An Intelligent Missing Data Imputation Techniques: A Review

Kimseth Seu a, Mi-Sun Kang b, HwaMin Lee c,*

a Department of Software Convergence, Soonchunhyang University, Asan, Republic of Korea
b Department of Computer Software Engineering, Soonchunhyang University, Republic of Korea
c Department of Medical Informatics, Korea University, Seoul, Republic of Korea

Corresponding author: hwamin@korea.ac.kr

Abstract—The incomplete dataset is an unescapable problem in data preprocessing that primarily machine learning algorithms could not employ to train the model. Various data imputation approaches were proposed and challenged each other to resolve this problem. These imputations were established to predict the most appropriate value using different machine learning algorithms with various concepts. Furthermore, accurate estimation of the imputation method is exceptionally critical for some datasets to complete the missing value, especially imputing datasets in medical data. The purpose of this paper is to express the power of the distinguished state-of-the-art benchmarks, which have included the K-nearest Neighbors Imputation (KNNImputer) method, Bayesian Principal Component Analysis (BPCA) Imputation method, Multiple Imputation by Center Equation (MICE) Imputation method, Multiple Imputation with denoising autoencoder neural network (MIDAS) method. These methods have contributed to the achievable resolution to optimize and evaluate the appropriate data points for imputing the missing value. We demonstrate the experiment with all these imputation techniques based on the same four datasets which are collected from the hospital. Both Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) are utilized to measure the outcome of implementation and compare with each other to prove an extremely robust and appropriate method that overcomes missing data problems. As a result of the experiment, the KNNImputer and MICE have performed better than BPCA and MIDAS imputation, and BPCA has performed better than the MIDAS algorithm.

Keywords—Data imputation technique; missing data; machine learning; deep learning

Manuscript received 27 Nov. 2021; revised 20 Jan. 2022; accepted 10 Mar. 2022. Date of publication 31 May 2022.

I. INTRODUCTION

In the machine learning field, many algorithms are exploited and proposed by the researcher into numerous estimation applications for the time ahead prediction and analyzing and drawing inference, such as medical systems [1]–[2], network fields, recommendation systems [3], robotics, and industries. To build up each trustable modeling application system, only one of the most significant technical algorithms is not enough to make the best application. Still, it also demands an essential full sample in the dataset to acquire effective estimation. In addition, the dataset resource requires definite observation and experiment though following the actual situation in the previous sample and measured absolutely from sample targets. Therefore, the intended data resource is the significantly principal element that is attentively collected and arranged to intend to train the prediction model appropriately. However, the arbitrary missing value pattern in data mining is an unescapable problem that makes many data analysis applications generate low-performance decisions and inferences. Furthermore, the ubiquity of missing data can also cause failure modeling or unreliable prediction. Before training the datasets, the robust method is raised as the essential principle requirement to manipulate the incomplete data for classification, regression, and time-series prediction [4]. In the missing value problem, the incomplete dataset was considered as three categories of missing patterns such as missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) [5]. MCAR is a type of missing value pattern in which data points have vanished independently of other values in each sample. MAR is also a missing pattern in which values in each feature of the dataset are vanished depending on values in another feature.

Furthermore, the value in feature in which is vanished rely on their value in feature is called MNAR. In the simple benchmarks, which have included the K-nearest Neighbors Imputation (KNNImputer) method, Bayesian Principal Component Analysis (BPCA) Imputation method, Multiple Imputation by Center Equation (MICE) Imputation method, Multiple Imputation with denoising autoencoder neural network (MIDAS) method. These methods have contributed to the achievable resolution to optimize and evaluate the appropriate data points for imputing the missing value. We demonstrate the experiment with all these imputation techniques based on the same four datasets which are collected from the hospital. Both Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) are utilized to measure the outcome of implementation and compare with each other to prove an extremely robust and appropriate method that overcomes missing data problems. As a result of the experiment, the KNNImputer and MICE have performed better than BPCA and MIDAS imputation, and BPCA has performed better than the MIDAS algorithm.

Keywords—Data imputation technique; missing data; machine learning; deep learning

Manuscript received 27 Nov. 2021; revised 20 Jan. 2022; accepted 10 Mar. 2022. Date of publication 31 May 2022.

International Journal on Informatics Visualization is licensed under a Creative Commons Attribution-Share Alike 4.0 International License.
But the method may perform in a significant loss of statistical information and precision under a complex multivariate analysis [7]. Moreover, the mean, mode, and median, which are general methods, can apply to the missing data. However, this method is not a suitable solution for some datasets which consist of many incomplete values. Although the machine learning algorithms were exploited in numerous estimation applications for the time ahead prediction and objective classification, various up-to-date imputation methods were also proposed to handle this problem effectively via using convenient machine learning algorithms such as the regression method [8], the k-nearest neighbor method [9], deep learning approach [10-11], the neural network-based method [12] with advanced statistics strategies [13], [14]. The most appropriate value estimation predicted by these imputations used incompatible algorithms. Moreover, the accuracy of imputation methods must be extremely critical to complete information into the dataset, especially imputing the value medical dataset. It is a conditional situation that inaccurately leads to classification estimation or failure prediction result.

The proposed methods have contributed to the achievable resolution to optimize and evaluate the appropriate data points for imputing the missing. The paper expresses the various powers of the missing data imputation algorithms by using machine learning approaches. In addition, we demonstrate the experiment with these imputation methods based on four datasets collected from the hospital. Furthermore, we would indicate which methods have done better to complete losing information in the data system by comparing outcome performance.

II. MATERIAL AND METHOD

A. Related Work

Each imputation approach has performed different situation-based to estimate missing value lost from the dataset. According to the related work acquired in the method, the missing data imputation was grouped into four categories: the global approach, local approach, hybrid approach, and knowledge assisted approach [15-16]. Imputation algorithms of the different categories are shown in Table 1.

<table>
<thead>
<tr>
<th>Class</th>
<th>Algorithm</th>
<th>Year</th>
<th>Remark</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global</td>
<td>SVDImpute</td>
<td>2001</td>
<td>Using singular value decomposition (SVD) to acquire eigengenes that is a set of patterns of mutually orthogonal expression [17]</td>
</tr>
<tr>
<td></td>
<td>BPCA</td>
<td>2015</td>
<td>Bayesian theory and principal component analysis technique are used to predict the missing value [18]</td>
</tr>
<tr>
<td></td>
<td>PPCA</td>
<td>2019</td>
<td>The missing value is predicted by probability methodology and principal component analysis technique [13]</td>
</tr>
<tr>
<td>Local</td>
<td>KNNImpute</td>
<td>2016</td>
<td>Estimate the missing value based on K-nearest neighbor methods [9]</td>
</tr>
<tr>
<td></td>
<td>SLLSimpute</td>
<td>2008</td>
<td>Missing value estimation is made by sequentially imputing based on the local least square technique [20]</td>
</tr>
<tr>
<td></td>
<td>LLSimpute</td>
<td>2013</td>
<td>Local least square is used to define coherent gene for imputing missing value [19]</td>
</tr>
<tr>
<td></td>
<td>IKNNImpute</td>
<td>2021</td>
<td>Estimate missing data by iterative k-nearest neighbor [8]</td>
</tr>
<tr>
<td>Hybrid</td>
<td>LinCmb</td>
<td>2005</td>
<td>Combination of five different imputation methods such as row average, KNNimpute, SVDimpute, BPCA, and GMCimpute to estimate the missing values [21]</td>
</tr>
<tr>
<td></td>
<td>DIFC</td>
<td>2019</td>
<td>Iterative fuzzy cluster and Decision tree are used for missing value estimation. [5]</td>
</tr>
<tr>
<td></td>
<td>HIMP</td>
<td>2021</td>
<td>A four-layer model estimates the missing value to develop the imputation focus on multi-pattern missing data [29]</td>
</tr>
<tr>
<td></td>
<td>FCM+GA</td>
<td>2015</td>
<td>Using the Integrate fuzzy C-means with genetic algorithm to estimate missing value [30]</td>
</tr>
<tr>
<td></td>
<td>GFCMI</td>
<td>2019</td>
<td>Combination of fuzzy c-means, mutual information-based feature selection, and regression model to estimate the missing value [23]</td>
</tr>
<tr>
<td>Knowledge</td>
<td>POCSimpute</td>
<td>2006</td>
<td>Theoretic framework and biological knowledge are used to discover the missing data [31]</td>
</tr>
<tr>
<td></td>
<td>GOimpute</td>
<td>2008</td>
<td>Gene ontology is used to investigate similarity originating to select relevant genes for missing data estimation. [32]</td>
</tr>
</tbody>
</table>

1) **Global Method:** The algorithm in this global category performs missing value imputation depending on global correlation information, which is computed derivation from whole data matrices. The methods assume the existence of the global covariance structure among all genes or samples in the expression matrices [15]. The assumption is not adequate when they have faced a particular condition. For instance, these outcome performances of the imputations decrease accuracy when each sample shows a dominant local similarity structure. The SVDimpute [17] and Bayesian principal component analysis (BPCA) imputation [18] are well-known as the missing data imputation in this category. For the SVD imputation, the eigengenes, a set of manually orthogonal expression patterns, are acquired by the singular vector decomposition algorithm. This algorithm makes approximately the expression of all values in the dataset from the linear combination of this algorithm. This method firstly makes the regression of the gene against the k most significant eigengenes. Then, the missing value is reconstructed from the linear combination of the k eigengenes by employing regression coefficients. For the BPCA imputation, the d-dimension gene indicates vectors as a linear combination principal axis vector with utilizes an Expectation-Maximization (EM)-like algorithm to estimate the posterior distributions of the model parameter and the missing data simultaneously. Based on the related structure of the data system, local value imputing approaches are performed.
2) Local Method: In the local approach, this category only exploits the local similarity structure in the dataset to reconstruct the value that is discarded. The missing values are only computed from the instance subset that is a high correlation with the sample that contains the missing values [15]. The k-nearest neighbor imputation (KNNimpute) and local least square imputation (LLSimpute) are widely used existing imputation methods are among this approach category [8], [19-20]. For The KNNimpute, this method has performed k-nearest neighbor algorithms depending on k number of high sample correlation with gene contained missing value to compute missing data in the dataset. The LLSimpute imputes the disappeared point from all k-related samples simultaneously by deploying a multiple regression model [20].

3) Hybrid method: In the hybrid approach, this algorithm is a unique imputation developed by converging the local method with the global method despite the performance of imputation algorithms based on the kind of correlation structure in the dataset. Local correlation structure between samples is influential for heterogeneous datasets, and localized imputation approaches such as LLSimpute or KNNimpute perform better than global imputation methods such as SVDimpute or BPCA. In contrast, the local imputation technique would be less performant than the global approach, such as SVDimpute or BPCA, if the dataset is homogeneous, which better captures global correlation information. A LinCmb imputation method is a hybrid approach that captures both the data's global and local correlation samples [21]. This method utilized a convex combination of different imputation algorithms such as row average, kNNimpute, SVDimpute, BPCA, and GMCimpute to estimate the missing values in the dataset. SVDimpute and BPCA deploy global correlation information in their imputing value, whereas local correlation information is utilized by kNNimpute, row average, and GMCimpute. Another hybrid imputation is a GFCMI imputation that evolved from the convergence of three algorithms such as the Grey System Theory, Fuzzy c-Mean (FCM), and Mutual Information (MI) [22]. This method utilizes the FCM technique to optimize the membership degrees and the cluster prototypes, whereas the grey system theory concept is used to measure the similarity between the input sample and previous value, which determines uncertain systems with known and unknown information. Moreover, MI method is applied to discover a set of strong correlation information. The MI assumes some dependence between variables, such as linear data.

4) Knowledge Assisted method: For the method in this category, this algorithm implements imputing missing value by integrating the domain knowledge or external information to apply to a specific data format. Domain knowledge is an extremely important part of the highly enhanced performance of the imputation method with a purely data-driven method, especially over the dataset's condition that contains a small number of samples, noisy, or high missing rate. The Projection onto Convex Set (POCS) [23] is an algorithm into the knowledge-assisted method that deploys the flexible set-theoretic framework, which is the biological occurrence of synchronization loss and correlation information between genes and arrays. Local least square regression [19], [20] were generated by the POCS method to perform PCA imputation to capture gene-wise correlation and array-wise correlation and restrict the squared power of the expressions profiles for synchronization loss capturing.

B. Methods

To enhance the accuracy of imputing model, various imputation methods were proposed and optimized with state-of-art algorithms to complete the missing data with appropriated value in the dataset. And these approaches have performed different accurate estimations for imputing missing values. Thus, we demonstrate the robust missing data imputation method in this article.

1) KNNimpute Algorithm: The KNNimpute is the algorithm in the local method category, which uses the KNN-based algorithm to compute missing values in the dataset. The KNN method selects samples as a subset with high similarities to the gene containing incomplete data to impute missing values. The KNN algorithm determined the subset using similarity methods such as the Euclidean distance matrix formula, Pearson correlation, and variance minimization to observe the whole non-information in the dataset and evaluate the closest points or highest similar value as an appropriate sample to impute relevant value in the incomplete dataset. The Euclidean distance was determined as a sufficiently accurate standard method to discover the similarity gene [17]. Suppose that A sample contained a missing value. The KNN then trains the remaining values in the sample with other genes without the missing value to evaluate expression most closely similar as a subset. The K number defines the number of genes in the subset. The subset is then utilized to estimate the missing value in the A sample. Finally, the A sample is imputed by the weighted average of the contributed subset [18].

2) BPCA Imputation Algorithm: The BPCA imputation is known as the global category, which uses the global correlation to compute the missing value. This BPCA-based method was utilized to estimate the incomplete data in the gene. This algorithm comprises three methodologies procedures: Bayesian estimation, principal component regression, and expectation-maximization repetitive algorithm [18]. The PCA represents the D-dimensional microarray expression vector Y as a linear combination of principal axis vectors $W_l$.

$$y = \sum_{l=1}^{K} x_l w_l + \varepsilon$$ \hspace{1cm} (1)

Where the linear coefficients $x_l (1 \leq l \leq K), (K < D)$ are namely factor scores and $\varepsilon$ is the residual error. As matrix $y$ is the existence of missing values in the matrix, $Y$. PC regression separates the missing part $y^{miss}$ and observed part $y^{obs}$ and the PCA is then used to estimate the missing part $y^{miss}$ from the observed part $y^{obs}$ in the expression vector $Y$. Respectively, To correspond to the observed part and missing part in $y, w_l^{miss}$ and $w_l^{obs}$ is acquired as parts of each principal axis $w_l$. The residual error, which is a well-known regression problem, then, is utilized to acquire the factor scores $x = (x_1, x_2, \ldots, x_K)$ for the expression vector $y$.

$$err = |y^{obs} - W^{obs}x|$$ \hspace{1cm} (2)
Thus, the missing party $y^{\text{miss}}$ is estimated:

$$y^{\text{miss}} = W^{\text{miss}} x$$ (3)

3) MICE Algorithm: Multiple imputations by chained equations (MICE) [24-26] is a particular multiple imputation algorithm that performs a better estimation for finding out the missing value in the dataset. This method has substantial flexibility in multiple imputing proceedings with be used in a broad range of settings. MICE’s procedure imputes the missing values depending on an iterative series of estimated models, which uses all other variables to predict the missing variable in the dataset. In addition, MICE cannot only impute numerical datatypes, but continuous, binary, unordered categorical, and ordered MICE can also impute categorical data. This method performs under certain assumptions of the missing data pattern, such as MAR and MCAR, in which the probability of the missing value is based on observed values, not on unobserved values [17]. The multiple imputations are generated the complete value by four general processes of the chained equation procedure. Firstly, statistical mean imputation is applied to complete the missing values in the whole dataset. The statistical mean imputation is considered a placeholder. The previous placeholder value is then set back to missing for the first column in the dataset. The Regression model is used to perform under the same assumptions, which can be linear, logistic, or poison regression model to estimate missing data [27]. In the regression model, the observed values set back are thought of as dependent variables, and the other variables are independent variables. After the regression model imputes the missing value, it is utilized as an independent variable in the regression models for all the other variables, which means that both observed values and imputed values are used until complete all missing values. All the procedures previously mentioned, except the first process, are repeated for each variable with incomplete data until all variables have been imputed.

4) MIDAS Algorithm: The MIDAS method is constructed from multiple imputations with denoising autoencoder neural networks [28]. The standard denoising autoencoder model is adjusted in two fundamental ways by MIDAS. Firstly, in the initial corruption procedure part, all missing value is transformed to 0. Thus, the denoising autoencoder predicts corrupted values, both originally missing values and observed initially. Secondly, MIDAS makes a regular denoising autoencoder with a dropout technique to decrease the overfitting problem. MIDAS alternates the sample thinned network, which thinned network is arbitrarily sampled to let dropout training proceeds to generate multiple imputations.

The imputation-generating encoder trained with dropout can be shown below:

$$y = f_{\theta}(x) = \sigma(W^{\text{p}}y^{\text{p}})[\sigma(W^{\text{e}}x^1 + b^1)] + b^2 \ldots$$ (4)

And the decoder is shown:

$$y = g_{\theta'}(y) = \phi(W^N[y][\sigma(W^{B+1})y + b^{B+1})][\sigma(W^{B+1})] + b^N] \ldots$$ (5)

Where $g$ is a Gaussian process (GP), a commonly used probability distribution over function, and $z$ represents fully observed vector predictions of the observed initially and originally missing values.

The full architecture of a MIDAS network utilizes an activation function, an exponential linear unit (ELU) used to simplify efficient training in deep learning networks. And activation function of the output layer is selected depending on continuous, categorical, or binary variables. For loss function, it is measured as the distance between $x$ and $z$: $L(x, z)$. In the reconstruction error for estimations of initially observed corrupted value, the missing indicator vector $r$ is, thus, multiplied with these functions. In addition, root mean squared error (RMSE) and cross-entropy loss functions are used for both categorical and continuous variables in the MIDAS model.

$$L(x, z, r) = \left\{ \begin{array}{ll}
\frac{1}{2} \sum_{j=1}^{n} [y(x_j) - z_j]^2 \\
\frac{1}{2} \sum_{j=1}^{n} [y(x_j)\log z_j + (1 - x_j)\log(1 - z_j)]
\end{array} \right. \quad (6)$$

C. Experiment

1) Datasets: The retrospective observational cohort study, which was collected from Soonchunhyang University Cheonan Hospital between 2015 and 2020, was conducted in the experiment. The datasets are measured from patients who have gotten a pesticide intoxication by the electronic medical records information on outcomes, such as the time of respiratory failure. The selected datasets are shown in detail in Table 3.

2) Implementation: In the experiment, we demonstrate the experimental performance of these methods and compare their performance to define which imputation performance of the approaches is a powerful method for imputing missing data.

<table>
<thead>
<tr>
<th>No</th>
<th>Dataset</th>
<th>Samples</th>
<th>Features</th>
<th>Categories</th>
<th>Admission Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DI_96_1hr</td>
<td>39,143</td>
<td>88</td>
<td>10</td>
<td>2016-03-01 ~ 2020-12-28</td>
</tr>
<tr>
<td>2</td>
<td>DI_combined_20211230</td>
<td>1,790</td>
<td>190</td>
<td>83</td>
<td>2015-01-02 ~ 2020-12-30</td>
</tr>
<tr>
<td>3</td>
<td>DI_laba20_data_20211109_all</td>
<td>77,732</td>
<td>98</td>
<td>13</td>
<td>2012-01-02 ~ 2020-02-14</td>
</tr>
<tr>
<td>4</td>
<td>DI_vital_20211109</td>
<td>146,647</td>
<td>12</td>
<td>3</td>
<td>2016-02-23 2020-12-30</td>
</tr>
</tbody>
</table>

In particular, four datasets that already contain the missing value are utilized to implement each strategy. Initially, these datasets are transformed the categorical variable to numerical data and then scales these datasets to the same scale between 0 and 1 because different scales of variables lead the training model to generate a biased replacement for the missing values. Then, the originally missing data were imputed by using each imputation algorithm. After assigning the missing
value, the datasets generate missingness randomly with a ratio of 30% of each dataset. Later, these missing data are imputed with each imputation again to evaluate. Finally, the root means square error (RMSE) and mean absolute error (MAE) are employed to calculate each approach's performance.

III. RESULTS AND DISCUSSION

The result is shown in Table 3, which gives the expression of each performance approach imputed with four missing datasets by using MAE and RMSE measurements.

**Table III**
The measurement performance of four imputation methods with four missing datasets.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Dataset</th>
<th>Accuracy</th>
<th>Performance</th>
<th>MAE</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNNimputer</td>
<td>DI_96_1hr</td>
<td>0.0047</td>
<td></td>
<td>0.0290</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DI_combined_20211230</td>
<td>0.0230</td>
<td></td>
<td>0.0870</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DI_labdata_20211109_all</td>
<td>0.0001</td>
<td></td>
<td>0.0050</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DI_vital_20211109</td>
<td>0.0220</td>
<td></td>
<td>0.0700</td>
<td></td>
</tr>
<tr>
<td>BPCA</td>
<td>DI_96_1hr</td>
<td>0.0196</td>
<td></td>
<td>0.0948</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DI_combined_20211230</td>
<td>0.0313</td>
<td></td>
<td>0.1012</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DI_labdata_20211109_all</td>
<td>0.0012</td>
<td></td>
<td>0.0090</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DI_vital_20211109</td>
<td>0.0248</td>
<td></td>
<td>0.0685</td>
<td></td>
</tr>
<tr>
<td>MICE</td>
<td>DI_96_1hr</td>
<td>0.0048</td>
<td></td>
<td>0.0241</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DI_combined_20211230</td>
<td>0.0104</td>
<td></td>
<td>0.0378</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DI_labdata_20211109_all</td>
<td>0.0015</td>
<td></td>
<td>0.0044</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DI_vital_20211109</td>
<td>0.0262</td>
<td></td>
<td>0.0880</td>
<td></td>
</tr>
<tr>
<td>MIDAS</td>
<td>DI_96_1hr</td>
<td>0.0322</td>
<td></td>
<td>0.0913</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DI_combined_20211230</td>
<td>0.0462</td>
<td></td>
<td>0.1234</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DI_labdata_20211109_all</td>
<td>0.0257</td>
<td></td>
<td>0.0438</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DI_vital_20211109</td>
<td>0.0470</td>
<td></td>
<td>0.1089</td>
<td></td>
</tr>
</tbody>
</table>

The result is compared with the same dataset, as shown in Fig. 1, which is a graphical representation of the comparison of the methods that perform on the DI_96_1hr dataset, which indicates that MSE and RMSE of KNNimputer and MICE are lower than other methods.

This comparison expresses that both MICE and KNNimputer algorithms perform better for imputing missing in the "DI_96_1hr" dataset. On the other hand, Fig. 2 indicates that the MICE method performs better on the "DI_combined_20211230" dataset, and MIDAS performs lower than the other method. Fig. 3 and Fig. 4 point out that the KNNimputer, MICE, and BPCA imputation method on both "DI_labdata_20211109" and "DI_vital_20211109" datasets are better performance than the MIDAS method. This comparison shows that KNNimputer, MICE are the best imputation method appropriate for imputing missing values in the dataset. However, MIDAS and BPCA are lower performance than other methods, and this algorithm is still a powerful technique with the result of MAE and RMSE measurement on our four datasets. However, depending on the amount of our data resources for implementation with these methods has limitations. That means our data resource is not enough for performance evaluation to define the best solution for imputing missing value or is generally utilized in the data preprocessing challenge. It is required to experience various or more complex datasets and contain different ratios of missing values in the dataset.

![Fig. 1](#) Comparison of MSE and RMSE measurement of KNNimputer, BPCA, MICE, and MIDAS perform on "DI_96_1hr" dataset.

![Fig. 2](#) Comparison of MSE and RMSE measurement of KNNimputer, BPCA, MICE, and MIDAS perform on "DI_combined_20211230" dataset.

![Fig. 3](#) Comparison of MSE and RMSE measurement of KNNimputer, BPCA, MICE, and MIDAS perform on "DI_labdata_20211109_all" dataset.

![Fig. 4](#) Comparison of MSE and RMSE measurement of KNNimputer, BPCA, MICE, and MIDAS perform on the "DI_vital_20211109" dataset.
Many imputation methods were proposed and optimized with states-of-art algorithms and statistical concepts to enhance the accurate imputing model against the missing data challenge. In this article, we have reviewed the various well-known artificial intelligence algorithm-based missing data imputation, aiming to evaluate which method performs perfectly to impute the missing data in the dataset. In addition, we implement the experiment of these imputation methods with four real datasets of the patient, which were collected from the hospital, to compare the performance. The comparison result of these methods indicates that KNNimputer and MICE perform the most excellent approach to imputing missing value. Moreover, the BPCA performs better than the MIDAS algorithm. However, for the evaluation performance of these methods, it is not enough yet to define which one is the best way to utilize in the data preprocessing challenge. It is required to experience various or more complex datasets and the ratio of missing values in the dataset. In the future, we will explore these missing data imputation algorithms with other datasets to find the best method.

ACKNOWLEDGMENT

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. NRF-2021R1A2C1009290).

REFERENCES