MJFAS
Malaysian Journal of Fundamental and Applied Sciences

Shantal *et al.* | Malaysian Journal of Fundamental and Applied Sciences, Vol. 19 (2023) 1052-1067

RESEARCH ARTICLE

Impact of Missing Data on Correlation Coefficient Values: Deletion and Imputation Methods for Data Preparation

Mohammed Shantal*, Zalinda Othman, Azuraliza Abu Bakar

Center for Artificial Intelligence Technology, Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor Darul Ehsan, Malaysia

Abstract The correlation coefficient is one of the essential statistical techniques used to discover relationships among variables. Various techniques can quantify correlation, such as Pearson's, Spearman's, and Kendall's correlation coefficients, depending on the data type. As with any use of data, missing data will impact the availability of data, reducing it and potentially affecting the results. Furthermore, the removal of missing-value data from the study when using complete case analysis or available case analysis may result in selection biases. In this paper, we investigate the impact of missing data on the correlation coefficient value by calculating the difference between the correlation coefficient of the original complete dataset and that of a dataset with missing data. Two deletion strategies (Listwise and Pairwise) and three imputation strategies (Mean, *k*-Nearest Neighbors (*k*-NN), and Expectation-Maximization) were used to prepare the data before calculating the correlation coefficient. Unique correlation coefficient values were created by converting unique values to a one-dimensional array, and RMSE metrics were used to evaluate the experiments. Eight UCI and Kaggle datasets with different sizes and numbers of attributes were used in this study. The experiment results demonstrate that the Pairwise strategy and *k*-NN give good results on the correlation coefficient, respectively, when the missing rate is moderate or less. Pairwise uses all the available values and discards only the missing values of the related attribute, while *k*-NN fills the missing values with new values that produce correlation coefficient values close to the actual values.

Keywords: Correlation Coefficient, Pearson's Correlation, Missing data, Mean Imputation, *k*-NN imputation, Expectation Maximization imputation.

Introduction

***For correspondence:**
p97711@siswa.ukm.edu.my

Received: 24 July 2023
Accepted: 7 Nov. 2023

Missing values are one of the prevalent issues in data analysis, as the results can be misleading if cases with missing values are consistently different from cases without missing values. Additionally, missing data may affect the precision of calculated statistics, as there is less information than initially expected [1, 2]. On the other hand, the correlation coefficient is a statistical method that provides information on the strength and direction of the relationship between two variables [3]. It is pivotal in many works, particularly in data mining [4], such as feature selection [5-7] and missing data imputation methods [8-11]. However, missing data can be an issue in finding the correlation coefficient, as complete pairs are needed to calculate it. The missing data rate is one factor that impacts the sample size, and the correlation structure, missing mechanism, entries distribution, and the percentage of values significantly impact the performance of imputation methods [12]. The missing rate is categorized based on its percentage, with less than 10% being described as minor missing, 10% to less than 25% as moderate, 25%-50% as high, and more than 50% as excessive [13]. The correlation coefficient is a measurement method used to explore the relationships among attributes. It is a bivariate study that measures the connection strength between two variables and the relationship direction [14]. Linear correlation, measured by Pearson's Correlation Coefficient

(PCC), determines whether two variables have a linear connection. PCC is proportional to covariance and can be interpreted in the same way. Its value fluctuates from +1 to -1 in terms of the strength of the relationship [3]. A correlation coefficient value close to one indicates a strong correlation, while a value approaching zero indicates a weak correlation. Additionally, the correlation coefficient shows the direction of the relationship, whether it is negative or positive, as illustrated in Figure 1 [15].

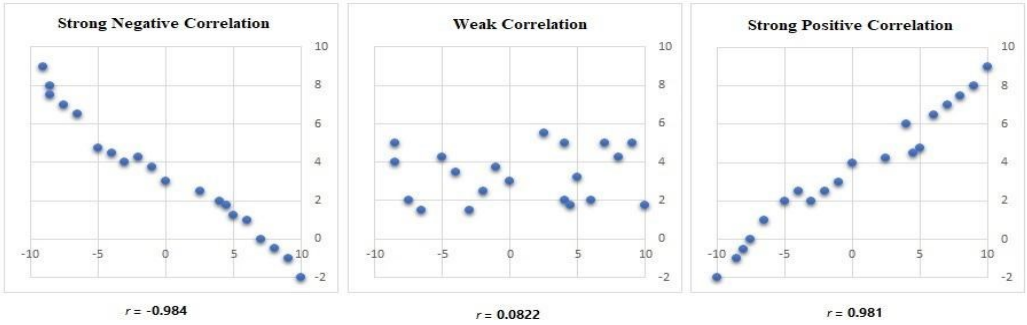

Figure 1. Type of correlation.

In contrast, the Spearman correlation can be used to measure a monotonic association for non-normally distributed continuous data, ordinal data, or data with relevant outliers. Both coefficients are scaled to vary from –1 to +1, with 0 indicating no linear or monotonous connection. This relationship strengthens and eventually approaches a straight line (Pearson correlation) or a constantly increasing or decreasing curve (Spearman correlation) as the coefficient approaches the total value of 1 [14]. Table 1 shows the type of data and the appropriate measurement metric for the relationship between them [16].

Table 1. Data type vs correlation measurement method

Variable Y	Variable X		
	Nominal	Ordinal	Continuous
Nominal	ϕ or λ	Rank biserial	Point biserial
Ordinal	Rank biserial	τb or Spearman	τb or Spearman
Continuous	Point biserial τ	τb or Spearman	Pearson or Spearman

ϕ = phi coefficient, λ = Goodman and Kruskal's lambda, τb = Kendall's τb

The measured correlation coefficient can be interpreted based on the strength categories. The degree of the relationship is categorized as negligible, weak, moderate, strong, and very strong correlation. As shown in Table 2, the correlation degree depends on the proximity of its value to one, and its weakness as it approaches zero [14].

Table 2. Correlation coefficient interpretation

The absolute magnitude of the observed correlation coefficient	Degree of correlation
0.00 - 0.10	Negligible
0.10 - 0.39	Weak
0.40 - 0.69	Moderate
0.70 - 0.89	Strong
0.90 - 1.00	Very strong

The point biserial correlation coefficient is a case where Pearson's correlation coefficient can be used to calculate the correlation between categorical and continuous variables. Specifically, the categorical variable is a dichotomous variable, meaning it has only two values that can be set as 0

or 1. The point biserial correlation coefficient can vary from -1 to +1, like the Pearson coefficient [17]. As with any data analysis, missing data is a critical issue that needs to be addressed. The calculation of the correlation coefficient is essential for exploring relationships among attributes and for many other methods, such as feature selection, regression, and imputation methods [18-20].

In many research studies, the correlation coefficient plays an important role in statistical analysis and various machine learning methods. Calculating the correlation coefficient is a simple step in complete data cases. However, the primary focus of this work is on addressing the issues related with missing data and the critical data preparation activities required prior to using it to calculate the correlation coefficient.

Materials and Methods

This section will provide a comprehensive account of the research methodology. It covers the handling of missing data, computation of correlation coefficients, experimental design, and the foundational steps that led to the results presented in the subsequent "Results and Discussion" section.

Dealing with Missing Values

In many areas of data analysis, missing data is one of the significant challenges for researchers. Neglecting or disregarding this missing data can result in significant biases in the findings [21]. Dealing with this issue has been the focus of much work. Many statistical and machine learning methods have been proposed to deal with the missing data problem to ensure that the maximum amount of data can be used to ensure that no misleading occurs in the results [22, 23]. There are two important strategies in dealing with missing data, namely, deletion and imputation strategy.

Deletion is a standard method used for handling missing values [24]. The main idea behind it is to discard incomplete information. There are two ways for deletion: Listwise deletion and Pairwise deletion [25]. Listwise deletion, also known as "case deletion," is a common way of dealing with missing values by deleting all cases/instances with one or more missing values and using only complete cases to analyze the data [26]. Meanwhile, the Pairwise strategy uses all observed information instead of deleting instances with missing values. Pairwise deletion eliminates missing data for each data point separately. All available information will be used to analyze further.

The imputation technique is one of the popular methods to deal with the missing values issue. Statistical and machine learning techniques are used to impute the missing values. In this study, Mean, *k*-Nearest Neighbors (*k*-NN), and Expectation-Maximization (EM) techniques will be used. Mean Imputation (MI) is the simplest way of dealing with missing values because it replaces the missing values with the mean of the observed values. The MI results are acceptable if the missing rate is minor; otherwise, it will result in bias [27, 28]. *k*-NN is one of the non-parametric methods in machine learning and is widely used to classify data due to its simplicity, generality, and relatively high accuracy. It has successfully been used for classification purposes such as data mining, machine learning, pattern recognition, text categorization, and object recognition [29-31]. The core of *k*-NN is finding *k* most similar instances in samples with a high probability of being in the same class [32]. In 1977, Dempster, *et al*. [33] proposed the expectation-maximization algorithm to solve the drawbacks of Maximum Likelihood approaches. The Expectation-Maximization Imputation (EMI) approach depends on estimating the dataset's mean and covariance matrices to impute missing numerical values. It begins with initial estimates of the mean and covariance matrix and iterates through the steps until the imputed values and the mean and covariance matrix estimates do not change significantly from one iteration to the next iteration [34, 35]. Mirzaei, *et al*. [36] summarize the impact of missing data on statistical analysis depends on the percentage of missing data. For less than 5% missing data, a simpler single imputation approach may be appropriate. For more than 10% missing data, there is a higher likelihood of bias, and multiple imputation techniques may be considered. When missing data exceeds 40%, imputation or likelihood methods may only generate hypothesis-generating results. The type of missing data (e.g., missing completely at random or not) should also be considered in deciding how to handle the missing observations. If missing data is completely at random, imputations or deletions can be done with minimal bias. Johnson and Young [37] suggest that both imputing missing data before analysis or handling missing data at each step of the analysis are acceptable strategies for handling missing values in family research. Little difference in substantive conclusions was found between the two approaches when using a commonly used dataset in family research, indicating flexibility in choosing the approach that best fits the research question and dataset. However, removing data in such cases may increase standard

error due to reduced sample size. Alternatively, imputation can be used to estimate the response that the respondent may have provided if the data was not missing. Also, it is important to ensure that the missing data model is as complete as the analysis model to obtain accurate results, and including auxiliary variables, especially those highly correlated with variables in the model, is recommended based on literature [38].

Correlation Coefficient with Missing Values

Musil, *et al*. [39] have been examines the impacts of different methods on descriptive statistics and correlations with other variables, utilizing a unique dataset for comparison. Listwise deletion, mean imputation, simple regression, regression with an error term and EM have been used to dealing with the missing values for MAR missing mechanism. The limitations of the study include low explanatory power of the variables used for imputation, which can affect the accuracy of parameter estimates, highlighting the importance of selecting predictors with substantial correlations with the missing variable for accurate imputations. Sefidian and Daneshpour [11] proposed methods for imputing missing values using correlation maximization-based techniques. They proposed ten correlation-based imputation methods that aim to maximize the correlation between a missing feature and other features. This is done by selecting data segments with strong correlations. The proposed methods involve three main steps: selecting a base set from all complete instances, generating data segments with strong correlations using the base set and the remaining complete instances, and imputing missing values using a linear model applied to the discovered data segments. However, using only the base set of complete instances may reduce the sample size, affecting the results of the correlation coefficient in earlier steps.

Liu, *et al*. [9] proposed a method to impute missing data by maximizing the correlation coefficient. Their work involves dividing the dataset into subsets, with each subset containing instances with the same number of missing data. The correlation coefficient is calculated in batches, with each sub-dataset starting with the subset with the fewest missing data and progressing to the most. The *k*-NN method is used to impute the missing values in the subset, with the correlation coefficient used to weigh the distances. The imputed subset is then added to the previous one to calculate the subsequent correlation. However, using only a small subset of the first subsets may affect the sample size and thus the results of the correlation coefficient.

ÜRESİN [40] proposed a method called Correlation Based Regression Imputation (CBRI), which uses simple regression-based imputation to impute missing values of each feature based on its correlation with others. Pearson's pairwise correlation matrix is used to determine the dependent and independent features, with only features with high correlation selected. While the results have shown improvement in imputation accuracy, the study used small datasets, and using Pearson pairwise correlation may not be ideal for large datasets with higher missing rates.

Nugroho, *et al*. [2] proposed the Class Center-based Imputation method, which considers the correlation among attributes to estimate missing values. The hybrid imputation stage uses both the firefly algorithm and the CCMVI algorithm. During the data preparation step, the data is divided into two subsets: incomplete data and complete data. Only the Listwise strategy is used to produce the correlation coefficient for use in the proposed method. While the study showed that using correlation as one of the parameters to estimate missing values improves the imputation method, using Listwise may dramatically reduce the available data, especially with a high missing rate. Using reduced data to calculate the correlation coefficient may not provide an accurate indicator of the relationship between attributes.

As correlation is a crucial factor in feature selection methods, obtaining an accurate correlation coefficient is essential in the process. However, missing data can affect the results, and most studies on feature selection assume that the data have no missing values [41]. Some researchers fill the missing values with the mean value in numeric features [42] or use a multiple regression model to impute the missing values and then use Pearson Correlation Coefficient to calculate the similarity among the attributes [43]. However, most previous studies do not consider the number of instances when calculating the correlation coefficient. Using a Listwise case to prepare data with all variable values available is still a problem, and getting the most out of all data is the researchers' biggest obstacle.

In the case of missing data, D'Angelo, *et al*. [44] extended the EM algorithm for the partial correlation and compared it to the Multi Imputation approach in a thorough simulation study. They used Pearson Correlation Coefficient to measure the correlation using full, complete cases, EM, and MI to evaluate their work. They discovered that the correlation values obtained with complete cases were most likely

deceptive. Furthermore, the EM exhibited the best statistical features of all the approaches. EM worked almost as well as MI. However, only EM and IM were used to estimate the missing values, and only one dataset was used in the study. Therefore, this study investigates the impact of missing values on the calculation of the correlation coefficient and suggests the best strategies to deal with this issue in various missing rates [44].

This work aims to study the impact of missing data on the correlation coefficient and how to deal with this issue. The main objective is to determine the best data preparation strategy (deletion and imputation strategy) without compromising the correlation coefficient among the attributes. The experimental design is described in Section 2, and Section 3 presents the experiment results of the various deletion and imputation methods. This paper contributes to the field of dealing with missing data by identifying the best strategies for data preparation and presenting the impact of different missing rates on the correlation coefficient values. The article concludes with a discussion.

Experimental Design

In this study, the work is divided into three phases, i.e., data preparation, handling missing values, and evaluation, as shown in Figure 2. It offers the experiment design conducted in this study.

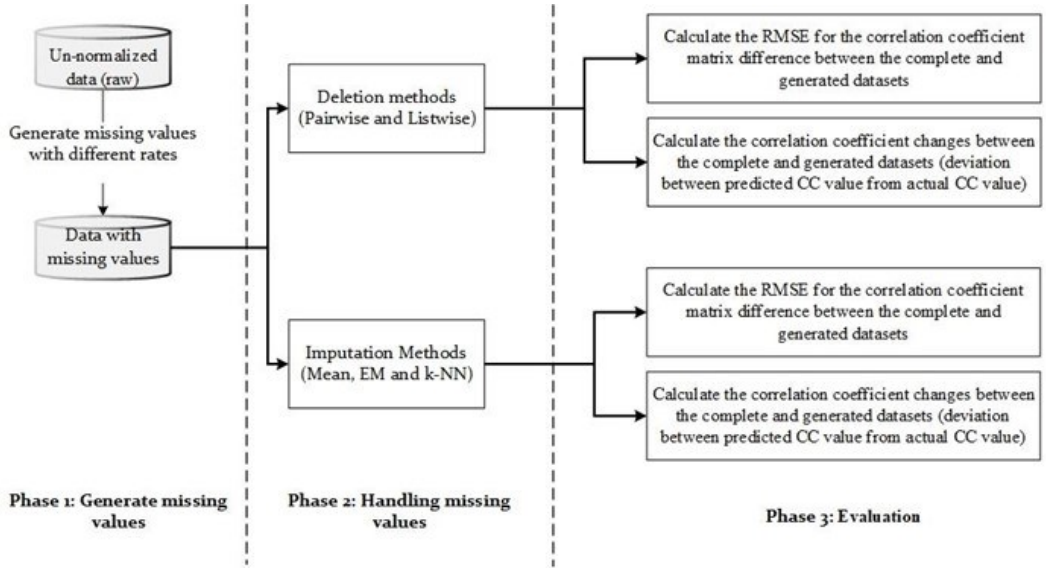

Figure 2. Experimental design

The experiment was initiated with a complete dataset comprising eight real-world datasets from UCI and Kaggle. Only numerical data, including discrete and continuous numbers, were included in the study, as shown in Table 3.

Table 3. List of datasets (label attribute is not counted in # of attributes)

Datasets	Type of Data	No of instances	No of attributes	No of classes	No of a unique correlation value
Blood	Integer	748	4	2	10
Breast cancer 1	real	569	30	2	465
Breast cancer 2	Integer	683	9	2	45
Parkinson	real	195	22	2	253
Ionosphere	real	351	34	2	595
QSAR	real	1055	41	2	861
Spam	real	4601	57	2	1653
Musk	Integer	6598	166	2	13861

In phase 1, the experiment was initiated with a complete dataset comprising eight real-world datasets from UCI and Kaggle. Only numerical data, including discrete and continuous numbers, were

included in the study, as shown in Table 3.

A dataset with missing values was generated using the Completely at Random missing (MCAR) type. Five types of datasets were used in the experiments, including:

i. Complete dataset – a complete dataset used to calculate the correlation matrix (CM) for evaluation purposes.
ii. Missing dataset – a dataset with missing values generated with various missing rates.
iii. Listwise dataset – a dataset with only complete instances values, with any instances containing at least one missing value discarded from the dataset.
iv. Pairwise dataset – a dataset that uses all available pairs of attributes to calculate the correlation coefficient.
v. Imputed dataset – a dataset generated using one of the imputation methods, i.e., Mean, *k*-NN, and Expectation-Maximization.

Binary class datasets were used as all datasets were numeric and normalized to (0,1) to give all attributes the same power in the dataset [45]. The experiment employed five different missing rates: 2%, 10%, 25%, 50%, and 80% [13]. In assessing the impact of the imputation method on the correlation coefficient value, the missing rate, in addition to the data preparation method for correlation analysis, is a factor. For missing data type, MCAR indicates no correlation between the missing mechanism and any attributes used [46]. Two types of missing models are there: Uniformly Distributed (UD) and overall. In UD, each feature has the same percentage number of missing values, whereas overall, the ratio of missing values in each variable has a different value than another. In this study, UD was used [47]. To implement the methods, codes were written in Python 3.7, Jupyter Notebook (Anaconda 3), and SciKitLearn Library. The reported results in this experiment are based on the average of 10 values obtained by repeating the experiment run. Each experiment was repeated 10 times to ensure reliability and reduce the impact of random fluctuations.

In phase 2, two strategies, i.e., deletion and imputation, are employed to handle missing data. In the deletion strategy, two approaches, namely Listwise and Pairwise, are being used. Meanwhile, *k*-NN, Mean, and Expectation Maximization are used for the imputation strategy. As a result, new datasets are created, which are used to evaluate the correlation coefficient between the complete dataset and the generated datasets.

The Pearson Correlation Coefficient (CC) is applied to calculate the correlation matrix (CM), as shown in Equation 1. Meanwhile, Point Biserial Correlation is used when only the class has two values.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (1)$$

Where r is a Pearson correlation coefficient between x and y, n is the number of observations, x_i is the value of x for i_{th} observation, and y_i is the value of y for i_{th} observation.

The assumption is that the dataset has m attributes, and the calculation of the CM produces an array of n x n. The unique correlation value (UCV) array will only include one unique correlation coefficient value between each pair of attributes. The CM upper-triangle values are converted into a one-dimensional array, and the correlation between any attributes (for example, Att1 to Att5) themselves will be ignored, as shown in Figure 3. The length of the UCV array is defined in Equation 2.

$$UCV\ Length = \frac{(n^2 - n)}{2} \quad (2)$$

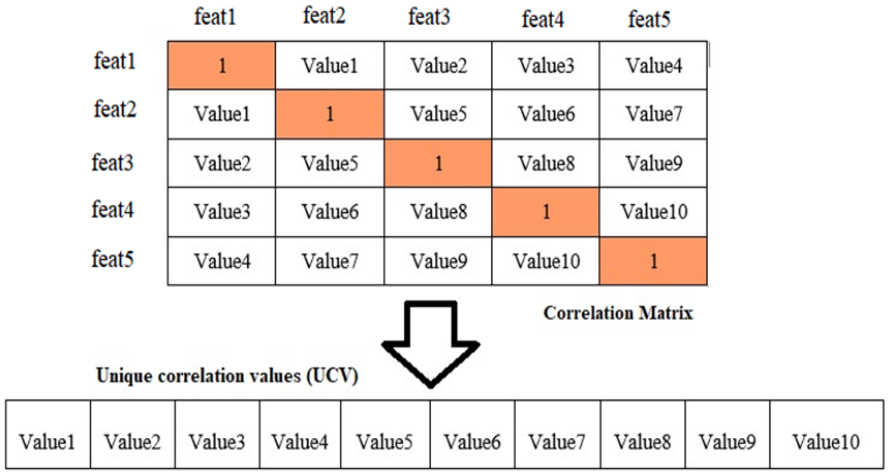

Figure 3. Unique correlation values (UCV) array

The impact of missing values on correlation coefficient values is studied in two different ways, i.e. the difference in correlation coefficient matrix between the complete dataset with the generated datasets as measured by Root Mean Square Error (RMSE), and the correlation coefficient changes between the complete dataset and the generated datasets (deviation between predicted CC value from actual CC value). Since the datasets used in the study consist of numeric data, most of the correlation coefficient (CC) values are calculated among numeric attributes. Imbalance issues within label attributes are not considered extensively, as they constitute a small proportion of the total correlation values.

RMSE is a standard statistical metric to determine the model's performance. In other words, it tells you how concentrated the data is around the line of best fit. It is commonly used in climatology, forecasting, and regression analysis to verify experimental results [48]. The equation that describes RMSE is as in Equation 3.

$$RMSE(x,\hat{x}) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \hat{x_i})^2}{n}} \quad (3)$$

$\sum_{i=1}^{n}(x_i - \hat{x_i})^2$: This part of the formula calculates the squared difference between each true value x_i and its corresponding predicted value $\hat{x}_i$, for all values in the dataset from i=1 to i=n. These squared differences are then summed up. Next, the sum of squared differences calculated in previous step is then divided by the total number of values in the dataset n. This division by n normalizes the squared differences to represent the average squared difference between the true and predicted values. Finally, the average squared difference obtained is square rooted. This is done to obtain the square root of the average squared difference, which gives us the root mean squared error (RMSE). Each coefficient correlation value change is calculated by taking the absolute difference between the actual and the predicted coefficient correlation. After calculating the absolute difference between the actual and predicted correlation coefficient values, these differences are categorized as small if they fall within ±0.05, moderately small if they fall within ±0.05-±0.10, moderately large if they fall within ±0.10-±0.15, large if they fall within ±0.15-±0.20, and very large if they exceed ±0.20. These categories provide information on the magnitude of the difference between predicted and actual correlation coefficients.

Results and Discussion

In general, missing data affect the number of complete instances, which are instances with no missing values. The number of complete instances is influenced by the missing rate and the number of attributes. Treating missing values at different levels of missing rates using different methods can demonstrate the ability to form complete instances.

Correlation Coefficient Analysis after Deletion Methods

In this section, a comparison between two deletion methods, Listwise and Pairwise, has been made. Figures 4a to 4d show the RMSE results of the deletion methods. Based on these figures, some Listwise results are not available due to the absence of complete instances in the datasets.

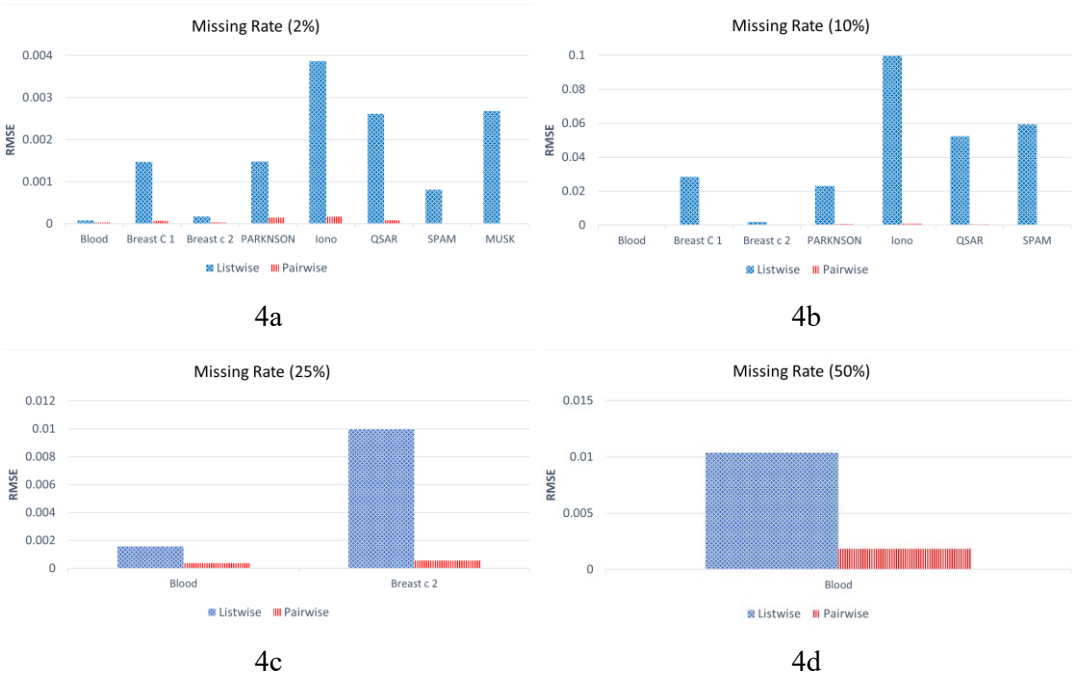

Figure 4 (a-d). Correlation coefficient RMSE of Listwise and Pairwise deletion method in different missing rate

Through Figures 4a to 4d, we can see the performance of Listwise and Pairwise, and based on RMSE, we observe that the error rates are high when using Listwise. While Pairwise outperforms Listwise, even though the RMSE increases as the missing rate increases. In addition, the number of complete instances decreases as the missing rate increases, as shown in Figures 4a to 4d. All datasets have complete instances with a 2% missing rate, as shown in Figure 4a. However, we begin to see no complete instances with the MUSK dataset at a 10% missing rate, as shown in Figure 4b. At a 25% missing rate, only two datasets, Blood and Breast Cancer 2, have complete instances due to their lower number of attributes. In Figure 4d, with a 50% missing rate, only the Blood dataset has complete instances. Finally, there are no complete instances at an 80% missing rate.

In deletion methods, the Pairwise strategy is the best way to deal with missing data as all available values are exploited, while the Listwise strategy loses much of its power because it deletes some data in instances with some missing values. For each dataset, Table 4 shows the percentage of CC values that deviate from the actual CC values for each deviation level. It shows that Listwise and Pairwise kept all CC values close to actual values when the missing rate was low (2%), while all predicted CC has deviated less than 0.05. At a 10% missing rate, Pairwise continues to keep all CC values in deviation less than 0.05. However, at the same missing rate, Listwise started producing some CC values that varied from actual values more than 0.20, as in the Ionosphere dataset, where 40% of CC values varied more than 0.20 points. Complete instances are unavailable in most cases starting at 25% of the missing rate. Only Blood and Breast Cancer 2 datasets are yielding results due to their lower number of attributes. At 25% missing rate, Pairwise still provides good results where most CC changes are less than 0.05 points. In a high missing rate of 50%, most of the predicted CC changes are less than 0.05, while some other changes are between 0.05 to 0.10 points. In the 80% missing rate, the differences between the predicted values and the actual values are distributed on many deviation levels, sometimes reaching more than 0.20. It can be seen in the Parkinson and Ionosphere datasets, where the percentage of values that deviated by more than 0.20 points is 23% and 26%, respectively. Using the pairwise strategy to deal with missing values and calculating the correlation coefficient with different missing rates gives CC values close to the actual CC values because it uses all available values.

Table 4. Percentage of CC value deviation level for deletion strategies

		Deletion methods									
Missing rate 2%		Listwise					Pairwise				
Dataset	# UCV	< ±0.05	±0.05 - ±0.10	±0.10 - ±0.15	±0.15 - ±0.20	>±0.20	< ±0.05	±0.05 - ±0.10	±0.10 - ±0.15	±0.15 - ±0.20	>±0.20
Blood	10	100%	0%	0%	0%	0%	100%	0%	0%	0%	0%
Breast C1	465	100%	0%	0%	0%	0%	100%	0%	0%	0%	0%
Breast C2	45	100%	0%	0%	0%	0%	100%	0%	0%	0%	0%
Parkinson	253	100%	0%	0%	0%	0%	100%	0%	0%	0%	0%
Ionosphere	595	93%	7%	0%	0%	0%	100%	0%	0%	0%	0%
QSAR	861	97%	2%	1%	0%	0%	100%	0%	0%	0%	0%
Spam	1653	99%	1%	0%	0%	0%	100%	0%	0%	0%	0%
Musk	13861	97%	3%	0%	0%	0%	100%	0%	0%	0%	0%
Missing rate 10%											
Blood	10	100%	0%	0%	0%	0%	100%	0%	0%	0%	0%
Breast C1	465	43%	30%	15%	9%	4%	100%	0%	0%	0%	0%
Breast C2	45	100%	0%	0%	0%	0%	100%	0%	0%	0%	0%
Parkinson	253	70%	18%	8%	4%	1%	100%	0%	0%	0%	0%
Ionosphere	595	19%	16%	14%	12%	40%	100%	0%	0%	0%	0%
QSAR	861	36%	23%	17%	10%	13%	99%	1%	0%	0%	0%
Spam	1653	35%	25%	16%	11%	13%	100%	0%	0%	0%	0%
Musk	13861	-	-	-	-	-	100%	0%	0%	0%	0%
Missing rate 25%											
Blood	10	100%	0%	0%	0%	0%	100%	0%	0%	0%	0%
Breast C1	465	-	-	-	-	-	100%	0%	0%	0%	0%
Breast C2	45	89%	9%	2%	0%	0%	100%	0%	0%	0%	0%
Parkinson	253	-	-	-	-	-	97%	2%	0%	0%	0%
Ionosphere	595	-	-	-	-	-	99%	1%	0%	0%	0%
QSAR	861	-	-	-	-	-	98%	1%	0%	0%	0%
Spam	1653	-	-	-	-	-	100%	0%	0%	0%	0%
Musk	13861	-	-	-	-	-	100%	0%	0%	0%	0%
Missing rate 50%											
Blood	10	90%	10%	0%	0%	0%	100%	0%	0%	0%	0%
Breast C1	465	-	-	-	-	-	92%	7%	0%	0%	0%
Breast C2	45	-	-	-	-	-	100%	0%	0%	0%	0%
Parkinson	253	-	-	-	-	-	80%	16%	3%	1%	0%
Ionosphere	595	-	-	-	-	-	75%	23%	2%	0%	0%
QSAR	861	-	-	-	-	-	93%	5%	1%	0%	0%
Spam	1653	-	-	-	-	-	98%	2%	0%	0%	0%
Musk	13861	-	-	-	-	-	100%	0%	0%	0%	0%
Missing rate 80%											
Blood	10	-	-	-	-	-	90%	0%	10%	0%	0%
Breast C1	465	-	-	-	-	-	54%	26%	12%	5%	3%
Breast C2	45	-	-	-	-	-	82%	11%	7%	0%	0%
Parkinson	253	-	-	-	-	-	40%	16%	11%	10%	23%
Ionosphere	595	-	-	-	-	-	24%	19%	18%	13%	26%
QSAR	861	-	-	-	-	-	55%	28%	11%	3%	3%
Spam	1653	-	-	-	-	-	85%	12%	2%	0%	0%
Musk	13861	-	-	-	-	-	98%	2%	0%	0%	0%

Correlation Coefficient Analysis after Imputation Methods

Using one of the imputation methods is the best way to keep all instances where all missing values will be filled. Three imputation methods were used to prepare the data with missing values before calculating the CC.

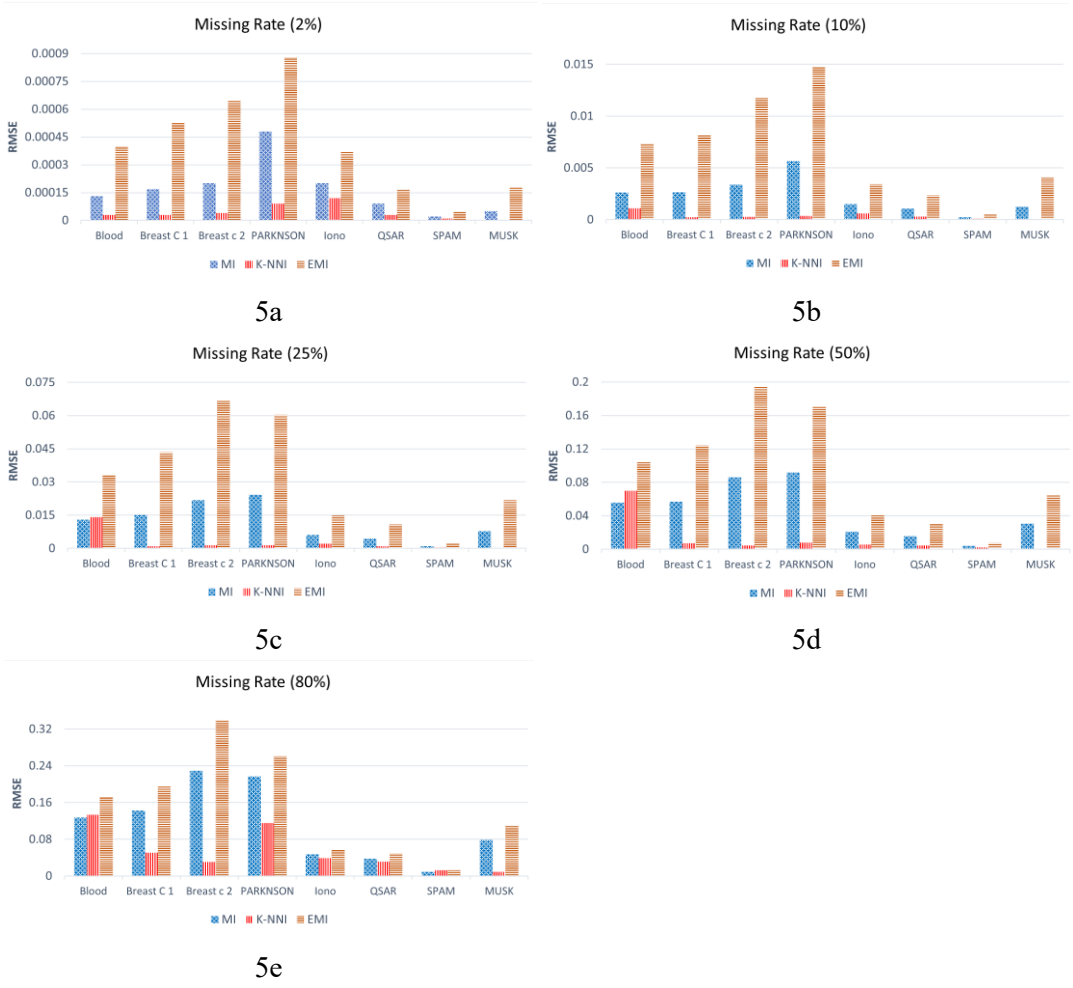

Figure 5 (a-e). Correlation coefficient RMSE of imputation methods in different missing rate

As shown in Figure 5a, all imputation methods perform well with a missing rate of 2%, where all RMSE values are less than 0.001. Although the RMSE values are low, *k*-NNI gives the lowest error rate followed by the MI method. With a missing rate of 10%, as seen in Figure 5b, all methods keep the RMSE rate under 0.02, but EMI shows a higher error rate compared to the other methods. When the missing rate is 25%, as in Figure 5c, EMI produces the worst results except for the Breast Cancer 2 dataset, where *k*-NNI has a higher RMSE rate. Additionally, in the Blood dataset, MI gives the lowest error rate. Overall, raising the missing rate increases the RMSE values, as shown in Figures 5a to 5d. The performance of the methods, from best to worst, is listed as *k*-NNI, MI, and EMI.

Table 5 shows that all methods keep all CC values close to the actual values when the missing rate is low (2%), with all predicted CC values deviating less than 0.05. At a missing rate of 10%, CC values calculated from the *k*-NNI imputed dataset keep most CC values in the low deviation level. In contrast, in the MI dataset, values are distributed between levels 0.05 and 0.05-0.10, meaning the maximum deviation is less than 0.10 points. The EMI dataset produced the worst CC values, with differences in some cases reaching 0.15 from the actual values. At a missing rate of 25%, the *k*-NNI imputed dataset is still better, with most values around 90% of CC values deviating less than 0.05. The EMI imputed dataset produced most CC values with a difference of more than 0.20 when the missing rate was high (50%). In the case of an 80% missing rate, all methods give some values with a high deviation from the actual CC values, with most values having more than 0.20 difference. However, *k*-NNI is the best in all imputation method results, where it produces fewer values with high deviation in most cases.

Table 5. Percentage of CC value deviation level for imputation methods

		Imputation methods														
Missing rate 2%		MI					k-NNI					EMI				
	# UCV	< ±0.05	±0.05 - ±0.10	±0.10 - ±0.15	±0.15 - ±0.20	>±0.20	< ±0.05	±0.05 - ±0.10	±0.10 - ±0.15	±0.15 - ±0.20	>±0.20	< ±0.05	±0.05 - ±0.10	±0.10 - ±0.15	±0.15 - ±0.20	>±0.20
Blood	10	100%	0%	0%	0%	0%	100%	0%	0%	0%	0%	100%	0%	0%	0%	0%
Breast C1	465	100%	0%	0%	0%	0%	100%	0%	0%	0%	0%	100%	0%	0%	0%	0%
Breast C2	45	100%	0%	0%	0%	0%	100%	0%	0%	0%	0%	100%	0%	0%	0%	0%
Parkinson	253	100%	0%	0%	0%	0%	100%	0%	0%	0%	0%	100%	0%	0%	0%	0%
Ionosphere	595	100%	0%	0%	0%	0%	100%	0%	0%	0%	0%	100%	0%	0%	0%	0%
QSAR	861	100%	0%	0%	0%	0%	100%	0%	0%	0%	0%	100%	0%	0%	0%	0%
Spam	1653	100%	0%	0%	0%	0%	100%	0%	0%	0%	0%	100%	0%	0%	0%	0%
Musk	13861	100%	0%	0%	0%	0%	100%	0%	0%	0%	0%	100%	0%	0%	0%	0%
Missing rate 10%																
Blood	10	70%	20%	10%	0%	0%	90%	10%	0%	0%	0%	70%	0%	20%	10%	0%
Breast C1	465	68%	31%	1%	0%	0%	100%	0%	0%	0%	0%	41%	33%	20%	6%	0%
Breast C2	45	40%	60%	0%	0%	0%	100%	0%	0%	0%	0%	2%	44%	51%	2%	0%
Parkinson	253	51%	41%	8%	0%	0%	100%	0%	0%	0%	0%	33%	23%	23%	18%	3%
Ionosphere	595	90%	10%	0%	0%	0%	100%	0%	0%	0%	0%	69%	25%	6%	0%	0%
QSAR	861	95%	5%	0%	0%	0%	99%	1%	0%	0%	0%	82%	14%	3%	1%	0%
Spam	1653	98%	2%	0%	0%	0%	100%	0%	0%	0%	0%	96%	2%	1%	0%	0%
Musk	13861	84%	16%	0%	0%	0%	100%	0%	0%	0%	0%	62%	25%	10%	3%	0%
Missing rate 25%																
Blood	10	70%	0%	10%	10%	10%	70%	0%	10%	10%	10%	10%	60%	0%	0%	0%
Breast C1	465	28%	30%	19%	14%	9%	96%	4%	0%	0%	0%	17%	19%	17%	0%	15%
Breast C2	45	0%	13%	44%	40%	2%	98%	2%	0%	0%	0%	0%	2%	7%	0%	27%
Parkinson	253	26%	19%	15%	23%	16%	95%	5%	0%	0%	0%	17%	14%	10%	0%	11%
Ionosphere	595	58%	26%	12%	4%	0%	88%	11%	1%	0%	0%	38%	26%	15%	0%	11%
QSAR	861	69%	20%	7%	2%	2%	99%	1%	0%	0%	0%	56%	19%	12%	0%	6%
Spam	1653	95%	3%	2%	0%	0%	99%	1%	0%	0%	0%	91%	5%	2%	0%	1%
Musk	13861	50%	27%	13%	7%	4%	100%	0%	0%	0%	0%	33%	23%	0%	16%	10%
Missing rate 50%																
Blood	10	20%	40%	10%	0%	30%	20%	30%	10%	10%	30%	10%	10%	50%	0%	30%
Breast C1	465	15%	14%	14%	15%	42%	49%	38%	10%	2%	1%	12%	9%	11%	8%	60%
Breast C2	45	0%	0%	2%	11%	87%	80%	20%	0%	0%	0%	0%	0%	0%	2%	98%
Parkinson	253	13%	14%	7%	10%	55%	52%	35%	12%	1%	0%	13%	8%	7%	5%	66%
Ionosphere	595	37%	22%	14%	12%	16%	62%	32%	6%	0%	0%	31%	15%	15%	10%	30%
QSAR	861	52%	18%	13%	7%	11%	78%	14%	4%	3%	1%	44%	18%	10%	10%	19%
Spam	1653	87%	8%	1%	2%	2%	92%	4%	1%	1%	1%	79%	12%	3%	1%	4%
Musk	13861	28%	21%	16%	11%	24%	98%	2%	0%	0%	0%	19%	18%	13%	11%	38%
Missing rate 80%																
Blood	10	10%	0%	60%	0%	30%	10%	30%	10%	10%	40%	10%	0%	20%	30%	40%
Breast C1	465	10%	9%	8%	10%	63%	16%	20%	18%	16%	30%	11%	8%	7%	8%	66%
Breast C2	45	0%	0%	0%	0%	100%	40%	40%	7%	9%	4%	0%	0%	0%	0%	100%
Parkinson	253	12%	9%	7%	3%	69%	7%	6%	16%	15%	55%	12%	9%	4%	2%	72%
Ionosphere	595	30%	15%	14%	8%	31%	30%	21%	16%	10%	22%	28%	14%	13%	10%	35%
QSAR	861	41%	17%	10%	9%	23%	39%	23%	15%	9%	14%	42%	15%	8%	7%	28%
Spam	1653	77%	13%	4%	2%	4%	73%	14%	7%	2%	4%	74%	13%	5%	2%	5%
Musk	13861	18%	16%	13%	11%	42%	60%	18%	10%	6%	5%	18%	14%	11%	10%	48%

Comparison of between Correlation Values among All Pairs of Attributes of the Complete and Generated Dataset

Figure 6 displays the impact of missing data on CC values and the effectiveness of deletion and imputation methods in data preparation. The deviation between the predicted CC of a prepared dataset and the actual CC reflects the effect of missing data on the dataset. The diagonal line in the graph represents the actual CC of the complete dataset, and each point represents the predicted CC. The closer the points are to the diagonal line, the better the CC. As only two datasets (Blood and Breast Cancer 2) with a limited number of attributes were used as examples, the findings are

not generalizable to all datasets. In the Blood dataset, Listwise showed some points far from the actual points, indicating high deviation, particularly in the case of 50% missing data. However, cases with less than 50% missing data still had acceptable CC values as they were closer to the diagonal line. Pairwise performed better in all cases of missing rates compared to Listwise. For the Breast Cancer 2 dataset, only the results for the missing rates of 2%, 10%, and 25% were calculated using the Listwise method. In some cases with 25% missing data, the computed CC significantly deviated from the actual values.

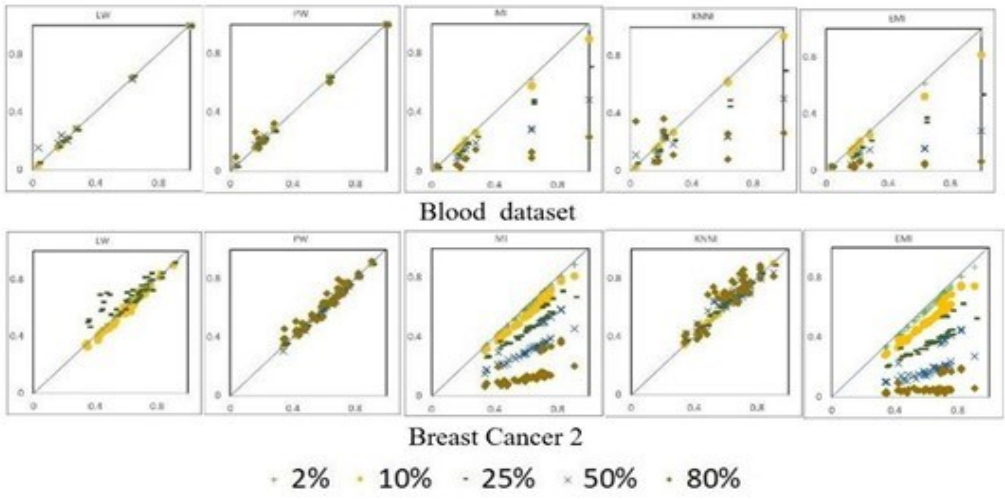

Figure 6. Comparison of CC values of complete and generated dataset**.** Note: Each dot represents a CC between two attributes. The diagonal line represents the actual CC of the complete dataset

Meanwhile, the distribution of CC calculated using imputed datasets by MI, *k*-NNI, and EMI varies in various cases of missing data. In the 80% missing rate, most points are in a high deviation range, which means raising the missing rates raises the deviation of the CC values. In the case of the Breast Cancer 2 dataset, it is apparent that the *k*-NNI method is the best method to deal with missing data since most of the predicted CC values remain close to the actual values even with high missing rates. This means that the imputed values kept the strengths of the relationships between the attributes close to the strengths of the real relationships.

Choosing the preparation method to prepare the data with missing data is a crucial step in considering the mining method. In Table 6, there are three comparisons: the first is between deletion methods (the lowest RMSE values are bolded). The second is between imputation methods (the lowest RMSE values are bolded). The third is the lowest RMSE among all methods (* indicates the lowest RMSE in all methods). As shown in Table 6, the Pairwise technique obtained the lowest RMSE values among deletion methods while *k*-NNI predicted the closest result to actual among imputation methods. Among all methods, the best result was obtained by both Pairwise and *k*-NNI, where filled values can be adopted as initial values if we need them in the further process that follows the data pre-processing stage.

Table 6. RMSE with various missing rates and various methods

Datasets\Method	Missing Rate	# Complete instance	RMSE (Standard Deviation) of the entire dataset				
			Listwise	Pairwise	MI	k-NNI	EMI
Blood	02%	688 (8)	0.00008(0.000)	**0.00003(0.000)***	0.00013(0.000)	**0.00003(0.000)***	0.00040(0.000)
	10%	484 (10)	0.00039(0.000)	**0.00014(0.000)***	0.00261(0.002)	**0.00111(0.001)**	0.00733(0.003)
	25%	237 (11)	0.00157(0.001)	**0.00037(0.000)***	**0.01289(0.006)**	0.01410(0.006)	0.03325(0.008)
	50%	45 (9)	0.01038(0.005)	**0.00185(0.001)***	**0.05559(0.015)**	0.06997(0.028)	0.10424(0.019)
	80%	1 (1)	-	0.00862(0.004)*	**0.12691(0.019)**	0.13347(0.030)	0.17224(0.011)
Breast Cancer 1	02%	305 (8)	0.00147(0.001)	**0.00007(0.000)**	0.00017(0.000)	**0.00003(0.000)***	0.00053(0.000)
	10%	24 (4)	0.02853(0.011)	**0.00038(0.000)**	0.00265(0.000)	**0.00024(0.000)***	0.00821(0.001)
	25%	0 (0)	-	0.00127(0.000)	0.01508(0.001)	**0.00101(0.000)***	0.04336(0.002)
	50%	0 (0)	-	0.00475(0.000)*	0.05690(0.001)	**0.00714(0.001)**	0.12408(0.003)
	80%	0 (0)	-	0.02816(0.002)*	0.14266(0.003)	**0.05089(0.007)**	0.19564(0.003)

Datasets\Method	Missing Rate	# Complete instance	RMSE (Standard Deviation) of the entire dataset				
			Listwise	Pairwise	MI	k-NNI	EMI
Breast cancer 2	02%	563 (11)	0.00017(0.000)	**0.00003(0.000)***	0.00020(0.000)	**0.00004(0.000)**	0.00065(0.000)
	10%	259 (8)	0.00183(0.001)	**0.00019(0.000)***	0.00338(0.001)	**0.00028(0.000)**	0.01183(0.001)
	25%	51 (8)	0.00995(0.006)	**0.00057(0.000)***	0.02177(0.002)	**0.00134(0.001)**	0.06661(0.007)
	50%	1 (1)	-	**0.00199(0.000)***	0.08601(0.005)	**0.00445(0.002)**	0.19466(0.011)
	80%	0 (0)	-	**0.01897(0.003)***	0.22905(0.010)	**0.03078(0.015)**	0.33903(0.007)
Parkinson	02%	125 (7)	0.00148(0.001)	**0.00015(0.000)**	0.00048(0.000)	**0.00009(0.000)***	0.00088(0.000)
	10%	21 (4)	0.02301(0.014)	**0.00078(0.000)**	0.00568(0.002)	**0.00037(0.000)***	0.01470(0.003)
	25%	0 (1)	-	0.00243(0.000)	0.02425(0.005)	**0.00145(0.000)***	0.06032(0.011)
	50%	0 (0)	-	0.00937(0.002)	0.09197(0.015)	**0.00799(0.002)***	0.17131(0.014)
	80%	0 (0)	-	0.07517(0.012)*	0.21677(0.008)	**0.11529(0.014)**	0.26033(0.004)
Iono (ionosphere)	02%	174 (10)	0.00387(0.001)	**0.00017(0.000)**	0.00020(0.000)	**0.00012(0.000)***	0.00037(0.000)
	10%	9 (3)	0.09983(0.031)	**0.00086(0.000)**	0.00150(0.000)	**0.00062(0.000)***	0.00343(0.000)
	25%	0 (0)	-	0.00272(0.000)	0.00609(0.000)	**0.00215(0.000)***	0.01482(0.001)
	50%	0 (0)	-	0.00998(0.001)	0.02112(0.001)	**0.00556(0.001)***	0.04097(0.002)
	80%	0 (0)	-	0.06912(0.007)	0.04763(0.001)	**0.03888(0.003)***	0.05857(0.001)
QSAR	2%	459 (18)	0.00261(0.001)	**0.00009(0.000)**	0.00009(0.000)	**0.00003(0.000)***	0.00017(0.000)
	10%	16 (6)	0.05232(0.015)	**0.00050(0.000)**	0.00106(0.000)	**0.00030(0.000)***	0.00235(0.000)
	25%	0 (0)	-	0.00155(0.000)	0.00449(0.000)	**0.00104(0.001)***	0.01076(0.000)
	50%	0 (0)	-	0.00420(0.000)*	0.01562(0.001)	**0.00463(0.000)**	0.03102(0.001)
	80%	0 (0)	-	0.02428(0.002)*	0.03787(0.001)	**0.03149(0.002)**	0.04947(0.000)
Spam	02%	1440 (37)	0.00081(0.000)	**0.00001(0.000)***	0.00002(0.000)	**0.00001(0.000)***	0.00005(0.000)
	10%	12 (5)	0.05940(0.016)	**0.00008(0.000)**	0.00023(0.000)	**0.00007(0.000)***	0.00055(0.000)
	25%	0 (0)	-	0.00035(0.000)*	0.00096(0.000)	**0.00036(0.000)**	0.00233(0.000)
	50%	0 (0)	-	0.00112(0.000)*	0.00394(0.000)	**0.00213(0.000)**	0.00737(0.000)
	80%	0 (0)	-	0.00634(0.001)*	0.00945(0.000)	0.01265(0.001)	0.01232(0.000)
Musk	02%	238 (21)	0.00268(0.001)	**0.00000(0.000)***	0.00005(0.000)	**0.00000(0.000)***	0.00018(0.000)
	10%	0 (0)	-	0.00003(0.000)	0.00125(0.000)	**0.00001(0.000)***	0.00406(0.000)
	25%	0 (0)	-	0.00008(0.000)	0.00774(0.000)	**0.00007(0.000)***	0.02178(0.000)
	50%	0 (0)	-	0.00032(0.000)*	0.03073(0.000)	**0.00047(0.000)**	0.06584(0.000)
	80%	0 (0)	-	0.00240(0.000)*	0.07835(0.000)	**0.00889(0.000)**	0.10945(0.000)

Conclusions

This paper demonstrates how to deal with missing data when calculating the correlation coefficient value. Several experiments were conducted on eight datasets of various sizes and attributes from the UCI and Kaggle repositories. Although Pairwise and *k*-NNI methods showed better results than other methods, all methods produced significant results for a small missing rate (2%). However, increasing the missing rate negatively affects the data by reducing the number of complete instances in the Listwise strategy and reducing the number of complete pairs in the Pairwise strategy. The results show that using Pairwise gives good results in most cases. Using all available pairs to calculate the CC of data with missing values or filling in missing values using the *k*-NNI method gives a more accurate CC. The Listwise strategy is the worst case for calculating CC with missing values since much data will be lost, particularly for higher missing rates.

Calculating the CC is crucial in data analysis research to recognize the relationship among attributes. However, missing data is a common challenge when using data for analysis purposes. This research presents an effective data preparation strategy to calculate the correlation coefficient when the data has missing values. Selecting the preparation method depends on the missing rate and data size, and if the research plan requires complete data, filling in the missing values becomes a necessary step. The choice of the preparation method still depends on the pre-processing strategy followed by the researcher. Pairwise deletion strategies yielded the best results. However, if the researcher's goal is to increase the sample size and impute missing values, k-NN will provide a significant result with the imputed dataset.

Conflicts of Interest

The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper.

Acknowledgment

We would like to thank Data Mining and Optimization Lab of Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia for the expert knowledge sharing. Thanks to Universiti Kebangsaan Malaysia and the Ministry of Higher Education for funding this research under Fundamental Research Grant Scheme (FRGS), FRGS/1/2021/ICT06/UKM/02/1.

References

[1] W.-C. Lin and C.-F. Tsai. (2020).Missing value imputation: A review and analysis of the literature (2006–2017). *Artificial Intelligence Review, 53*(2), 1487-1509.

[2] H. Nugroho, N. P. Utama, and K. Surendro. (2021). Class center-based firefly algorithm for handling missing data. *Journal of Big Data, 8*(1), 1-14.

[3] B. Ratner.(2009). The correlation coefficient: Its values range between +1/−1, or do they? *Journal of Targeting, Measurement and Analysis for Marketing, 17*(2), 139-142. Doi: 10.1057/jt.2009.5.

[4] I. Swesi and A. Abu Bakar. (2019). Feature clustering for PSO-based feature construction on high-dimensional data. *Journal of Information and Communication Technology, 18*. Doi: 10.32890/jict2019.18.4.3.

[5] M. A. Hall. (2000). Correlation-based feature selection of discrete and numeric class machine learning. Hamilton, New Zealand: University of Waikato, Department of Computer Science.

[6] H.-H. Hsu and C.-W. Hsieh. (2010). Feature selection via correlation coefficient clustering. *J. Softw., 5*(12), 1371-1377.

[7] R. Saidi, W. Bouaguel, and N. Essoussi. (2019). Hybrid feature selection method based on the genetic algorithm and pearson correlation coefficient. *Machine Learning Paradigms: Theory and Application*, A. E. Hassanien Ed. Cham: Springer International Publishing, 3-24.

[8] X. Chen, Z. Wei, Z. Li, J. Liang, Y. Cai, and B. Zhang. (2017). Ensemble correlation-based low-rank matrix completion with applications to traffic data imputation. *Knowledge-based Systems, 132*, 249-262.

[9] X. Liu, X. Lai, and L. Zhang. (2019). A hierarchical missing value imputation method by correlation-based K-nearest neighbors.*Proceedings of SAI Intelligent Systems Conference*, 486-496.

[10] G. Rahman and Z. Islam. (2011). A decision tree-based missing value imputation technique for data pre-processing. *Proceedings of the Ninth Australasian Data Mining*, *121*, 41-50.

[11] A. M. Sefidian and N. Daneshpour. (2020). Estimating missing data using novel correlation maximization based methods. *Applied Soft Computing, 91*, 106249. Doi: https://doi.org/10.1016/j.asoc.2020.106249.

[12] R. Armina, A. Mohd Zain, N. A. Ali, and R. Sallehuddin. (2017). A review on missing value estimation using imputation algorithm. *Journal of Physics: Conference Series, 892*, 012004. Doi: 10.1088/1742-6596/892/1/012004.

[13] K. F. Widaman. (2006). Missing data: what to do with or without them. *Monographs of the Society for Research in Child Development, 71*(3), 42-64. Doi: 10.1111/j.1540-5834.2006.00404.x.

[14] P. Schober, C. Boer, and L. A. Schwarte. (2018). Correlation coefficients: appropriate use and interpretation. *Anesthesia & Analgesia, 126*(5), 1763-1768.

[15] M. Baak, R. Koopman, H. Snoek, and S. Klous. (2020). A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics. *Computational Statistics & Data Analysis, 152*. Doi: 10.1016/j.csda.2020.107043.

[16] H. Khamis. (2008). Measures of association: how to choose? *Journal of Diagnostic Medical Sonography, 24*(3), 155-162.

[17] D. Kornbrot. (2014). Point biserial correlation. *Wiley StatsRef: Statistics Reference Online.*

[18] C. Arunkumar and S. Ramakrishnan. (2016). A hybrid approach to feature selection using correlation coefficient and fuzzy rough quick reduct algorithm applied to cancer microarray data. *2016 10th International Conference on Intelligent Systems and Control (ISCO)*. 1-6. Doi: 10.1109/ISCO.2016.7726921.

[19] A. Alhroob, W. Alzyadat, I. Almukahel, and H. Altarawneh. (2020). Missing data prediction using correlation genetic algorithm and SVM approach. *Population, 11*(2).

[20] S. Plancade, M. Berland, M. B. Nicolas, O. Langella, A. Bassignani, and C. Juste. (2021). A combined test for feature selection on sparse metaproteomics data-alternative to missing value imputation. *bioRxiv.*

[21] J. M. Brick and G. Kalton. (1996). Handling missing data in survey research. *Statistical methods in medical research, 5*(3), 215-238.

[22] O. Rado, M. Al Fanah, and E. Taktek. (2019). Performance analysis of missing values imputation methods using machine learning techniques. *Intelligent Computing-Proceedings of the*

Computing Conference, 738-750.
[23] P. S. Raja and K. Thangavel. (2020). Missing value imputation using unsupervised machine learning techniques. *Soft Computing, 24*(6), 4361-4392. Doi: 10.1007/s00500-019-04199-6.
[24] J. T. McCoy, S. Kroon, and L. Auret. (2018). Variational autoencoders for missing data imputation with application to a simulated milling circuit. *IFAC-PapersOnLine, 51*(21), 141-146. Doi: https://doi.org/10.1016/j.ifacol.2018.09.406.
[25] J. R. Cheema. (2014). A review of missing data handling methods in education research. *Review of Educational Research, 84*(4), 487-508.
[26] S. Yenduri and S. S. Iyengar. (2007). Performance evaluation of imputation methods for incomplete datasets. *International Journal of Software Engineering and Knowledge Engineering, 7*(01), 127-152.
[27] A. T. S. Dhevi. (2014). Imputing missing values using Inverse distance weighted interpolation for time series data. *2014 Sixth International Conference on Advanced Computing (ICoAC)*, 255-259. Doi: 10.1109/ICoAC.2014.7229721.
[28] I. Eekhout, H. C. de Vet, J. W. Twisk, J. P. Brand, M. R. de Boer, and M. W. Heymans. (2014). Missing data in a multi-item instrument were best handled by multiple imputation at the item score level. *Journal of Clinical Epidemiology, 67*(3), 335-342.
[29] N. García-Pedrajas, J. A. R. d. Castillo, and G. Cerruela-García. (2017). A proposal for local k values for k-nearest neighbor rule. *IEEE Transactions on Neural Networks and Learning Systems, 28*(2), 470-475. Doi: 10.1109/TNNLS.2015.2506821.
[30] A. B. Hassanat, M. A. Abbadi, G. A. Altarawneh, and A. A. Alhasanat. (2014). Solving the problem of the K parameter in the KNN classifier using an ensemble learning approach. *arXiv preprint arXiv:1409.0919*. Doi: 10.48550/ARXIV.1409.0919.
[31] S. Zhang. (2012). Nearest neighbor selection for iteratively kNN imputation. *Journal of Systems and Software, 85*(11), 2541-2552. Doi: https://doi.org/10.1016/j.jss.2012.05.073.
[32] J. Chen and J. Shao. (2000). Nearest neighbor imputation for survey data. *Journal of Official Statistics, 16*, 113-131.
[33] A. P. Dempster, N. M. Laird, and D. B. Rubin. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological), 39*(1), 1-22.
[34] W. Jiang, J. Josse, M. Lavielle, and G. TraumaBase. (2020). Logistic regression with missing covariates-parameter estimation, model selection and prediction within a joint-modeling framework. *Computational Statistics & Data Analysis, 145*, Art no. 106907. Doi: 10.1016/j.csda.2019.106907.
[35] T. Schneider. (2001). Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *Journal of Climate, 14*(5), 853-871.
[36] A. Mirzaei, S. R. Carter, A. E. Patanwala, and C. R. Schneider. (2022). Missing data in surveys: Key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy, 18*(2), 2308-2316. Doi: https://doi.org/10.1016/j.sapharm.2021.03.009.
[37] D. R. Johnson and R. Young. (2011). Toward best practices in analyzing datasets with missing data: Comparisons and recommendations. *Journal of Marriage and Family, 73*(5), 926-945. Doi: https://doi.org/10.1111/j.1741-3737.2011.00861.x.
[38] L. M. Collins, J. L. Schafer, and C.-M. Kam. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological methods, 6*(4), 330.
[39] C. M. Musil, C. B. Warner, P. K. Yobas, and S. L. Jones. (2002). A comparison of imputation techniques for handling missing data. *Western Journal of Nursing Research, 24*(7), 815-829.
[40] U. ÜRESİN. (2021). Correlation based regression imputation (CBRI) method for missing data imputation. *Turkish Journal of Science and Technology, 16*(1), 39-46.
[41] A. Bommert, X. Sun, B. Bischl, J. Rahnenführer, and M. Lang. (2020). Benchmark for filter methods for feature selection in high-dimensional classification data. *Computational Statistics & Data Analysis, 143,* 106839. Doi: https://doi.org/10.1016/j.csda.2019.106839.
[42] S. Egea, A. R. Mañez, B. Carro, A. Sánchez-Esguevillas, and J. Lloret. (2018). Intelligent IoT traffic classification using novel search strategy for fast-based-correlation feature selection in industrial environments. *IEEE Internet of Things Journal, 5*(3), 1616-1624. Doi: 10.1109/JIOT.2017.2787959.
[43] S. Rakshit, P. Das, and A. K. Das. (2018). Importance of Missing value estimation in feature selection for crime analysis. *Intelligent Communication and Computational Technologies*, Singapore, Y.-C. Hu, S. Tiwari, K. K. Mishra, and M. C. Trivedi, Eds. Springer Singapore. 97-105.
[44] G. M. D'Angelo, J. Luo, and C. Xiong. (2012). Missing data methods for partial correlations. *Journal of Biometrics & Biostatistics, 3*(8).
[45] D. Singh and B. Singh. (2020). Investigating the impact of data normalization on classification performance. *Applied Soft Computing, 97*, 105524. Doi: https://doi.org/10.1016/j.asoc.2019.105524.
[46] I. Pratama, A. E. Permanasari, I. Ardiyanto, and R. Indrayani. (2016). A review of missing values

handling methods on time-series data. *2016 International Conference on Information Technology Systems and Innovation (ICITSI)*. 1-6. Doi: 10.1109/ICITSI.2016.7858189.

[47] R. Razavi-Far, B. Cheng, M. Saif, and M. Ahmadi. (2020). Similarity-learning information-fusion schemes for missing data imputation. *Knowledge-Based Systems, 187*, 104805. Doi: https://doi.org/10.1016/j.knosys.2019.06.013.

[48] T. Chai and R. R. Draxler. (2014). Root mean square error (RMSE) or mean absolute error (MAE). *Geoscientific Model Development Discussions, 7*(1), 1525-1534.