure 2 A methodological process for comparative assessment of imputation techniques

3.3. Evaluation method

Performance was assessed using the **Normalized Root Mean Squared Error (NRMSE)**, which measures the error of imputation by comparing imputed dataafter post-processingwith true values, as described in (7) (Dahmani et al., 2016; , Lee and Styczynski, 2018; Jadhav et al., 2019).

NRMSE can be viewed as a standardized version of the **Root Mean Squared Error (RMSE)** (Nugroho et al., 2021), as shown in Equation (8). This normalization makes the error scale independent, allowing for a fair comparison of model performance across datasets that vary in range or units.

$$NRMSE = \frac{RMSE}{|mean\,(X)|} \tag{7}$$

where, X is a set of the actual value (not missing data).

$$RMSE\ = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2}, \tag{8}$$

where, N is the number of observations (i.e., missing data), x_i is the actual (observed) value of observation i, y_i is the imputed (predicted) value of observation i.

3.4. Test statistic used to compare the performance of imputation techniques

To assess the consistency of the performance of imputation techniques across various missing rates, Kendalls W test was utilized. Kendalls coefficient of concordance (W) ranges from 0 to 1, with higher values indicating a stronger agreement among the rankings. The hypotheses for the test were stated as follows (Memon et al., 2023):

Null hypothesis H_0: There is no agreement in the rankings of imputation techniques across missing rates $(W = 0)$

Alternative hypothesis H_1: There is a significant agreement in the rankings of imputation techniques across different missing rates $(W = 1)$

4. Results and Discussion

This study compared 11 different imputation methods, categorized into four groups: Simple, Regression, KNN, and Others. The results, assessed using NRMSE across nine datasets, are initially discussed and analyzed by a group.

Within the Simple group, MI consistently outperformed MEI, demonstrating a lower NRMSE across all datasets. As a result, only the NRMSE values for MI are displayed in Table A1 in the Appendix. In the Regression group, DLR consistently outperformed all other regression techniques across every dataset. Since DLR produces the same results each time it is applied, its outcomes are reliable and predictable. This contrasts with methods such as SLR, BLR, and BTLR, which incorporate randomness and may yield different results with each execution. The NRMSE values for DLR are also presented exclusively in Table A1 in the Appendix.

In the KNN group, the KNNW method outperformed KNNM. This improvement is attributed to KNNW's ability to consider the values of all selected neighbors while applying weights based on their distances, which allows it to give more importance to closer neighbors. As a result, KNNW provides a more accurate representation of the data compared to KNNM, which only considers the median value of the selected neighbors. Therefore, only the results of KNNW are presented and discussed in detail, as shown in Table 3. In this research, the values of k were set to 3, 5, 7, and 9. Table 3 indicates that the lowest imputation error for most datasets occurs when k is set to 9. However, for the Rice and Banknote Authentication datasets, the best results are achieved with k values of 5 and 7. Setting k to 3 is not recommended, as it does not perform well in any of the cases studied.

In the case of KNN, it was observed that in small group datasets, as the amount of missing data increasedparticularly at higher missing ratesit became impossible to estimate the nearest missing values, known as N/A (not applicable), especially when the specified value was set to 9. The presence of N/A was noticeable throughout all datasets in the small group, but it was particularly pronounced in the smallest dataset, Breast Cancer Coimbra.

The issue of not applicable (N/A) values arises because the KNN algorithm requires a sufficient number of observation records to calculate the distances for the missing records. As the rate of missing data increases, the number of available observation records decreases. Consequently, this may result in an insufficient number of records to determine the nearest distances, particularly when k is large. This problem, however, does not occur in the medium and large size datasets, where the datasets contain a significant number of records.

Table 3 NRMSE values for KNNW imputation with various of k

Size	Datasets	k	Missing rates				
			10	20	30	40	50
Small	Breast Cancer Coimbra	3	1.4815	1.5522	1.5343	N/A	N/A
		5	1.4149	1.5274	1.2345	N/A	N/A
		7	1.3603	**1.5153**	N/A	N/A	N/A
		9	**1.3363**	N/A	N/A	N/A	N/A
	BUPA liver disorders	3	0.4710	0.4465	0.4787	0.4679	0.5062
		5	0.4511	0.4216	0.4618	0.4551	0.5141
		7	0.4389	0.4125	0.4569	0.4519	N/A
		9	**0.4338**	**0.4092**	**0.4551**	**0.4501**	N/A
	Pima	3	1.0049	0.9302	0.9266	0.9971	N/A
		5	0.9686	0.8873	0.8930	0.9596	N/A
		7	0.9507	0.8637	0.8783	**0.9462**	N/A
		9	**0.9363**	**0.8586**	**0.8753**	0.9894	N/A
Medium	Banknote authentication	3	2.5396	3.3014	3.8850	4.6422	5.3187
		5	2.5050	**3.2041**	3.7559	4.4485	5.1654
		7	**2.4701**	3.2193	**3.7267**	**4.4329**	5.0704
		9	2.4830	3.2163	3.7309	4.4383	**5.0417**
	svmguide1	3	0.5894	0.6306	0.6603	0.7018	0.7538
		5	0.5576	0.5994	0.6308	0.6730	0.7159
		7	0.5459	0.5873	0.6170	0.6649	0.7029
		9	**0.5393**	**0.5820**	**0.6111**	**0.6577**	**0.6962**
	Rice	3	0.0213	0.0310	0.0474	0.0716	0.1134
		5	**0.0211**	**0.0300**	**0.0465**	**0.0713**	**0.1123**
		7	**0.0211**	**0.0300**	0.0475	0.0720	0.1137
		9	0.0213	0.0307	0.0486	0.0740	0.1174
Large	Magic	3	1.0527	1.0935	1.1377	1.1984	1.3017
		5	1.0037	1.0450	1.0901	1.1545	1.2632
		7	0.9853	1.0270	1.0709	1.1378	1.2479
		9	**0.9757**	**1.0175**	**1.0623**	**1.1302**	**1.2412**
	Codrna	3	1.1344	1.1264	1.1627	1.2158	1.3060
		5	1.0623	1.0698	1.1100	1.1642	1.2510
		7	1.0313	1.0451	1.0868	1.1428	1.2284
		9	**1.0128**	**1.0325**	**1.0748**	**1.1310**	**1.2182**
	Skin Segmentation	3	0.3114	0.3746	0.4243	0.4644	0.4973
		5	0.3053	0.3707	0.4212	0.4611	0.4916
		7	0.3024	0.3688	0.4191	0.4572	0.4880
		9	**0.3009**	**0.3675**	**0.4167**	**0.4545**	**0.4826**

To summarize, Table A1 presents the NRMSE values for the best-performing variations of impu-

tation techniques within each group: Simple, Regression, and KNN. It also includes three techniques from the Others: MI, DLR, PMM, EM, KNNW, and RF.

Regarding PMM imputation, the NRMSE values demonstrated inconsistency in small group datasets. However, in medium and large datasets, the NRMSE values generally increased as the percentage of missing data rose. Overall, PMM did not perform well across all nine datasets.

When applying EM imputation to small group datasets, the NRMSE values exhibited slight fluctuations. In contrast, for medium and large datasets, the NRMSE values tended to rise as the percentage of missing data increased. Overall, EM imputation was found to perform effectively on smaller datasets. Regarding RF imputation, the NRMSE values varied for both small and large size datasets.

However, for medium datasets, NRMSE values generally increased with a higher percentage of missing data. Overall, RF imputation demonstrated strong performance with large group datasets.

Among the various imputation techniques, EM produced better results than PMM. This advantage stems from EM's capacity to systematically refine parameter estimates with each iteration, leading to more accurate and realistic predictions of missing data. In contrast, PMM relies on matching predicted values from a regression model, which may not effectively represent the underlying distribution of the data.

Table 4 presents the two most effective imputation techniques for each dataset. The leading methods are RF, EM, and KNNW. However, in the Breast Cancer dataset, MI proved effective in two instances. Among the smaller datasets, EM performed best on several occasions. For the medium datasets, both KNNW and EM demonstrated strong performance. In the large datasets, RF outperformed both EM and KNNW. Overall, as illustrated in Table 4, the three most effective imputation techniques are EM, RF, and KNNW, although KNNW can be complex due to the difficulty in determining the appropriate value for certain parameter k.

Table 4 Two most effective techniques for each dataset by missing rate

Size	Datasets	Missing rates				
		10	20	30	40	50
Small	Breast Cancer Coimbra	**RF** KNNW(9)	**RF** EM	**KNNW(5)** RF	**Mean** RF	**Mean** RF
	BUPA liver disorders	**EM** DLR	**EM** RF	**EM**, DLR	**EM** RF	**EM** RF
	Pima	**RF** DLR	**EM** DLR	**EM**, RF	**EM** KNNW(7)	**EM** Mean
Medium	Banknote authentication	**KNNW(7)** RF	**KNNW(5)** RF	**KNNW(7)** RF	**KNNW(5)** EM	**EM** KNNW(9)
	svmguide1	**KNNW(9)** RF	**KNNW(9)** RF	**KNNW(9)** RF	**KNNW(9)** EM	**KNNW(9)** EM
	Rice	**EM** RF	**EM** RF	**EM** RF	**EM** RF	**EM** RF
Large	Magic	**RF** KNNW(9)	**RF** KNNW(9)	**RF** KNNW(9)	**RF** EM	**RF** EM
	Codrna	**EM** KNNW(9)	**KNNW(9)** EM	**RF** KNNW(9)	**RF** KNNW(9)	**RF** EM
	Skin Segmentation	**RF** KNNW(9)	**RF** EM	**RF** EM	**EM** RF	**EM** DLR

Table 5 Rank of imputation techniques by size of dataset

Dataset by size	Kendall's W	Chi-square	$p-value$
Small (3 datasets)	0.8263	20.7	0.0009
Medium (3 datasets)	0.8743	21.9	0.0006
Large (3 datasets)	0.9257	23.1	0.0003
All datasets (9)	0.8730	13.1	0.0225

Kendalls W test was employed to evaluate the consistency of various imputation techniques across different rates of missing data. Table 5 presents the ranking of these techniques based on dataset size. For small datasets, EM technique achieved the lowest average rank and consistently performed best across nearly all missing rates, except for the 10% missing rate. This finding suggests that the EM technique is the most effective for imputing missing data in small datasets. The results demonstrate strong agreement in the ranking of imputation techniques across the different missing rates, indicated by W = 0.8263, Chi-square = 20.7, p = 0.0009. Additional details can be found in Table A3 in the Appendix.

For medium-sized datasets, the KNNW technique achieved the lowest average rank and performed best across nearly all missing rate scenarios, except at 50%. The second best-performing technique was EM, indicating that KNNW and EM are the most suitable methods for imputing missing data in medium datasets. The results demonstrate a high level of consistency across various missing rates, showcasing strong agreement in the rankings of techniques (W = 0.8743, Chi-square = 21.9, p = 0.0006). Additional details can be found in Table A4.

For large datasets, the RF technique demonstrated the lowest average rank and performed best across all missing data rates. The next best-performing techniques were EM and KNNW, respectively. This indicates that RF is the most effective technique to impute missing data in large datasets. However, due to its relatively long processing timeas shown in Figure 3 and Table A2 in AppendixEM and KNNW are recommended as more practical alternatives. Overall, RF, EM, and KNNW consistently rank as the top-performing techniques for large datasets. Additional details are provided in Table A5.

Furthermore, when analyzing all datasets categorized by size and averaging across all five missing data rates for each dataset, the three most effective imputation techniques were identified as EM, RF, and KNNW. These results are consistent across various dataset sizes, indicating strong agreement in the ranking of the imputation techniques (W = 0.8730, Chi-square = 13.1, p = 0.0225). Additional details can be found in Table A6.

We also aim to facilitate a discussion centered on the time required for each technique, as summarized in Table A2 in Appendix, which presents the average processing times across all five missing rates for each dataset. The computational experiments were conducted on a dedicated system equipped with an Intel Xeon Gold 5318Y CPU (2.10 GHz) and 8.00 GB of RAM. All data analyses were carried out using the R statistical environment (version 4.2.1). The findings indicate that the size of the dataset has a more pronounced impact on the processing time for each technique than the missing rate itself. Both RF and KNNW demanded substantial processing time, particularly with larger datasets, often taking several hours, while the other techniques generally completed processing in just seconds regardless of dataset size.

Figure 3 illustrates how the processing time for each technique increases as the dataset size grows, using a logarithmic scale (base 10) to highlight the complexity of each method. According to the figure, MI technique shows the slowest increase in complexity with rising data size. In comparison, the techniques of DLR, EM, and PMM display moderate complexity, while RF and KNNW exhibit the highest levels of complexity.

It is important to note that while RF performs well with large datasets, it is also time-consuming and resource-intensive because it involves building multiple decision trees. However, the missRanger package provides an alternative to RF imputation and is reported to process data more quickly than the missForest method used in this study Suh and Song (2023). Similarly, KNN method is time-

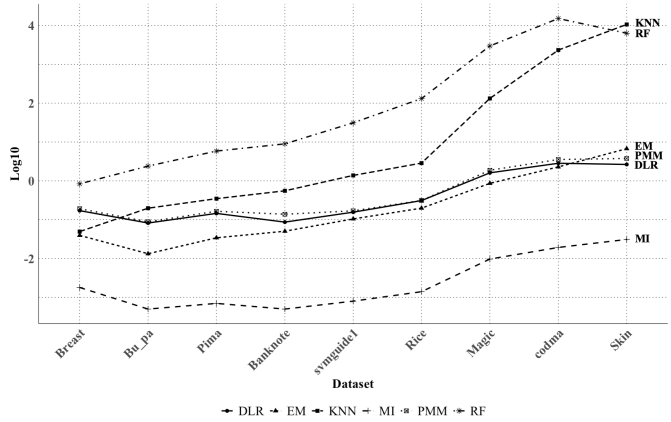


Figure 3 Time required (in Log10 scale) for each imputation technique by dataset

consuming, as it requires calculating the distances between missing values and all observation records in the dataset. Despite efforts by many researchers to improve its speed, KNN remains resource heavy. In contrast, EM typically performs slightly worse but requires significantly less processing time compared to both KNN and RF in most cases.

5. Conclusion and Further Work

Based on our study, the results indicate that MI Imputation outperforms Median Imputation. Additionally, DLR demonstrates superior performance compared to other linear regression imputation variants, while KNNW consistently exceeds KNNM across all cases. We propose the following guidelines for imputing missing data based on dataset size, performance, and processing time considerations:

1. For small datasets, we recommend using the EM method due to its accuracy and speed.
2. For medium-sized datasets, both KNNW and EM show strong performance.
3. For large datasets, although RF outperforms both EM and KNNW, it requires a significant amount of processing time.

Overall, RF, KNNW, and EM have been identified as the most effective techniques. However, given KNNWs dependence on the choice of k, as well as its longer processing times and associated challenges, EM often proves to be a more practical option in various scenarios.

In future work, we plan to conduct a more comprehensive investigation into the performance of EM, KNNW, and RF in the context of classification modeling. Specifically, we aim to evaluate these imputation techniques in combination with various classification algorithms to gain a deeper understanding of their impact on predictive performance. The analysis will place particular emphasis on improving model accuracy, while also considering other important aspects such as MAR (Missing at Random) and MNAR (Missing Not at Random) mechanisms across diverse datasets. Through this extended evaluation, we hope to identify potential strengths and limitations of each method, providing useful guidelines for researchers and practitioners when selecting suitable imputation approaches for real-world classification tasks.

Acknowledgements

The author is deeply grateful to the Development and Promotion of Science and Technology Talents Project (DPST) and the Faculty of Science at Naresuan University for the financial support that made this research possible.

References

Batista GE, Maria CM. (2002). A Study of K-Nearest Neighbour as an Imputation Method. HIS : Soft Computing Systems - Design, Management and Applications, HIS 2002 Dec 1-4; Santiago, Chile, 2002; 30: 251-260.

Breiman L. Random forests. Mach Learn. 2001; 45: 5-32.

Buuren SV. Flexible Imputation of Missing Data, Chapman and Hall/CRC; 2018.

Chhabra G, Vashisht V, Ranjan J. A Comparison of Multiple Imputation Methods for Data with Missing Values. IndianJ. Sci Technol. 2017; 10(9): 1-7.

Dahmani K, Notton G, Voyant C, Dizene R, Nivet ML, Paoli C, Tamas W. Multilayer Perceptron approach for estimating 5-min and hourly horizontal global irradiation from exogenous meteorological data in locations without solar measurements. Renew Energ. 2016; 90: 267-282.

Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J Roy Stat Soc B Met. 1997; 39(1): 1-38.

Ezzine I, Benhlima L. A Study of Handling Missing Data Methods for Big Data. 2018 IEEE 5th International Congress on Information Science and Technology (CiSt), Marrakech, Morocco, 2018. pp. 498-501.

Garca LPJ, Sancho GJL, Figueiras VAR, Verleysen M. K nearest neighbours with mutual information for simultaneous classification and missing data imputation. Neurocomputing. 2009; 72(7-9): 1483-1493.

Gopal K.M, Durgaprasad N, Deepa KS, Sravan RG, Revanth RD. Comparative Analysis Of Different Imputation Techniques For Handling Missing Dataset. Int J Innnov Technol Explor Eng. 2019; 8(7): 347-351.

Han JW, Kamber M, Pei J. Data Mining Concepts and Techniques. 3rd Edition, Morgan Kaufmann Publishers, Waltham; 2012.

Hedderley D, Wakeling I. A comparison of imputation techniques for internal preference mapping using Monte Carlo simulation. Food Qual Prefer. 1995; 6(4): 281-297.

Jadhav A, Pramod D, Ramanathan K. Comparison of Performance of Data Imputation Methods for Numeric Dataset. Appl Artif Intell. 2019; 33(10): 913-933.

Jamshidian M, Bentler PM. ML Estimation of Mean and Covariance Structures with Missing Data Using Complete Data Routines. J Educ Behav Stat. 1999; 24(1): 21-41.

Jerez JM, Molina I, Garcia LPJ, Alba E, Ribelles N, Martin M, Franco L. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artif Intell Med. 2010; 50(2):105-115.

Lee JY, Styczynski MP. NS-kNN: a modified k-nearest neighbors approach for imputing metabolomics data. Metabolomics. 2018; 14(12): 153.

Le TD, Beuran R, Tan Y. Comparison of the Most Influential Missing Data Imputation Algorithms for Healthcare. 10th International Conference on Knowledge and Systems Engineering (KSE); 2018 Nov 1-3; Vietnam. Ho Chi Minh City: IEEE; 2018. pp. 247-251.

Little RJA. Missing-data adjustments in large surveys. J Bus Econ Stat. 1988; 6(3): 287-296.

Little RJA, Rubin DB. Statistical Analysis with Missing Data. Wiley; 2002.

McLachlan GJ, Krishnan T. The EM Algorithm and its Extensions. Wiley; 1997.

Memon SMZ, Wamala R, Kabano IH. A comparison of imputation methods for categorical data. Inform Med Unlocked. 2023; 42: 101382.

Mitra M, Samanta RK. A Study on Missing Data Management. Int J Recent Res Sci Eng Technol. 2017; 5(2): 2347-2693.

Mohammed M, Zulkafli H, Adam M, Ali N, Baba I. Comparison of five imputation methods in handling missing data in a continuous frequency table. AIP Conference Proceeding; 2020 Dec 1-2; Malaysia. Pagoh: AIP Conf. Proc; 2021. pp.61-69.

Nookhong J, Kaewrattanapat N. Efficiency Comparison of Data Mining Techniques for Missing Value Imputation. J Ind Intell Inf. 2015; 3(4): 305-309.

Nugroho H, Utama NP, Surendro K. Class center-based firefly algorithm for handling missing data. J Big Data. 2021; 8(1):37-50.

Paisanwarakiat R, Na-udom A, Rungrattanaubol J. Comparing Statistical and K Nearest Neighbor Imputation Techniques in Diabetes Dataset. Proceeding of The International Conference on Applied Statistics 2022, 2022 Nov 3-4; Thailand. Bangkok: ICAS; 2022. pp. 179-185.

Rahman HAA, Hidayat T, Rahman AA, Razif AM. Comparisons of imputation methods on different types of survey research data: A continuous variable. Proceeding of The 38th International Conference of The Polymer Processing Society 2023, 2023 Sep 19- 20; Shah Alam. Malaysia: AIP; 2023. pp. 050001-050007.

Schmitt P, Mandel J, Guedj M. A Comparison of six methods for missing data imputation. J Biomet Biostat. 2015; 6(1): 224-229.

Shah AD, Bartlett JW, Carpenter J, Nicholas O, Hemingway H. Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. Am J Epidemiol. 2014; 179(6): 764-774.

Song Q, Shepperd M. Missing data imputation techniques. Int J Bus Intell Data Min. 2007; 2(3): 261-291.

Suh H, Song J. A comparison of imputation methods using machine learning models. Commun Stat Appl Methods. 2023; 30(3): 331-341.

Tang F, Ishwaran H. Random Forest Missing Data Algorithms. Stat Anal Data Min. 2017; 10(6):363-377.

Thongsri, T, Samart K. Development of Imputation Methods for Missing Data in Multiple Linear Regression Analysis. Lobachevskii J Math 2022; 43: 33903399.

Torgo L. Data Mining with R: Learning with Case Studies, Second Edition. Chapman and Hall/CRC; 2016.

Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB. Missing value estimation methods for dna microarrays. Bioinformatics. 2001; 17(6): 520-525.

Zhang S. Nearest neighbor selection for iteratively KNN imputation. J Syst Softw. 2012; 85(11): 2541-2552.

Zhang Z. Missing data imputation: focusing on single imputation. Ann Transl Med. 2016; 4(1): 1-8.

Appendix

Table A1 Average NRMSE values for MI, DLR, PMM, EM, KNNW and RF techniques

Size	Datasets	Imputation Techniques	Missing rates				
			10	20	30	40	50
Small	Breast Cancer Coimbra	MI	1.3471	1.4915	1.3500	**1.3239**	**1.3303**
		DLR	1.4571	1.5701	1.5487	1.6691	1.6164
		PMM	1.8599	1.8431	1.6415	1.7314	1.6743
		EM	1.3642	1.4048	1.3327	1.4221	1.5131
		KNNW	1.3363	1.5153	**1.2345**	N/A	N/A
		RF	**1.1997**	**1.3276**	1.3246	1.3439	1.3502
	BUPA liver disorders	MI	0.4775	0.4519	0.4807	0.4595	0.4802
		DLR	0.4241	0.3958	0.4423	0.4876	0.4820
		PMM	0.6004	0.5654	0.5434	0.5667	0.5819
		EM	**0.4203**	**0.3868**	**0.4174**	**0.4164**	**0.4294**
		KNNW	0.4338	0.4092	0.4551	0.4501	0.5062
		RF	0.4316	0.4041	0.4494	0.4450	0.4663
	Pima	MI	1.0674	0.9494	0.9320	0.9643	0.9999
		DLR	0.9189	0.8554	0.8998	1.0276	1.0051
		PMM	1.2748	1.1768	1.1831	1.2461	1.2572
		EM	0.9191	**0.8411**	**0.8240**	**0.8772**	**0.9267**
		KNNW	0.9363	0.8586	0.8753	0.9462	N/A
		RF	**0.9107**	0.8703	0.8721	0.9634	1.0206

Table A1 (continued) Average NRMSE values for MI, DLR, PMM, EM, KNNW and RF techniques

Size	Datasets	Imputation Techniques	Missing rates				
			10	20	30	40	50
Medium	Banknote authentication	MI	6.2834	6.2139	6.3255	6.3160	6.3070
		DLR	4.0126	4.3232	4.5401	4.9175	5.3121
		PMM	5.2359	5.7914	6.0502	6.5352	6.8593
		EM	4.0027	4.2766	4.4218	4.7222	**4.9933**
		KNNW	**2.4701**	**3.2041**	**3.7267**	**4.4329**	5.0417
		RF	2.6572	3.6850	4.0815	4.7716	5.2409
	svmguide1	MI	0.8276	0.8312	0.8300	0.8309	0.8332
		DLR	0.6039	0.6348	0.6575	0.6987	0.7443
		PMM	0.8240	0.8649	0.8923	0.9234	0.9728
		EM	0.6036	0.6320	0.6487	0.6773	0.6986
		KNNW	**0.5393**	**0.5820**	**0.6111**	**0.6577**	**0.6962**
		RF	0.5413	0.6044	0.6394	0.6977	0.7474
	Rice	MI	0.2470	0.2490	0.2477	0.2482	0.2499
		DLR	0.0225	0.0331	0.0501	0.0769	0.1128
		PMM	0.0205	0.0331	0.0525	0.0782	0.1150
		EM	**0.0092**	**0.0166**	**0.0324**	**0.0472**	**0.0722**
		KNNW	0.0211	0.0300	0.0465	0.0713	0.1123
		RF	0.0127	0.0195	0.0363	0.0509	0.0775
Large	Magic	MI	1.2555	1.2552	1.2577	1.2587	1.2613
		DLR	1.0688	1.0829	1.1608	1.2652	1.2459
		PMM	1.4822	1.4908	1.5060	1.5398	1.5574
		EM	1.0673	1.0770	1.0883	1.1023	1.1191
		KNNW	0.9757	1.0175	1.0623	1.1302	1.2412
		RF	**0.9176**	**0.9556**	**0.9983**	**1.0498**	**1.1108**
	Codrna	MI	1.4882	1.4847	1.484	1.4845	1.4854
		DLR	1.0131	1.065	1.1741	1.2524	1.2847
		PMM	1.4069	1.4791	1.5467	1.6031	1.6889
		EM	**1.0110**	1.0534	1.0929	1.1371	1.1869
		KNNW	1.0128	**1.0325**	1.0748	1.1310	1.2182
		RF	1.1232	1.0861	**1.0691**	**1.0913**	**1.1352**
	Skin Segmentation	MI	0.5131	0.5134	0.5132	0.5132	0.5134
		DLR	0.3265	0.3489	0.3743	0.4018	0.4275
		PMM	0.5132	0.5213	0.5404	0.5628	0.5862
		EM	0.3249	0.3430	0.3613	**0.3812**	**0.4021**
		KNNW	0.3009	0.3675	0.4167	0.4545	0.4826
		RF	**0.2498**	**0.3058**	**0.3489**	0.3902	0.4666

Table A2 Average time for each imputation technique by dataset

Size	Datasets	Mean	DLR	PMM	EM	KNNW	RF
Small	Breast Cancer Coimbra	0.0018 s	0.1710 s	0.1918 s	0.0394 s	0.0494 s	0.8371 s
	BUPA liver disorders	0.0005 s	0.0821 s	0.0886 s	0.0133 s	0.1978 s	2.3963 s
	Pima	0.0007 s	0.1453 s	0.1644 s	0.0342 s	0.3480 s	5.8447 s
Medium	Banknote authentication	0.0005 s	0.0866 s	0.1383 s	0.0505 s	0.5540 s	8.9559 s
	svmguide1	0.0008 s	0.1562 s	0.1689 s	0.1047 s	1.3802 s	31.2413 s
	Rice	0.0014 s	0.3106 s	0.3133 s	0.1981 s	2.8634 s	2.1885 m
Large	Magic	0.0097 s	1.6095 s	1.8757 s	0.8617 s	2.2074 m	49.5032 m
	Codrna	0.0193 s	2.8353 s	3.5431 s	2.2860 s	38.7772 m	4.1977 h
	Skin Segmentation	0.0310 s	2.6444 s	3.7772 s	6.7211 s	2.9511 h	1.7525 h

Note: The unit time is measured in second (s), minute (m), and hour (h).

Table A3 Rank of imputation techniques for small-sized datasets

Imputation	10	20	30	40	50	Rank of method by mean
MI	4.3333	4.3333	4.6667	3.0000	2.0000	3.6667
DLR	3.0000	3.0000	3.6667	4.6667	3.6667	3.6000
PMM	6.0000	6.0000	6.0000	5.6667	5.3333	5.8000
EM	2.6667	1.3333	1.6667	1.6667	1.6667	**1.8000**
KNNW	3.3333	3.6667	2.6667	2.5000	5.0000	3.4333
RF	1.6667	2.6667	2.3333	2.3333	2.6667	2.3333
Kendalls Statistics			W = 0.8263, Chi-sq = 20.7, p = 0.0009			

Table A4 Rank of imputation techniques for medium-sized datasets

Imputation	10	20	30	40	50	Rank of method by mean
MI	6.0000	5.6667	5.6667	5.3333	5.3333	5.6000
DLR	4.3333	4.0000	4.0000	4.0000	3.6667	4.0000
PMM	4.3333	5.0000	5.3333	5.6667	5.6667	5.2000
EM	2.3333	2.3333	2.3333	1.6667	1.3333	2.0000
KNNW	2.0000	1.6667	1.6667	1.6667	2.0000	**1.8000**
RF	2.0000	2.0000	2.0000	2.6667	3.0000	2.3333
Kendalls Statistics			W = 0.8743, Chi-sq = 21.9, p = 0.0006			

Table A5 Rank of imputation techniques for large-sized datasets

Imputation	10	20	30	40	50	Rank of method by mean
MI	5.3333	5.3333	5.0000	4.6667	5.0000	5.0667
DLR	3.6667	3.3333	3.6667	4.0000	3.3333	3.6000
PMM	5.6667	5.6667	6.0000	6.0000	6.0000	5.8667
EM	2.3333	2.3333	2.6667	2.0000	1.6667	2.2000
KNNW	2.0000	2.3333	2.6667	3.0000	3.3333	2.6667
RF	2.0000	2.0000	1.0000	1.3333	1.6667	**1.6000**
Kendalls Statistics			W = 0.9257, Chi-sq = 23.1, p = 0.0003			

Table A6 Rank of imputation techniques based on average missing rates

Imputation	Small	Medium	Large	Rank of method by mean
MI	3.6667	5.6000	5.0667	4.7778
DLR	3.6000	4.0000	3.6000	3.7333
PMM	5.8000	5.2000	5.8667	5.6222
EM	1.8000	2.0000	2.2000	2.0000
KNNW	3.4333	1.8000	2.6667	2.6333
RF	2.3333	2.3333	1.6000	2.0889
Kendalls Statistics		W = 0.8730, Chi-sq = 13.1, p = 0.0225		