A Novel Clustering-Based Purity and Distance Imputation for Handling Medical Data With Missing Values

Ching-Hsue Cheng (chcheng@yuntech.edu.tw)
National Yunlin University of Science and Technology
Shu-Fen Huang
Chihlee University of Technology

Research Article

Keywords: Health records, Data imputation, Clustering, Missing values, Purity-based k nearest neighbors imputation (PkNNI), distance-threshold nearest neighbors imputation (DNNI)

Posted Date: May 17th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-520996/v1

A Novel Clustering-Based Purity and Distance Imputation for Handling Medical Data with Missing Values

Ching-Hsue Cheng[1,*], Shu-Fen Huang[2]

[1]Department of Information Management, National Yunlin University of Science & Technology, Touliou, Yunlin 64002, Taiwan.
[2]Department of Multi Media Design, Chihlee University of Technology, New Taipei City 22050, Taiwan.

Abstract

Nowadays, people pay increasing attention to health, and the integrity of medical records has been put into focus. Recently, medical data imputation has become a very active field because medical data usually have missing values. Many imputation methods have been proposed, but many model-based imputation methods such as expectation-maximization and regression-based imputation based on the variables data have a multivariate normal distribution, which assumption can lead to biased results. Sometimes this becomes a bottleneck, such as computationally more complex than model-free methods. Furthermore, directly remove instances with missing values, this approach has several problems, and it is possible to lose the important data, produce ineffective research samples, and cause research deviations, and so on. Therefore, this study proposes a novel clustering-based purity and distance imputation method to improve the handling of missing values. In the experiment, we collected eight different medical datasets to compare the proposed imputation methods with the listed imputation methods with regard to the results of different situations. In imputation measures, the area under the curve (AUC) is used to evaluate the performance of the imbalanced class datasets in MAR and MCAR experiments, and accuracy is applied to measure its performance of the balanced class in MNAR experiment. Finally, the root-mean-square error (RMSE) is also used to compare the proposed and the listing imputation methods. In addition, this study utilized the elbow method and the average silhouette method to find the optimal number of clusters for all datasets. Results showed that the proposed imputation method could improve imputation performance in the accuracy, AUC, and RMSE of different missing degrees and missing types.

Keywords: Health records; Data imputation; Clustering; Missing values; Purity-based k nearest neighbors imputation (PkNNI); distance-threshold nearest neighbors imputation (DNNI).

*Corresponding author:
Ching-Hsue Cheng
Department of Information Management, National Yunlin University of Science & Technology, 123, section 3, University Road, Touliou, Yunlin 64002, Taiwan
Email: chcheng@yuntech.edu.tw

1. Introduction

Nowadays, everyone is concerned about the state of their health and pays increasing attention to healthcare issues, with physicians confirming disease diagnosis from the relevant medical records of the patient. However, the medical records usually have missing values due to, for example, privacy of patients, evading medical examination, and avoiding medical treatment, etc. Missing data have a limitation when interpreting research results because a missing data problem is typically used by removing the observation with missing values. This may exclude the important information that is of theoretical interest and increase the misclassification cost, such as type I and Type II errors. In addition, the main drawbacks of removed instances with missing values will reduce the number of samples; the estimates will have larger standard errors and possible analysis bias. Hence, in medical research, data imputation has become an important research issue. Furthermore, medical data usually are imbalanced classes where the class distributions are not represented equally, which will produce a misclassification cost problem.

To handle missing values, many researchers have proposed various types of techniques, but many model-based imputation methods such as expectation-maximization and multiple imputation based on the variables data have multivariate normal distribution [1]. More assumptions can lead to biased results, and the model becomes difficult to learn; sometimes, this is a bottleneck, such as computationally more complex than model-free methods. In addition, most of them have not implemented a complete evaluation of all missing degrees and missing types. The most common method is to remove the records with missing values, making it possible to lose important data, produce ineffective samples, and cause research deviations. In previous work on handling missing values, there are several common methods to replace the missing values [2], such as the average of other complete values and average of same class values (class average imputation). After replacing the missing values, they usually have better accuracy than that of the original datasets [3], and these replaced missing values methods are usually called imputation methods [4]. However, a common imputation approach has a shortcoming of deleted missing values, potentially removing something important and causing research bias.

From the mentioned problems above, this study proposes a novel clustering-based purity and distance imputation method to improve the handling of missing values. Because the proposed imputation is a hybrid method, and the attributes data do not obey multivariate normal distribution. The proposed approach is combined k-mean with Purity-based k Nearest

Neighbors Imputation (PkNNI) and Combined clustering with Distance-threshold Nearest Neighbors Imputation (DNNI). The main difference between the two approaches (PkNNI and DNNI) and the proposed method is to find the optimal number of clusters and adapt the two kNN related approaches to obtain the optimal results for the eight medical datasets. Medical data have the higher MDs; hence the research works need to check a higher percentage of complete data (e.g., normally more than 50% of the total cases) and needed to impute less than three values per patient [5]. Therefore, this study simulated a massive missing degree (in the traditional definition is 360%, but the MD definition of this paper is 20%) to handle missing values. In evaluating metrics of imputation methods, most previously used accuracy and RMSE, but medical data usually are imbalanced classes. Therefore, this paper applied accuracy, the area under the curve (AUC), and root-mean-square deviation (RMSE) to handle the problem of imbalanced classes.

In summary, the objectives and contributions of this study are listed as follows:

(1) propose two novel algorithms based on the combined clustering with Purity-based k Nearest Neighbors Imputation (PkNNI) and combined clustering with Distance-threshold Nearest Neighbors Imputation (DNNI) to estimate the missing values,
(2) apply the elbow and average silhouette methods to find the optimal number of clusters consistently,
(3) compare the proposed imputation methods with the listing imputation methods in different missing degrees and missing types,
(4) evaluate the performance of the proposed imputation method by using Root-Mean-Square Error (RMSE), classification accuracy, and AUC, and
(5) apply the proposed imputation method to medical data with missing values.

The rest of this paper is organized as follows: Section 2 describes the related work, including medical data imputation, missing values type, clustering technique, and imputation methods. Section 3 introduces the concept of the proposed method and the proposed procedure. Section 4 outlines experiments and results. The discussions and findings are provided in Section 5. At last, the conclusions are given in Section 6.

2. Related work

This section introduced the related literature, including medical data imputation, missing value types, clustering technique, and imputation method.

2.1 Medical data imputation

Medical data imputation is an active area of research because medical data are relevant to

the health of patients and physician decision making, and therefore it is more important than any other type of data. The physician needs to refer to the medical record of patients for diagnosing their patients, and if the medical records have some errors, it may lead to diagnostic errors. It is very hard to achieve a complete record of medical data because there are various reasons for data with missing values, such as patient privacy, human negligence, and medical equipment dysfunction.

Missing values are most common when clinical trials are carried out because medical data are from various clinical trials, field experiments, or any other traditional mechanism. Proper care must be taken to handle missing values suitably and accurately. Hence, there are many challenges of missing value imputation in medical data; these challenges are listed in the follows. (1) A simple removal to handle missing values would be to discard the whole records, which essentially contains the missing value of an attribute, and this does not solve the problem. Measuring missing value incurs additional cost, whereas previously reported statistical methods result in reduced performance compared to when all variables are measured. (2) The higher missing degrees in the electrical health records have been previously reported from 20% to 80% [6]. Hence the majority of the research works normally presented a higher percentage of complete data (e.g., normally more than 50% of the total cases) and needed to impute less than three values per patient [5]. (3) The selection of the imputation method is a challenge, using the model-based or model-free method? (4) Evaluating imputation performance is also a challenge; most previously used accuracy and RMSE, but medical data usually are imbalanced classes.

In summary, missing medical values could face imputation, deletion, or classification of missing type. To avoid bias, classifying the pattern of missing values is an efficient step to select appropriate imputation methods to improve data consistency. Many imputation methods have been used in the medical field [4, 7-9], such as multiple imputation, expectation-maximization imputation, k nearest neighbor imputation, and Chained Equation Multiple Imputation (MICE). These studies are briefly introduced as follows. Sterne et al. [10] used MI in epidemiological and clinical research. Jerez, Molina, Subirats, and Franco [11] used artificial neural networks combined data imputation to prognosis breast cancer. García-Laencina et al. [12] proposed using k-nearest neighbor, mode, and expectation-maximization imputation for five-year survival prediction of breast cancer patients with unknown discrete values. Pombo, Rebelo, Araújo, and Viana [13] combined data imputation and statistics to design a clinical decision support system; the next year, they also proposed a patient-oriented method of a pain evaluation system [14] that produced tailored alarms, reports, and clinical guidance on the basis of collected patient-reported data, which was a clinical decision support systems.

2.2 Missing values type

In the real world, the collected datasets are usually incomplete because of reasons such as human negligence, equipment failure, network disconnection, data not originating from the same source, postprocessing errors, and collection phase having noise; we called these cases of non-structures missingness. Rubin [15] reported that there are three types of missing values, Missing At Random (MAR), Missing Completely At Random (MCAR), and Missing Not At Random (MNAR). This study focuses on the nonstructured missingness problem, which has three types: MAR, MCAR, and MNAR.

MAR type can be defined as a missing data point that does not depend on missing data but depends on some of the observed data. *i.e.*, the missingness is conditional on another variable. MCAR type that if missingness does not depend on either the observed data points or missing data, that is, a missing data point, is completely random. MNAR type is when missing values in a variable are related to the values of the variable itself, even after controlling for other variables. MNAR can have two origins, missingness depending on attributes of the instance of other missing data or a missing element dependent on its own value [15].

In practice, there are two ways of handling missing values, one is marginalization [16], and the other is imputation. Marginalization is deleting or ignoring missing values, and it is the most common approach in handling missing values. Marginalization may cause research bias by deleting data records/instances with missing values; data imputation can maintain the original dataset and avoid the problem of research bias.

2.3 Clustering

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). It is the main task of exploratory data mining and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.

Clustering is a machine learning technique that involves the grouping of data points. Given a set of data points, we can use a clustering algorithm to classify each data point into a specific group. In theory, data points that are in the same group should have similar properties (features), while data points in different groups should have highly dissimilar properties. Clustering is a method of unsupervised learning, which is a common technique for statistical data analysis used in many fields. In data science, we can use clustering analysis to gain valuable insights from

the collected data and visualize which group data points fall.

There are many clustering methods, such as connectivity-based clustering (hierarchical clustering), centroid-based clustering, distribution-based clustering, and density-based clustering. This study uses k-means clustering, which is centroid-based clustering, where clusters are represented by a central vector, and the center of the cluster may not necessarily be a member of the data set. When the number of clusters is fixed to k, we find k cluster centers and assign the objects to the nearest cluster center so that the squared distances from the cluster are minimized [17]. K-means clustering is defined as follows.

Suppose there is a set of d-dimensional data:

$$\boldsymbol{X_i \in R^d, i = 1,2, \ldots, n} \tag{1}$$

We set k (for k ≤ n) clusters$\{S_1, S_2, \ldots, S_k\}$, k-means clustering minimizes the error square between the data and the cluster center within the cluster. The mathematical formula is defined as follows:

$$\boldsymbol{arg_\mu \ min \sum_{c=1}^{K} \sum_{i=1}^{n} \|x_i - \mu_c\|^2} \tag{2}$$

where $\boldsymbol{\mu_c}$ denotes the cluster center, and $\|\boldsymbol{x_i - \mu_c}\|$ is Euclidean distance. Its algorithm steps are listed as follows:

1. Initially, randomly set the k cluster center.

$$\mu_c^{(0)} \in R^d, c = 1,2,\ldots,k \tag{3}$$

2. Calculate the instances classified into each cluster, where (t) is the t^{th} operation

$$\boldsymbol{s_c^{(t)} = \left\{ x_i : \left\| x_i - \mu_c^{(t)} \right\| \leq \left\| x_i - \mu_{c^*}^{(t)} \right\|, \forall i = 1, \ldots, n \right\}} \tag{4}$$

3. Update the cluster center (n_c data in cluster c.)

$$\boldsymbol{\mu_c^{(t+1)} = \frac{sum\left(s_c^{(t)}\right)}{n_c} = \sum_{i=1}^{n_c} x_i \Big|_{x_i \in s_c^{(t)}}} \tag{5}$$

4. Repeat Steps 2 and 3 until the cluster center no longer changes:

$$\boldsymbol{S_c^{(t+1)} = S_c^{(t)}, \forall_c = 1, \ldots, k} \tag{6}$$

Determining the best number of clusters is an important problem in the k-means clustering algorithm. There is much-related research, but two of the most famous and common methods are as follows:

(1) Elbow method

The purpose of clustering is to minimize the total variation within the cluster and maximize the total variation between clusters; the elbow method [18] determines the number of clusters. First, set a value of k. When the data are divided into k clusters, the sum of squares error (SSE) within the cluster is the smallest; then, k is the best number of clusters:

$$\text{Min}\sum_{k=1}^{k} W(C_k) \quad (7)$$

where C_k denotes the *k-th* cluster, and $W(C_k)$ is the error sum of squares (SSE) within the cluster.

(2) Average silhouette method

In addition to calculating the SSE, another method of measuring the effect of clustering is the average silhouette method [19]. The silhouette coefficient measures the effect of clustering on the basis of the cohesion and dispersion of each data point(*i*); its equation is as follows:

$$s(i) = \frac{|b(i)-a(i)|}{\max\{a(i),b(i)\}} \quad (8)$$

where

a(i) denotes the average distance between data point (*i*) and other data points in the cluster;

b(i) is the average distance between data point (*i*) and the data points of the other clusters, taking the minimum value;

$|b(i)-a(i)|$ means to take an absolute value for $b(i)-a(i)$;

$s(i)$ is silhouette coefficient $0 \leq s(i) \leq 1$, which can be considered as a data point*(i)* if the indicator is appropriate within the cluster to which it belongs, $s(i)$ close to 1 means that the data are properly clustered, and when $s(i)$= 0, it denotes that the number of clusters is 1.

2.4 Imputation method

This section introduced the related imputation methods, including simple, multiple, kNN family, and computational intelligence imputation.

2.4.1 Simple imputation

The simple imputation is common imputation methods [4, 16]; these include zero imputation, average imputation, and class average imputation. The Zero Imputation (ZI) is the simplest imputation function that fills the missing values with zero. Average Imputation (AI) [16] replaces a missing value with the averages of the corresponding attribute on the entire dataset. The class average imputation (CAI) or concept mean imputation replaces the missing value with the average of the attribute over all instances within the same class label.

2.4.2 Multiple imputation

Multiple imputation methods have been applied in medical studies [20-21], including multivariate imputation by chained equation (MICE) and expectation-maximization imputation. Zhang [20] introduced a multivariate imputation by chained equation (MICE) by using the R package in medical research. Ondeck et al. [21] used multiple imputation in arthroplasty research and presented the results of comparisons between the demographic characteristics of

patients with and without missing preoperative albumin and hematocrit values.

2.4.3 kNN family imputation

Recently, the k-Nearest Neighbor Imputation (kNNI) has become widely applied in medical imputation [22]. In kNNI, a dataset is divided into two sub-datasets: one is the dataset that contains incomplete data with missing values, and the other has complete data without any missing values. The missing values of the incomplete data are replaced by the average of the corresponding attribute of its kNN, and the kNN average is computed with the complete data. However, this method tends to cover noise and outliers to be part of the predictive value; it may affect the accuracy of the imputation. In kNNI family, Troyanskaya et al. [23] proposed a weighted kNN imputation (WkNNI); Keerin et al. [24] proposed a cluster-directed framework with neighbor-based imputation; and Lee and Styczynski [25] proposed a new no-skip kNN to impute MNAR values. Additionally, Cheng et al. [26] proposed the Purity k-Nearest Neighbors Imputation (PkNNI) is an extension of kNNI, which is based on purity training and purity imputation. Furthermore, Cheng et al. [27] proposed a distance nearest neighbors Imputation (DNNI) was also expanded by kNNI.

2.4.4 Computational intelligence imputation

There are many Computational intelligence imputations [28], such as neural networks, random forests, and fuzzy c-means (FCM), etc. Awan et al. [29] proposed a class-specific distribution by adapting the popular conditional generative adversarial networks (CGAN) to impute the missing data. We briefly introduce FCM imputations because this study is based on clustering imputation techniques. Hathaway and Bezdek [30] proposed four FCM imputation techniques, and they pointed out that whole data strategy and partial distance strategy are faster to end, but the optimal completion strategy (OCS) and nearest prototype strategy (NPS) were outperformed over the first two methods based on accuracy and misclassification errors. In addition, Al Shami et al.[31] applied FCM-based OCS and NPS to compare four statistical imputation methods in their work for accurately substitute missing scores when producing the intelligent synthetic composite indicators. Therefore, we briefly introduced FCM-based OCS and NPS methods as follows. (1) In OCS approach, the missing values are viewed as additional attributes to be optimized and then impute missing values at each iteration till it reaches the best estimates, (2) NPS is a OCS modification, which computes the partial distances, and missing values are estimated by their nearest prototype counterparts during each iteration. In the hybrid clustering- based imputation method, Dinh et al. [32] proposed a framework of clustering mixed numerical and categorical data with missing values, it used the decision-tree-based method to find the set of correlated data instance and used the mean and kernel-based methods to obtain

cluster centers at numerical and categorical attributes, and they applied the dissimilarity measure to calculate the distances between instance and cluster centers.

3. Proposed imputation method

Many previous studies directly deleted data with missing values, which may remove the key dataset information and cause research bias. Recently, many studies have applied statistical methods, such as hot-deck imputation [33] and cold deck imputation [34], to estimate the missing values. So far, there have been many studies that directly removed outliers from a collected dataset, but the outliers usually had their practical meanings in the real world, such as traffic and popular holiday sightseeing sites. Jerez et al. [35] reported that artificial intelligence imputation is better than traditional statistical imputation; therefore, this study proposes an artificial intelligence imputation method to estimate missing values. *i.e.*, the proposed method is based on conditional attributes by k-means to cluster the dataset. Then, we applied PKNNI and DNNI imputation to estimate the missing values. This study used PkNNI and DNNI imputations because PkNNI can help calculate the purity of the same class, and DNNI can calculate the weight distance to find the nearest-neighbor group. Furthermore, this study employed elbow and average silhouette methods to determine the best clustering number.

3.1 PkNNI and DNNI imputations

The PkNNI and DNNI imputations are the main techniques in the proposed method. Hence, the following introduces the computational steps of PkNNI and DNNI imputations and equations to explain the mathematics and meaning.

(A) The PkNNI Imputation [26]: The PKNNI can be divided into two parts, one is purity training, and the other is purity imputation.

(1) Purity training is to compute the purity of each complete instance to obtain the purity of the i-*th* instance, and the purity training $P_t(\ i\ )$ is defined as:

$$P_t(\ i\ ) = \sum_{s=0}^{k_1} V(C(i), N_s) \tag{9}$$

where the C(i) is the i-th complete instance from datasets X, k_1 is the number of the nearest neighbors for purity calculations. N_s denotes the s-*th* nearest neighbor instance, and the function $V(\ )$ returns the class label between instance C(i) and N_s to identify whether they are the same or not. The function vote() is expressed as:

$$V(C(i), C(j)) \begin{cases} 1\ , if\ \ L(i) = L(j) \\ -1\ ,\ if\ \ L(i) \neq L(j) \end{cases} \tag{10}$$

where the $L(i)$ denotes the i-*th* class label from datasets X, to deduce whether the instance $C(i)$ is pure or not by comparing the class label $L(i)$ and $L(j)$.

(2) Purity imputation is based on the complete instances and purity values to predict the missing values, and the imputation equation $M(i,j)$ is defined as follows:

$$M(i,j) = \frac{\sum_{S=1}^{k_2} R(S,j)}{k_2} \quad (11)$$

where $M(i,j)$ represents the i-*th* instance and its j-*th* attribute, which is the missing value. k_2 is the number of nearest neighbors, and $R(S,j)$ represents the s-*th* nearest instance and its j-*th* attribute, which is a collection that contains all positive purity instance information from the complete instance.

(B) DNNI imputation [27]: The computational steps of DNNI are introduced as follows.

(1) Calculate the weight set of the distance between each incomplete instance (using no missing values) and all complete instances; the weight set is defined as Equation (12):

$$W = [w_1, w_2, \dots, w_m] = \left[\frac{|X|}{2\sum_{x\epsilon X}||y_1 - x_1||^{\frac{2}{(p-1)}}}, \frac{|X|}{2\sum_{x\epsilon X}||y_2 - x_2||^{\frac{2}{(p-1)}}}, \dots, \frac{|X|}{2\sum_{x\epsilon X}||y_m - x_m||^{\frac{2}{(p-1)}}}\right] \quad (12)$$

where w_m is the weight of m-*th* incomplete data, and |X| represents the cardinality of the complete data points for training. Additionally, y_1 is the first instance of incomplete data, x_1 is the first instance of complete data, and p is the weighted distance parameters.

(2) Apply the W weight set to compute the weighted distance D_{ij} between x and y. The calculated equation is defined as follows:

$$D_{ij} = [d_{i1}, d_{i2,} \dots d_{in}] = [(w_i(y_i - x_1)^2), (w_i(y_i - x_2)^2), \dots, (w_i(y_i - x_n)^2)] \quad (13)$$

where d_{ij} is the weighted distance between y_i and x_i, w_i is the weight of y_i incomplete data, y_i is the ith incomplete data, and x_i is the i-*th* complete data.

(3) Imputation: The set of weighted distance D_{ij} has an adapted threshold to determine the set of nearest neighbors. The distance threshold is utilized to adjust the optimal nearest neighborhood for estimating missing values; hence the proposed method does not set the k value of the nearest neighborhood. That is, d_{ij} is less than the threshold, and then x_i will be added to the set of nearest neighbors. The imputation method is based on the adapted threshold to obtain the reference points, then apply central tendencies to estimate missing values from the set of nearest neighbors. The central tendencies have average, median, and geometric mean.

3.2 Proposed computational procedure

To easily understand the computational procedure from data collection to evaluation, we

used Figure 1 to show a visual flow. The computational procedure includes the important clustering data and data imputation. The five steps are introduced in detail as follows.

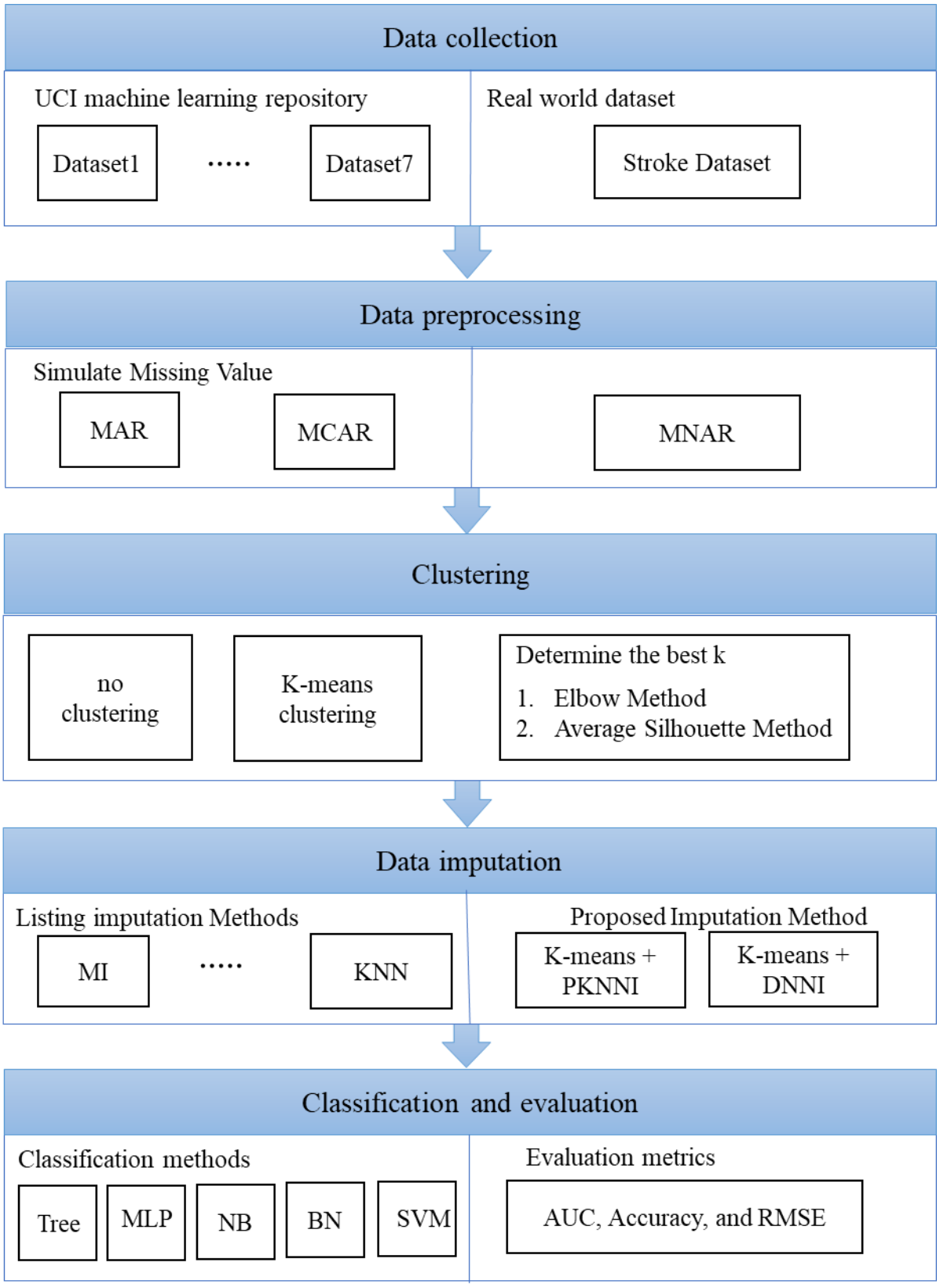

Figure 1. Computational procedure of this study

Step 1. Data collection

This step collected eight datasets to verify whether the proposed method was better than the listing imputation methods. The eight datasets were seven medical datasets from the UCI dataset repository and one stroke dataset collected from The International Stroke Trial database [36], which is a real-world dataset. The stroke dataset had missing values, but the seven other UCI datasets had no missing values. Therefore, this study uses different missing degrees to simulate missing values and verify the proposed imputation.

Step 2. Data preprocessing

Before the imputation step, the class attribute was converted from numeric to a character (symbol), and we merged multi-column class attributes into the one-column class attribute for the subsequent experiments. Next, on the basis of the different ratios of missing values, we generated missing values; that is, we randomly removed some data values from the original datasets on the basis of different missing degrees. "Missing degree" means the percentage of missing values in the datasets. Missing values were simulated in all conditional variables except the class label. In MAR and MCAR missing values types, the missing degree is defined as follows:

$$\textbf{Missing degree(MD)} = \frac{number\ of\ missing\ values}{number\ of\ (instances \times attributes)} \tag{14}$$

This study chooses 5%, 10%, 15%, and 20% MDs in MAR and MCAR types to verify the performance of the proposed imputation method and compare it with that of listed imputation methods. The 20% MD in this study has been a very large missing ratio; we used the climate model dataset with 18 attributes and 540 instances as an example, as shown in Table 3. From equation (12), we set 20% as MD, and then there are 1944 missing values ($0.2 = 1944/(18 \times 540)$. However, we used the MD of previous studies to calculate this case; the MD was 1944/540=360%. This study used a larger MD (20%) to simulate the experiment in MAR and MCAR types, the main difference that this study had considered the number of attributes. After the simulation of different MDs, each different MD dataset was partitioned into a complete and incomplete sub-dataset. Hence, each dataset had 16 sub-datasets: 5%, 10%, 15%, and 20% complete sub-dataset and 5%, 10%, 15%, and 20% incomplete sub-datasets in MAR and MCAR types.

Step 3. Clustering

From Step 2, this step employed k-means to cluster the 5%, 10%, 15%, and 20% complete sub-dataset in MAR and MCAR types for each dataset. Clustering meant that each datum in the

same cluster might have had some relationship with another. This step used k-means clustering to cluster all complete sub-dataset in MAR and MCAR types for each dataset.

To find the best number of clusters, this step used two more famous and common methods to determine the best k. One was the elbow method [18] to find the best k, where the data were divided into k clusters, and the SSE within the cluster was the smallest (as Figure 2). The other was the average silhouette method [19]; the silhouette coefficient measures the clustering effect on the basis of cohesion and dispersion of each data point(i) based on Equation (6). When s(i) was close to 1, it meant that the data were properly clustered (Figure 3).

Step 4. Data Imputation

In this step, all incomplete sub-datasets in MAR, MCAR, and MNAR types for each dataset were replaced by the proposed imputation method and the listed imputation methods. From Step 3, after optimal clustering, the complete sub-datasets were divided into three, five clusters, and the optimal number of clusters because this study wanted to confirm whether the best performance was the best k clusters; the number of clusters was set as odd numbers (3, 5, etc.). Next, this step used the computational steps of PkNNI (Equations (9) - (11)) and DNNI (Equations (12) and (13)) in Section 3.1 to estimate all incomplete MAR and MCAR sub-datasets.

Step 5. Classification and Evaluation

After imputation, the accuracy, AUC, and RMSE were applied to evaluate the performance of the proposed imputation. This study used the area under the receiver operating characteristic curve (AUC) because the receiver operating characteristic curves (ROC) has the diagnostic ability in imbalanced classes, and it can treat between positive detection rates and false alarm rates. Besides, the AUC is a measure of discriminative strength between these two rates without considering misclassification costs or class prior probabilities [37].

First, this step used C4.5, MLP, NB, BN, and LibSVM classifiers to evaluate accuracy and AUC. The classification results were calculated by using the confusion matrix [38]; the accuracy is defined as follows:

$$\text{Accuracy} = \frac{tp+tn}{tp+fn+fp+tn} \times 100 \qquad (15)$$

where tp is true positive, fp denotes false positive, fn represents false negative, and tn is true negative.

After imputation, this study used eight datasets to evaluate the accuracy, AUC, and RMSE. We used accuracy when data with class balance; if the dataset was imbalanced classes, then this study applied AUC to measure their performance.

Second, this study employed RMSE as the evaluation criteria to see whether there was bias in each imputation method. RMSE is an evaluation criterion to measure the bias between estimated values and real values. The RMSE formula is defined as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2} \tag{16}$$

where e_i, $i = 1, 2, \ldots, n$ denotes the bias between estimated values and the real values.

4. Experiments and results

This section presents the experimental environment, datasets, and experimental results.

4.1 Experimental environment and dataset

In this study, the experimental environment was Python (Python 2.7) on AMD Ryzen™ 7 2700X, 3.7 GHz 8-cores 16-threads CPU with a Windows 10 Home operating system to implement missing values imputation and comparison. Furthermore, the imputation methods and parameters are listed in Table 1. The k-means parameter was set as k= $\{2, 3, 5\}$, and kNNI was k= $\{3, 5, 7, 9\}$. In PkNNI, we set $k_1 = \{3,5,7,9\}$ and $k_2 = \{3,5,7,9\}$as the parameters, and set $\text{threshold} = \{0.1,0.2, \ldots, 1.9\}$ as DNNI parameters.

After imputation, this study implemented data classification and evaluated the performance of the proposed and listing imputation methods, and the parameter settings of classifiers are shown in Table 2. In the experiment, this study selected eight datasets, seven medical datasets from the UCI Machine Learning Repository, and one dataset from a real-world stroke dataset. The stroke dataset from IST [36] contains data between 1991 and 1996, its pilot phase was between 1991 and 1993, and it has 100% baseline data and over 99% complete follow-up data. It has 19,436 instances, 112 feature attributes with acute stroke. After screening and removing the irrelevant attributes, the stroke dataset had 39 attributes and 4241 instances for this study.

Table 1. Imputation method and parameters

Imputation method	Parameter	Reference
AI	None	Donders, van der Heijden [3]
CAI	None	Donders, van der Heijden [3]
ZI	None	Donders, van der Heijden [3]
MI	None	Rubin [15]
KNNI t1	K = {3,5,7,9}	Batista & Monard [22]
PKNNI t1	$K_1 = \{3,5,7,9\}$ $K_2 = \{3,5,7,9\}$	Cheng et al. [26]
DNNI	threshold={0.1,0.2,…,1.9}	Cheng & Haung [27]
kMeans + PkNNI		Proposed
kMeans + DNNI		Proposed
best k + PkNNI		Proposed
best k + DNNI		Proposed

Note. "best k" denotes the optimal cluster number based on the elbow method and average silhouette methods.

Table 2. Parameter settings of classifiers

Classifier	Parameters	Reference
C4.5	Confidence Factor:0.25	Quinlan [39]
MLP	Hidden Layers: (attributes + classes)/2 Learning Rate: 0.3 Momentum Rate:0.2 Validation Threshold:20	Mitra and Pal [40]
Naïve Bayes	None	John and langley [41]
Bayes Network	None	Pearl and Russell [42]
LibSVM	Cost: 1.0 Gamma: 0 Kernel Type: radial basis function	Chang and Lin [43]

We employed the eight datasets to verify whether the proposed imputation method was better than the listing imputations. The numbers of classes, attributes, imbalanced class, and instances of the UCI medical and stroke datasets are listed in Table 3. From Table 3, we see that all of the selected datasets are imbalanced classes, except for stroke datasets. Therefore, the AUC metric is used in the experiment of seven UCI datasets.

Table 3 Experimental dataset information.

Dataset	Number of classes	Number of attributes	Imbalanced class (ratio of class instances)	Number of instances
Liver disorders	2	7	Yes (145/200=0.73)	345
Acute inflammations	4	6	Yes (20:40:31:19, 40/19=2.1)	120
ILPD	2	10	Yes (416/167=2.49)	583
Banknote	2	5	Yes (762/610=1.25)	1372
Blood transfusion	2	5	Yes (570/178=3.20)	748
Climate model	2	18	Yes (46/494=0.09)	540
Haberman's survival	2	4	Yes (225/81=2.78)	306
Stroke	2	39	No (2096/2145=0.977)	4241

4.2 MAR and MCAR experiments

This experiment was divided into an internal and an external experiment; the internal experiment was to verify the performance of the proposed imputation method in different cluster numbers, MDs, and missing values types. After imputation, we employed the C4.5, MLP, NB, BN, and LibSVM classifiers to compare the accuracy of the best number of clusters with three and five clusters. The different imputation comparison was to compare the proposed

imputation with the listing imputation methods in different cluster numbers, MDs, and missing value types. In this experiment, the stroke dataset itself had missing values that belonged to the MNAR type, while the seven UCI datasets had no missing values. Therefore, this study used 5%, 10%, 15%, and 20% MDs to simulate missing values and verify the proposed imputation.

4.2.1 Internal comparisons

This section outlines applying elbow and the average silhouette methods to obtain a consistent optimal cluster number and compares PkNNI and DNNI with regard to two (the best number of clusters), three, and five clusters in different MDs, and missing value types. After implementing elbow and the average silhouette methods, the best number of clusters of the collected eight datasets is two clusters. Here, we only show the results of the liver disorders and ILPD datasets in MAR and MCAR types, and the five other datasets are listed in Appendix A.

(1) Liver disorders dataset

First, we implemented the elbow and average silhouette methods to obtain the best number of clusters, shown in Figures 2 and 3. Figures 2 and 3 obtained two clusters as the consistent best number of clusters for the liver-disorders dataset. Then, we applied the best number of clusters (two clusters), three, and five clusters, combining PkNNI and DNNI to impute the missing values in MAR and MCAR types.

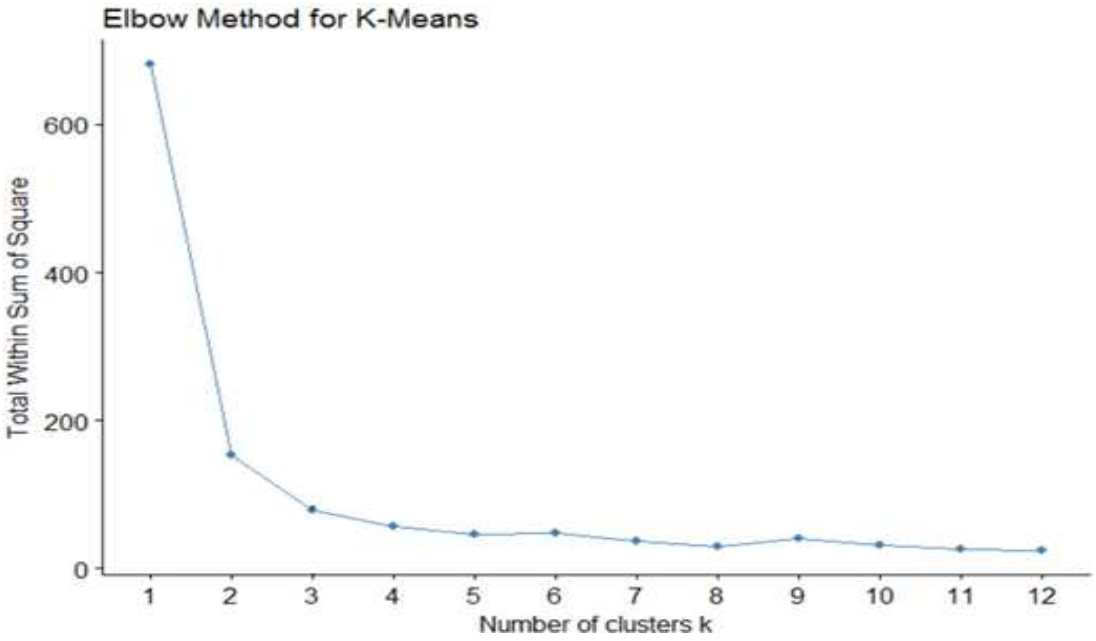

Figure 2. Elbow method showing the best k (liver-disorders dataset).

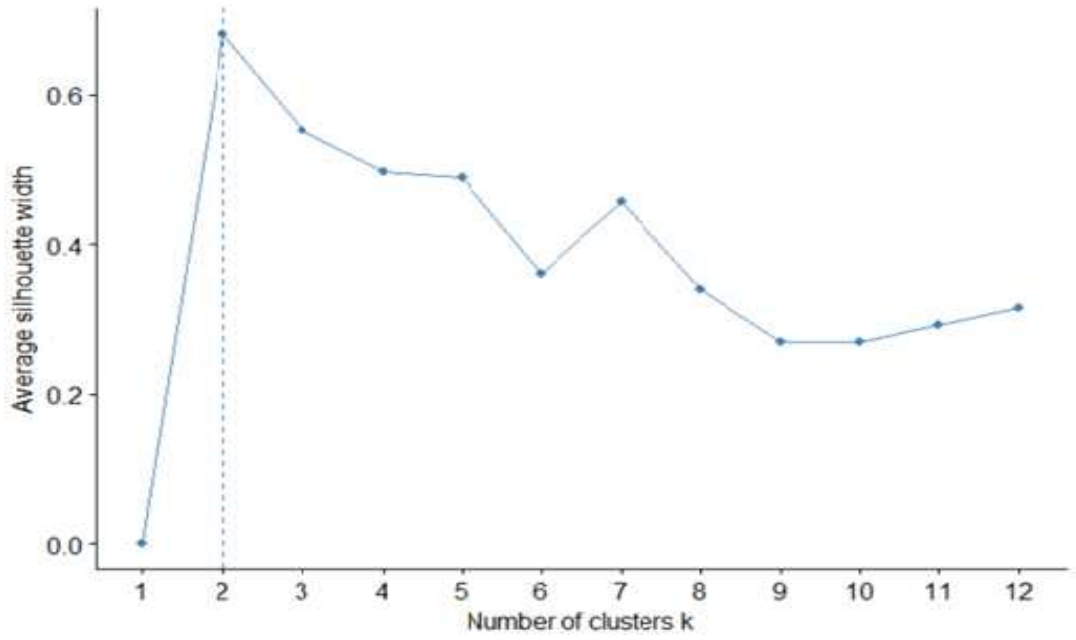

Figure 3. s(i) of different k for the average silhouette method (liver-disorders dataset).

After imputation, the C4.5, MLP, NB, BN, LibSVM classifiers were employed to compute and compare their accuracy. Table 4 shows that the three clusters combined with PkNNI had the best average AUC for all MDs in MAR-type missing values.

Figure 4 shows the no clustering results compared with those with clustering; the three clusters combined with PkNNI had the best AUC, and the five clusters combined with PkNNI had the worst result in different MD. Table 5 shows that the three clusters combined with PkNNI had the best result for all MDs in MCAR. Figure 5 shows the no clustering results compared with those with clustering; the three clusters combined with PkNNI had the best AUC in 5% and 15% MDs, and the no clustering combined with PkNNI had the better AUC in 10% and 20% MD. The best number of clusters combined with PkNNI was the worst result in all different MDs.

Table 4. AUC results of proposed imputation in MAR (liver disorders dataset).

MD	proposed	C4.5	MLP	NB	BN	LibSVM	Average
5%	best k +PkNNI	0.69	0.76	0.66	0.54	0.70	0.67
	best k +DNNI	0.69	0.76	0.65	0.59	0.69	**0.68**
	3 clusters +PkNNI	0.70	0.75	0.65	0.58	0.70	**0.68**
	3 clusters +DNNI	0.67	0.71	0.63	0.58	0.66	0.65
	5 clusters +PkNNI	0.69	0.74	0.66	0.55	0.72	0.67
	5 clusters +DNNI	0.68	0.72	0.66	0.55	0.70	0.66
10%	best k +PkNNI	0.71	0.77	0.65	0.69	0.72	0.71
	best k +DNNI	0.70	0.75	0.65	0.68	0.69	0.69
	3 clusters +PkNNI	0.72	0.79	0.70	0.64	0.74	**0.72**
	3 clusters +DNNI	0.72	0.74	0.67	0.67	0.71	0.70
	5 clusters +PkNNI	0.68	0.73	0.64	0.54	0.70	0.66
	5 clusters +DNNI	0.69	0.72	0.64	0.60	0.68	0.67
15%	best k +PkNNI	0.71	0.78	0.68	0.71	0.73	0.72
	best k +DNNI	0.71	0.74	0.65	0.72	0.66	0.70
	3 clusters +PkNNI	0.73	0.83	0.70	0.68	0.77	**0.74**
	3 clusters +DNNI	0.72	0.79	0.68	0.72	0.73	0.73
	5 clusters +PkNNI	0.69	0.75	0.64	0.54	0.72	0.67
	5 clusters +DNNI	0.70	0.73	0.63	0.66	0.69	0.68
20%	best k +PkNNI	0.73	0.78	0.69	0.69	0.74	0.73
	best k +DNNI	0.73	0.74	0.66	0.74	0.69	0.71
	3 clusters +PkNNI	0.77	0.80	0.68	0.74	0.76	**0.75**
	3 clusters +DNNI	0.78	0.76	0.64	0.76	0.69	0.73
	5 clusters +PkNNI	0.68	0.77	0.67	0.63	0.73	0.70
	5 clusters +DNNI	0.72	0.71	0.66	0.69	0.67	0.69

Note: "best k" denotes the optimal number of clusters where k=2.

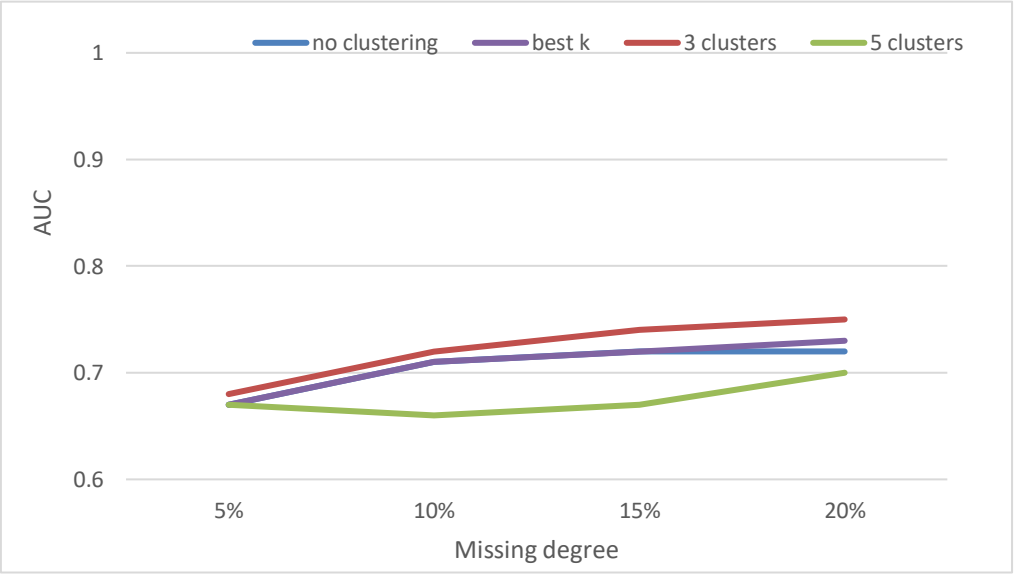

Figure 4 Average AUC of proposed method for different MDs and cluster number in MAR (liver disorders).

Table 5. AUC results of proposed imputation in MCAR (liver disorders dataset).

MD	proposed	C4.5	MLP	NB	BN	LibSVM	Average
5%	best k +PkNNI	0.65	0.73	0.64	0.53	0.70	0.65
	best k +DNNI	0.64	0.71	0.64	0.55	0.69	0.65
	3 clusters +PkNNI	0.65	0.74	0.65	0.55	0.71	**0.66**
	3 clusters +DNNI	0.66	0.73	0.64	0.57	0.70	**0.66**
	5 clusters +PkNNI	0.65	0.73	0.64	0.53	0.64	0.65
	5 clusters +DNNI	0.65	0.71	0.64	0.53	0.68	0.64
10%	best k +PkNNI	0.68	0.74	0.66	0.57	0.66	0.66
	best k +DNNI	0.68	0.72	0.66	0.57	0.64	0.65
	3 clusters +PkNNI	0.68	0.77	0.65	0.61	0.72	**0.69**
	3 clusters +DNNI	0.66	0.72	0.64	0.53	0.65	0.64
	5 clusters +PkNNI	0.68	0.74	0.66	0.56	0.66	0.68
	5 clusters +DNNI	0.65	0.72	0.66	0.57	0.64	0.65
15%	best k +PkNNI	0.68	0.71	0.62	0.65	0.69	0.67
	best k +DNNI	0.73	0.69	0.62	0.66	0.69	0.68
	3 clusters +PkNNI	0.73	0.77	0.65	0.68	0.73	**0.71**
	3 clusters +DNNI	0.69	0.71	0.62	0.69	0.67	0.68
	5 clusters +PkNNI	0.69	0.76	0.65	0.68	0.7	0.69
	5 clusters +DNNI	0.68	0.69	0.62	0.66	0.65	0.66
20%	best k +PkNNI	0.72	0.76	0.61	0.66	0.71	0.69
	best k +DNNI	0.71	0.74	0.60	0.73	0.63	0.68
	3 clusters +PkNNI	0.72	0.77	0.66	0.7	0.71	**0.71**
	3 clusters +DNNI	0.69	0.69	0.62	0.71	0.63	0.67
	5 clusters +PkNNI	0.71	0.79	0.7	0.66	0.72	**0.71**
	5 clusters +DNNI	0.72	0.74	0.68	0.73	0.70	**0.71**

Note: “best k” denotes the optimal number of clusters where k=2.

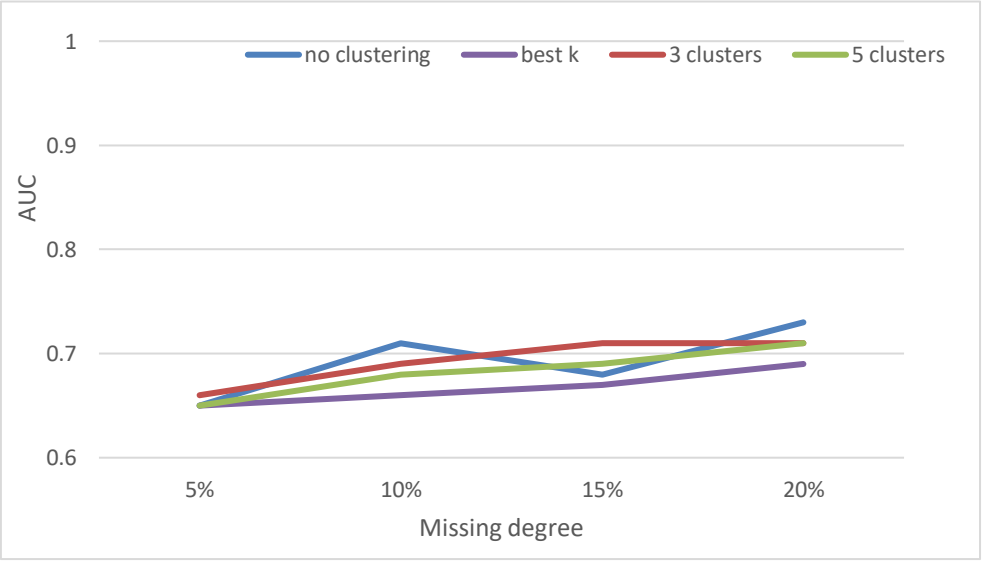

Figure 5. Average AUC of proposed method for different MDs and cluster numbers in MCAR (liver-disorders data)

(2) ILPD dataset

The best number of clusters of the elbow and average silhouette methods were two clusters for the ILPD dataset. Next, we used two (the best number of clusters), three, and five clusters combined with PkNNI and DNNI to impute the missing values in MAR and MCAR types. After imputation, the C4.5, MLP, NB, BN, and LibSVM classifiers were applied to calculate and compare their accuracy. Table 6 shows that the three clusters combined with DNNI had the best AUC for the four MDs in MAR type. Figure 6 indicates that the three clusters combined with DNNI had the best average AUC in 5%, 10%, and 15% MDs.

Table 6. AUC results of proposed imputation for different cluster mumbers in MAR (ILPD).

MD	proposed	C4.5	MLP	NB	BN	LibSVM	Average
5%	best k +PkNNI	0.68	0.76	0.73	0.73	0.50	0.68
	best k +DNNI	0.67	0.73	0.75	0.78	0.50	**0.69**
	3 clusters +PkNNI	0.67	0.74	0.75	0.76	0.50	0.68
	3 clusters +DNNI	0.68	0.73	0.75	0.78	0.50	**0.69**
	5 clusters +PkNNI	0.61	0.74	0.73	0.73	0.50	0.66
	5 clusters +DNNI	0.67	0.75	0.75	0.76	0.50	**0.69**
10%	best k +PkNNI	0.65	0.76	0.72	0.81	0.61	0.71
	best k +DNNI	0.64	0.76	0.72	0.82	0.60	0.71
	3 clusters +PkNNI	0.71	0.79	0.75	0.77	0.63	0.73
	3 clusters +DNNI	0.73	0.81	0.77	0.82	0.61	**0.75**
	5 clusters +PkNNI	0.65	0.76	0.73	0.75	0.52	0.68
	5 clusters +DNNI	0.72	0.77	0.77	0.81	0.55	0.72
15%	best k +PkNNI	0.73	0.81	0.78	0.77	0.59	0.74
	best k +DNNI	0.75	0.82	0.81	0.83	0.60	0.76
	3 clusters +PkNNI	0.73	0.80	0.75	0.82	0.60	0.74
	3 clusters +DNNI	0.77	0.82	0.78	0.84	0.64	**0.77**
	5 clusters +PkNNI	0.72	0.83	0.75	0.77	0.69	0.75
	5 clusters +DNNI	0.75	0.83	0.78	0.83	0.63	0.76
20%	best k +PkNNI	0.76	0.79	0.81	0.81	0.64	0.76
	best k +DNNI	0.77	0.79	0.83	0.87	0.64	**0.78**
	3 clusters +PkNNI	0.76	0.80	0.78	0.79	0.63	0.75
	3 clusters +DNNI	0.77	0.79	0.80	0.86	0.57	0.76
	5 clusters +PkNNI	0.74	0.79	0.76	0.79	0.57	0.73
	5 clusters +DNNI	0.77	0.81	0.81	0.86	0.65	**0.78**

Note: “best k” denotes the optimal number of clusters where k=2.

In MCAR, Table 7 shows that the three clusters combined with PkNNI had a better result for the l0%, 15%, and 20% MDs in MCAR. Figure 7 shows that the three clusters combined with PkNNI had the best average AUC in 10% and 15% MDs.

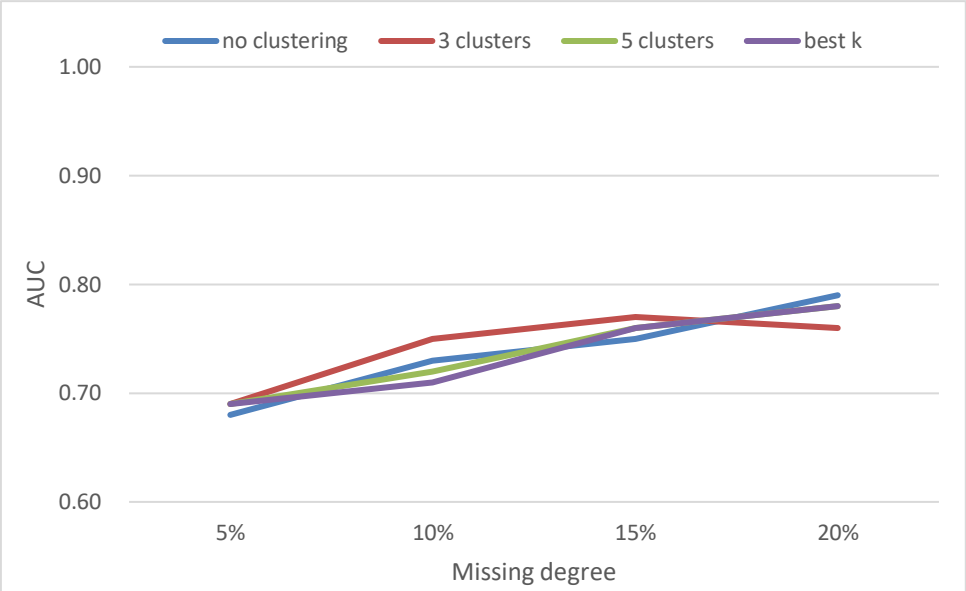

Figure 6. Average AUC of proposed method for different MDs and cluster numbers in MAR (ILPD data).

(3) All seven UCI datasets

We now summarize the results of all datasets in the following.

(a) In MAR, each dataset has 24 average AUC (4 different MDs × 3 different clusters × 2 imputation methods). The comparison is based on the same MD to count the number of wins, but the average AUC of banknote and Acute dataset could not distinguish their differences, then we only made 30 comparison results. The best k clusters won 11 times (included eight even), the three clusters won 12 times (contained six even), and the five clusters succeeded four times (including three even) in four different MDs of the five datasets, as shown in Table A of the appendix. Then, the proposed "three clusters + PkNNI" had better AUC in the seven UCI datasets.

(b) In MCAR, the comparison was made 42 results (3 different clusters × 2 imputation methods × 7 datasets); the best k clusters won 17 times (included 10 even), the three clusters won 20 times (contained 12 even), and the five clusters succeeded 13 times (including 12 even) in four different MD of seven datasets, as shown in Table B of the appendix. Then, the proposed "three clusters + PkNNI" had better accuracy in the seven UCI datasets.

Table 7. AUC results of proposed imputation with the best k, three, and five clusters in MCAR (ILPD).

MD	Proposed	C4.5	MLP	NB	BN	LibSVM	Average
5%	best k +PkNNI	0.69	0.75	0.74	0.75	0.50	0.69
	best k +DNNI	0.68	0.75	0.74	0.76	0.50	0.69
	3 clusters +PkNNI	0.66	0.72	0.74	0.75	0.50	0.67
	3 clusters +DNNI	0.66	0.72	0.76	0.75	0.50	0.68
	5 clusters +PkNNI	0.67	0.75	0.75	0.75	0.53	0.69
	5 clusters +DNNI	0.69	0.76	0.77	0.77	0.51	**0.70**
10%	best k +PkNNI	0.65	0.76	0.76	0.78	0.57	0.70
	best k +DNNI	0.64	0.74	0.77	0.80	0.52	0.69
	3 clusters +PkNNI	0.67	0.76	0.76	0.78	0.56	**0.71**
	3 clusters +DNNI	0.70	0.74	0.78	0.80	0.52	**0.71**
	5 clusters +PkNNI	0.66	0.73	0.74	0.76	0.51	0.68
	5 clusters +DNNI	0.68	0.73	0.77	0.79	0.51	0.70
15%	best k +PkNNI	0.70	0.77	0.78	0.83	0.52	0.72
	best k +DNNI	0.73	0.76	0.78	0.84	0.54	0.73
	3 clusters +PkNNI	0.70	0.78	0.78	0.83	0.61	**0.74**
	3 clusters +DNNI	0.74	0.71	0.75	0.87	0.51	0.72
	5 clusters +PkNNI	0.69	0.76	0.78	0.78	0.51	0.70
	5 clusters +DNNI	0.73	0.76	0.81	0.83	0.54	0.73
20%	best k +PkNNI	0.72	0.81	0.77	0.88	0.69	**0.77**
	best k +DNNI	0.71	0.76	0.79	0.88	0.67	0.76
	3 clusters +PkNNI	0.73	0.79	0.77	0.88	0.59	0.75
	3 clusters +DNNI	0.72	0.73	0.76	0.84	0.51	0.71
	5 clusters +PkNNI	0.74	0.8	0.77	0.85	0.66	0.76
	5 clusters +DNNI	0.71	0.76	0.74	0.83	0.58	0.72

Note: “best k” denotes the optimal number of clusters where k=2.

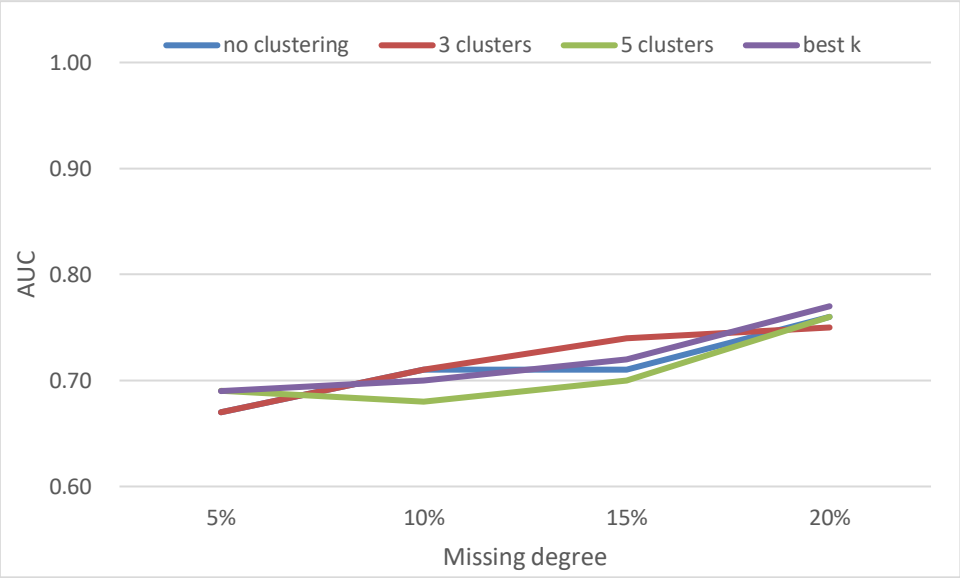

Figure 7. Average AUC of proposed method for different MDs and cluster numbers in MCAR (ILPD data).

4.2.2 Different imputation comparisons

After the internal experiment, most datasets show that no clustering had the worst result. Hence, this section outlines the external experiments to compare the proposed imputation with the listing imputation methods in the different cluster numbers, MDs, and missing types.

(1) AUC metric:

The results are shown in Tables 8 and 9 for the open seven datasets in MAR and MCAR. From the experimental results, we summarize the results of the seven open UCI datasets in the following.

(a) In MAR, the proposed imputation is better than the listing imputation methods for the seven UCI datasets in all different MDs, as shown in Table 8. In addition, the proposed combined clustering with PkNNI is slightly better than the combined clustering with DNNI.

(b) In MCAR, similarly, the proposed method has a better average AUC than the listing imputation methods for the seven UCI datasets in all different MDs, as shown in Table 9. In addition, the proposed combined clustering with PkNNI is better than the combined clustering with DNNI.

Table 8. Average AUC of proposed imputation and listing methods in MAR.

Dataset	MD	PkNNI	DNNI	NPS	OCS	kNN	MI	MICE
Acute inflammations	5%	**1.00**	**1.00**	0.99	**1.00**	0.99	0.98	0.98
	10%	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	15%	**1.00**	**1.00**	0.96	0.98	0.98	0.97	0.97
	20%	**1.00**	**1.00**	0.98	0.97	0.96	0.96	0.98
Banknote	5%	**0.99**	**0.99**	0.96	0.96	0.97	0.97	0.97
	10%	**0.99**	**0.98**	0.95	0.95	0.95	0.95	0.94
	15%	**0.98**	**0.98**	0.93	0.93	0.94	0.93	0.93
	20%	**0.98**	**0.99**	0.92	0.92	0.92	0.92	0.92
Blood transfusion	5%	**0.71**	**0.71**	0.68	0.68	0.67	0.68	0.68
	10%	**0.72**	**0.72**	0.67	0.67	0.67	0.67	0.67
	15%	**0.76**	**0.76**	0.67	0.67	0.67	0.67	0.67
	20%	**0.78**	**0.78**	0.67	0.67	0.66	0.67	0.67
Climate model	5%	**0.82**	0.81	0.79	0.79	0.78	0.79	0.79
	10%	**0.85**	0.84	0.78	0.78	0.77	0.78	0.78
	15%	0.81	**0.82**	0.78	0.78	0.78	0.78	0.77
	20%	0.82	**0.86**	0.77	0.77	0.77	0.77	0.77
Haberman survival	5%	**0.63**	0.61	0.61	0.60	0.60	0.60	0.60
	10%	**0.64**	0.61	0.60	0.60	0.60	0.60	0.60
	15%	**0.63**	0.65	0.60	0.62	0.60	0.60	0.60
	20%	0.66	**0.67**	0.62	0.62	0.61	0.61	0.61
ILPD	5%	0.68	**0.69**	0.54	0.54	0.54	0.54	0.54
	10%	0.73	**0.75**	0.54	0.54	0.54	0.54	0.54
	15%	0.75	**0.77**	0.54	0.54	0.54	0.54	0.54
	20%	0.76	**0.78**	0.54	0.54	0.54	0.54	0.54
Liver-disorders	5%	**0.68**	**0.68**	0.62	0.62	0.61	0.62	0.61
	10%	**0.72**	0.70	0.59	0.60	0.59	0.58	0.59
	15%	**0.74**	0.73	0.60	0.60	0.58	0.59	0.59
	20%	**0.75**	0.73	0.59	0.59	0.60	0.60	0.60

Note: where the best average AUC is listed for each MD.

Table 9. Average AUC of proposed imputation and listing method in MCAR

Dataset	MD	PkNNI	DNNI	NPS	OCS	kNN	MI	MICE
Acute inflammations	5%	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	10%	**1.00**	**1.00**	**1.00**	**1.00**	0.99	0.99	0.99
	15%	**1.00**	**1.00**	**1.00**	**1.00**	0.99	0.99	0.99
	20%	**1.00**	**1.00**	0.99	0.99	0.97	0.98	0.98
Banknote	5%	**0.98**	**0.98**	0.97	0.97	**0.98**	0.97	0.97
	10%	**0.98**	**0.98**	0.96	0.96	0.96	0.96	0.96
	15%	**0.98**	**0.98**	0.95	0.95	0.95	0.94	0.94
	20%	**0.98**	**0.98**	0.94	0.94	0.94	0.93	0.93
Blood transfusion	5%	**0.71**	**0.71**	0.67	0.67	0.67	0.68	0.68
	10%	**0.72**	**0.72**	0.67	0.67	0.67	0.67	0.67
	15%	0.73	**0.74**	0.66	0.66	0.67	0.67	0.67
	20%	**0.77**	**0.77**	0.66	0.65	0.66	0.66	0.66
Climate model	5%	**0.81**	0.80	0.79	0.79	0.78	0.78	0.78
	10%	**0.83**	0.80	0.78	0.77	0.76	0.77	0.77
	15%	**0.84**	0.79	0.76	0.76	0.76	0.76	0.76
	20%	**0.82**	0.77	0.75	0.75	0.72	0.75	0.74
Haberman survival	5%	**0.63**	0.62	0.61	0.61	0.61	0.61	0.61
	10%	**0.62**	0.61	0.61	0.61	0.60	0.60	0.60
	15%	**0.64**	0.62	0.61	0.61	0.61	0.60	0.60
	20%	**0.63**	0.62	0.61	0.60	0.60	0.60	0.60
ILPD	5%	0.69	**0.70**	0.54	0.54	0.54	0.54	0.54
	10%	**0.71**	**0.71**	0.54	0.54	0.53	0.54	0.54
	15%	**0.74**	0.73	0.54	0.54	0.53	0.53	0.53
	20%	**0.77**	0.76	0.54	0.54	0.54	0.54	0.53
Liver disorders	5%	**0.66**	**0.66**	0.60	0.60	0.59	0.60	0.60
	10%	**0.69**	0.65	0.60	0.60	0.59	0.60	0.60
	15%	**0.71**	0.68	0.57	0.57	0.56	0.56	0.56
	20%	**0.71**	**0.71**	0.58	0.58	0.58	0.57	0.57

Note: where the best average AUC is listed for each MD.

(2) RMSE metric

To verify the performance of the proposed imputation, we used RMSE as the other evaluation criterion. From the MAR and MCAR experiments, we only experimented on clustering combined with the different imputation methods in RMSE evaluation. In MAR and MCAR, we repeated 10 times to randomly remove some data values on the basis of different

MDs in each UCI dataset. Then, we could calculate the error between actual values and imputation values. In each dataset, we took the minimal average RMSE from 12 average RMSE (4 different MDs × 3 different cluster numbers) for each imputation method in MAR and MCAR, where the average RMSE was the average of 10 imputed datasets for each MD. The minimal RMSE of seven imputation methods in MAR and MCAR is shown in Table 10. In MAR, we could see that the combined clustering with PkNNI imputation had the best performance (minimal average RMSE) in seven imputation methods, as shown in Table 10 and Figure 8. In MCAR, the combined clustering with DNNI imputation had the minimal average RMSE in the seven imputation methods, as shown in Table 10 and Figure 9.

Table 10. Comparison of the minimal average RMSE in MAR and MCAR

Dataset	**DNNI**	**PkNNI**	**NPS**	**OCS**	**MI**	**MICE**	**kNN**
			MAR				
Acute Inflammations	0.0889	**0.0883**	0.7143	0.69475	0.2804	0.3042	**0.1009**
Banknote	0.4178	**0.4070**	3.2272	3.0593	0.4186	0.4179	0.5441
Blood	**1.0392**	**1.0392**	5.5560	5.5463	1.9776	2.2471	**1.1858**
Climate model	**0.2372**	**0.2372**	0.2886	0.2885	1.6205	1.6328	1.6983
Haberman survival	**3.3265**	**3.3265**	13.6485	13.6005	4.7795	4.7767	4.3562
ILPD	88.7349	**88.5346**	177.7185	171.8210	89.7255	93.2407	290.6026
Liver disorders	0.2816	**0.2815**	19.4455	18.6196	0.2864	0.2941	0.3296
			MCAR				
Acute inflammations	0.2541	**0.2441**	0.8229	0.7607	**0.2875**	0.2958	0.3311
Banknote	**3.0017**	**3.0017**	3.2178	3.0992	10.2733	10.2629	12.6589
Blood transfusion	**7.5191**	7.5292	732.0909	738.0240	7.9024	7.8543	10.8998
Climate model	**0.2369**	0.2373	**0.2919**	0.2893	109.6393	109.3239	120.3876
Haberman survival	**7.3265**	**7.3265**	9.9404	9.9862	106.5238	116.7198	131.5385
ILPD	**10.9346**	10.9846	138.8616	135.2440	11.0020	11.0953	12.8745
liver-disorders	**10.7165**	10.7166	17.2383	17.2417	10.9865	10.8172	13.1283

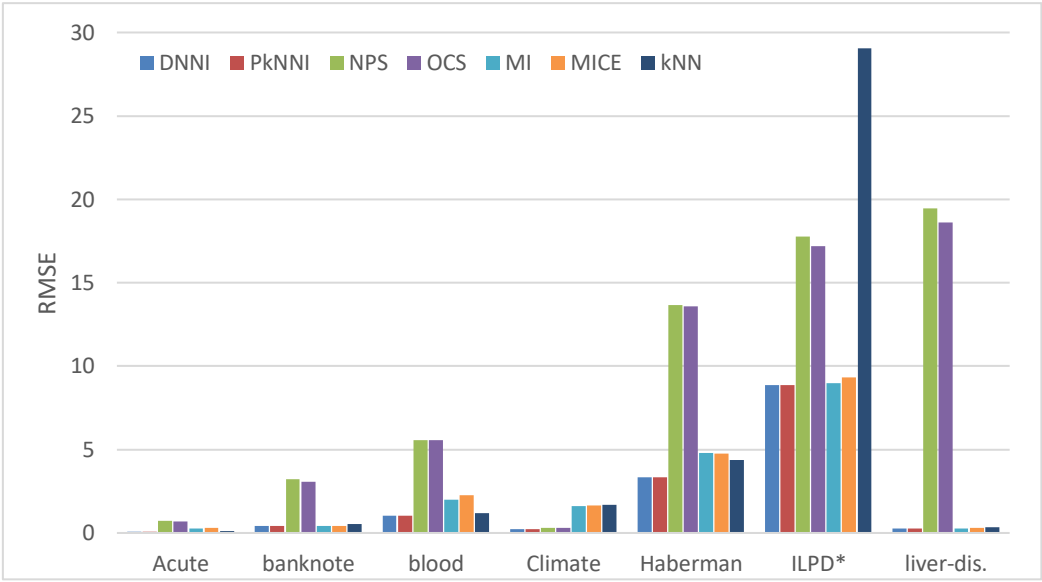

Figure 8. Minimal RMSE of seven imputation methods in MAR.

Note: ILPD* denotes that the RMSE of ILPD dataset is processed by (RMSE/10) to plot this figure because it has a large RMSE.

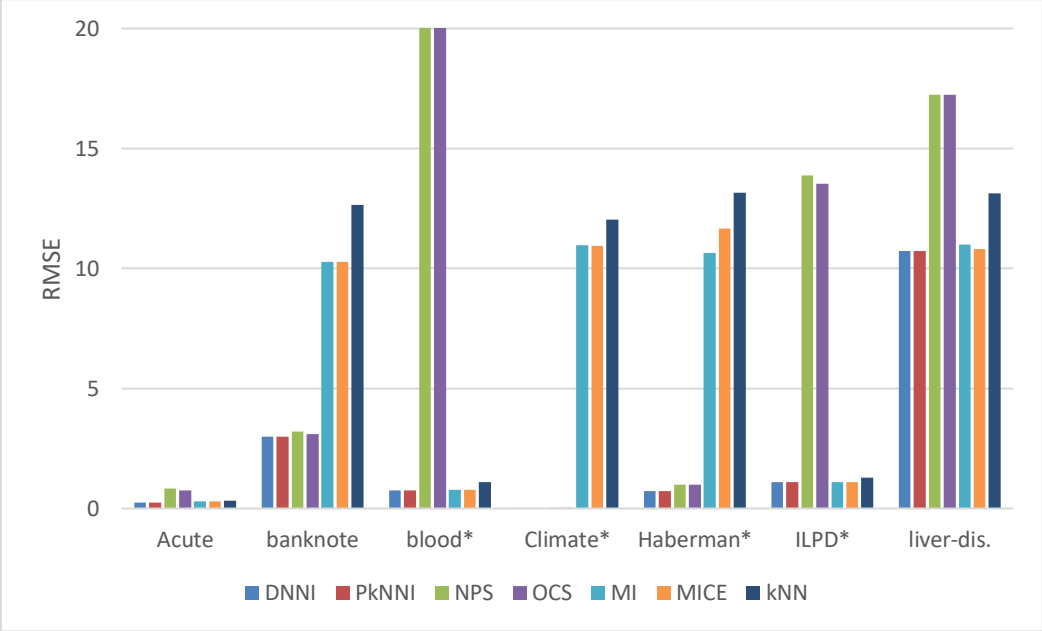

Figure 9. Minimal RMSE of seven imputation methods in MCAR.

Note: blood*, Climate*, Haberman*, and ILPD* denote that the RMSE of the four datasets is processed by (RMSE/10) to plot the figure because they have a large RMSE.

4.3 MNAR experiment

The MNAR refers to the patients who are not willing to provide the relevant data to the physician due to her privacy, and the missing values are subject to the unobserved patient attributes. The stroke dataset itself had missing values that belonged to the MNAR type. This

study directly clustered the complete dataset and imputed the incomplete dataset, and then classified the imputed dataset by five classifiers to verify the proposed imputation. From Table 3, the stroke dataset is balanced classes, and this study uses classification accuracy and AUC to measure comparison results.

In an internal comparison, the results of the stroke dataset were shown in Table 11, indicating that the three clusters combined with DNNI had the best average accuracy. In no clustering and different cluster numbers, Figure 10 shows that DNNI was better than PkNNI in the average accuracy of five classifiers, and clustering imputation was better than no clustering imputation. In an external comparison, Table 12 shows that the proposed DNNI imputation had the best average accuracy and AUC among the seven imputation methods

Table 11. Proposed method in MNAR for stroke dataset (best k and three, and five clusters).

Method	C4.5	MLP	NB	BN	LibSVM	Average
best k+PkNNI	91.33	80.60	65.24	65.96	75.40	75.71
best k+DNNI	94.05	83.02	65.16	78.48	74.25	78.99
3 clusters+PkNNI	91.35	80.98	65.17	65.05	75.01	75.51
3 clusters+DNNI	94.00	83.23	65.25	78.17	74.55	**79.04**
5 clusters+PkNNI	91.67	81.05	65.28	65.36	75.54	75.78
5 clusters+DNNI	94.29	81.97	65.64	78.65	74.55	79.02

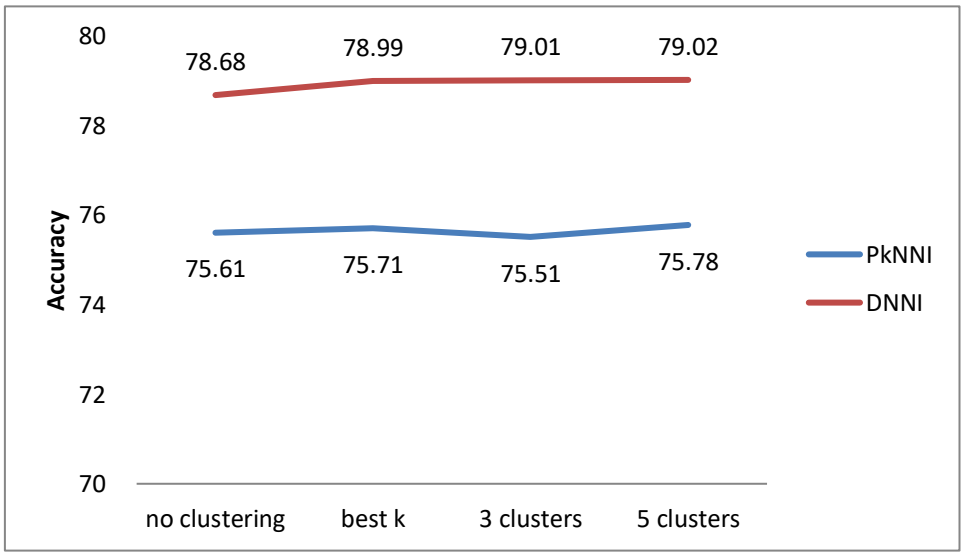

Figure 10. Average of the accuracy of the proposed method for different number of clusters in MNAR (stroke dataset)

Table 12. Average accuracy of proposed and the listing methods in MNAR (stroke dataset)

	PkNNI	DNNI	NPS	OCS	MICE	MI	kNN
Best average accuracy	75.78	**79.04**	70.58	70.66	70.50	70.53	70.64
Best average AUC	0.66	**0.69**	0.64	0.64	0.64	0.63	0.64

5 Discussions and findings

After the MAR, MCAR, and MNAR experiments, we provide some discussions and findings as follows.

(A) Cluster numbers effect

After determining the number of clusters by the elbow and the average silhouette methods, the best number of clusters in the collected eight datasets is two clusters. However, all best AUC and RMSE do not belong to two clusters from internal and different imputation comparisons. Figures 4-7 and 10 show that the combined clustering with PkNNI and DDNI can obtain better AUC than no clustering for the collected datasets. Overall, the best combination of the proposed imputation is the combined three clusters with PkNNI because the different datasets have different numbers of classes, data properties, and disadvantages of k-means. The disadvantages of k-means clustering are (1) selecting appropriate k, (2) dependent on initial values, (3) outliers impacting bias, and (4) assuming all variables have the same variance.

Similarly, FCM-based imputation also has some disadvantages. For example, OCS [30-31] used initial values and available feature values to calculate FCM cluster prototypes, and the missing values were estimated based on these biased cluster prototypes. Hence, the computation of the FCM cluster prototype and the imputation of the missing value would influence each other. In addition, Li et al. [44] developed a fuzzy clustering algorithm based on the nearest neighbor interval (FCM-NNI) to enhance the performance. Therefore, we suggest that FCM-based imputation can add some methods to enhance its performance.

(B) RMSE metric

The selected seven UCI datasets had no missing values, and then this study simulated four different MDs to handle missing values. Therefore, we can use RMSE to evaluate the performance of different imputation methods because the seven UCI datasets have actual values to calculate the RMSE. We experimented on the combined clustering with different imputation methods in RMSE evaluation, as shown in Table 10. For MAR, we can see that the combined clustering with PkNNI imputation had the minimal average RMSE in seven imputation methods.

In MCAR, the combined clustering with DNNI imputation had the minimal average RMSE in the seven imputation methods. Results showed that the proposed imputation was better than the listed imputation methods because clustering would aggregate similar data points in one cluster.

From Table 10 and Figures 8-9, we see that the imputation method has large RMSE in MAR of ILPD dataset because the attribute values of age ranged from 4 to 90 and its average was 45, and the attribute values of Alkphos Alkaline Phosphatase range from 62 to 2110 and its average=208. Similarly, the NPS and OCS imputations have large RMSE in MCAR of ILPD and Blood transfusion dataset because the ILPD attributes range as above mentioned and the attribute range of the Blood transfusion dataset are listed as follows: recency (months since the last donation) range is [0, 74] and average=9.5; frequency (total number of donation) range is [1, 50] and average=5.515; monetary (total blood donated in c.c.) range is [250, 12500] and average=1378.676; time (months since the first donation) range is [2, 98] and average=34.282.

(C) MNAR imputation

MNAR missing values is a non-ignorable non-response, and its data is neither MAR nor MCAR. That is, the missing values of the variable have their reason based on privacy [45]. The stroke dataset is collected by questionnaire of stroke patients, which had missing values that belonged to MNAR type. In MNAR, we could not calculate their RMSE because the missing values did not have actual values, and this study only used accuracy and AUC to evaluate their performance. After the experiments, the proposed imputation (three clusters + DNNI) had the best average accuracy and AUC from the seven imputation methods in the stroke dataset. The stroke dataset has many category attributes, but the proposed imputation also has the best performance, indicating that the proposed imputation method is viable. In this experiment, the best performance is three clusters combined with DNNI because the best fitting clusters combining with the imputation methods is an important factor.

(D) Multiple-combined imputation methods

In recent years, there are many hybrid imputation methods. Zhang et al. [46] proposed a flexibly combines three techniques: self-organizing feature map clustering, the fruit fly optimization algorithm, and the least squares support vector machine to impute spatiotemporal missing values. Dubey and Rasool [47] presented the combined k-means clustering with the weighted KNN to impute the missed value, and their results outperformed mean substitution and FCM imputation. This study proposed the combined k-means clustering with PkNNI and DNNI imputation, and the experimental results showed that the proposed imputation was better than the listing imputation methods in the eight datasets. We find that the combined local similarity structure of the dataset (using k-means, or self-organizing map, or FCM clustering)

with an imputation method can enhance the performance. Therefore, we suggest that FCM, MI, MICE, and kNN imputation methods can be added some advanced techniques to strengthen their imputation ability.

6. Conclusion

This study has proposed clustering-based purity and distance imputation methods to improve performance. After MAR, MCAR, and MNAR experiments, the proposed imputation was found to be better than the other imputation methods in seven UCI datasets and a stroke dataset, except for the acute inflammations dataset. In the RMSE of MAR and MCAR experiments, the combined clustering with PkNNI (DNNI) imputation had the minimal average RMSE in the seven imputation methods. Results showed that the proposed imputation was better than the listed imputation methods, mainly because clustering would aggregate similar data points in one cluster. To obtain the optimal number of clusters, we applied the elbow method and average silhouette method to obtain a consistent optimal number of clusters, and the best cluster number was two in the eight datasets. The number of classes of the seven datasets was two classes except for the acute inflammations dataset with four classes, as shown in Table 3, and the acute inflammations dataset almost had category attributes. From the results and findings, we summarized the contribution and applicability of the proposed method as follows.

(1) The proposed imputation is a hybrid method, and the attributes data do not obey multivariate normal distribution. *i.e.*, the proposed approach is combined k-mean with the two kNN related approaches. The main difference between the two approaches (PkNNI and DNNI) and the proposed method is to find the optimal number of clusters and adapt the two kNN related approaches to obtain the optimal results for the eight medical datasets. Because PkNNI with two parameters needs to train for optimal results, DNNI with different averages and thresholds must adapt to obtain the optimal results.

(2) Medical data have the higher MDs; hence the research works need to check a higher percentage of complete data (e.g., normally more than 50% of the total cases) and needed to impute less than three values per patient [6]. However, this study simulated 360% MD (MD=20% in this paper) to handle missing values. Especially in the practical stroke dataset, there were 69 attributes with 4242 instances and 6578 missing values; hence traditional MD is 155% (MD=2.25% in this paper). In addition, the largest MD of the single attribute was 95% and needed to impute 14 values per patient.

(3) In selecting the imputation method, Sim et al. [48] suggested, according to the

characteristics of the dataset (especially the patterns of missing values). This study proposed a hybrid imputation to handle missing values in MAR, MCAR, and MNAR types; the different imputation method has its advantages and disadvantages; hence a set of optimal combinations may be derived using the estimated results.

(4) How to evaluate imputation performance is an important issue. Most previously used accuracy and RMSE, but medical data usually are imbalanced classes; this paper applied accuracy, RMSE, and AUC to overcome the problem of imbalanced classes.

In future work, we can utilize the frequency-based or neighbor-count imputation to improve the performance of the category dataset and combine it with other model-free imputation methods to conduct massive experiments.

Compliance with Ethical Standards

Funding: The authors received no financial support for the research.

Conflict of interest: The authors declare that they have no conflicts of interest.

Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Schafer J.L. (1997), Analysis of incomplete multivariate data, New York, Chapman & Hall.
2. Amiri M., Jensen R. (2016), Missing data imputation using fuzzy-rough methods, Neurocomputing, 205 152-164.
3. Donders A.R., van der Heijden G.J., Stijnen T., Moons K.G. (2006), Review: a gentle introduction to imputation of missing values, Journal of Clinical Epidemiology, 59, 1087-1091.
4. Ondeck N.T., Fu M.C., Skrip L.A., McLynn R.P., Su E.P., Grauer J.N. (2018), Treatments of Missing Values in Large National Data Affects Conclusions: the Impact of Multiple Imputation on Arthroplasty Research, The Journal of Arthroplasty, 33(3), 661-667.
5. Jerez J.M., Molina I., Garcia-Laencina P.J., Alba E., Ribelles N., Martin M., Franco L. (2010), Missing data imputation using statistical and machine learning methods in a real breast cancer problem, Artif. Intell. Med., 50 (2), 105-115.
6. Kharrazi H., Wang C., Scharfstein D. (2014). Prospective EHR-based clinical trials: the challenge of missing data, J. Gen. Intern. Med., 29 (7) (2014), pp. 976-978.
7. M€uhlenbruch K., Kuxhaus O., Giuseppe R.d., Boeing H., Weikert C., Schulze M.B. (2017), Multiple imputation was a valid approach to estimate absolute risk from a

prediction model based on caseecohort data, Journal of Clinical Epidemiology, 84, 130-141.

8. Enders C.K. (2017), Multiple imputation as a flexible tool for missing data handling in clinical research, Behaviour Research and Therapy, 98, 4-18.
9. Galan C.O., Lasheras F.S., Juez F.J.de, Sanchez A.B. (2017), Missing data imputation of questionnaires by means of genetic algorithms with different fitness functions, Journal of Computational and Applied Mathematics, 311, 704–717.
10. Sterne J., White I.R., Carlin J.B., Spratt M., Royston P., Kenward M.G., Wood A.M., Carpenter J.R. (2009), Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls, BMJ. BMJ, Vol. 338, 157 - 160.
11. Jerez J.M., Molina I., Subirats J.L., Franco L. (2006), Missing Data Imputation in Breast Cancer Prognosis, Proceedings of the 24th IASTED international conference on Biomedical engineering, p.323-328, February 15-17, 2006, Innsbruck, Austria.
12. García-Laencina P.J., Abreu P.H., Abreu M.H., Afonoso N. (2015), Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values, Computers in biology and medicine 59, 125-133
13. Pombo N., Rebelo P., Araújo P., Viana J. (2015), Combining Data Imputation and Statistics to Design a Clinical Decision Support System for Post-Operative Pain Monitoring, Procedia Computer Science, 64, 1018-1025.
14. Pombo N., Rebelo P., Araújo P., Viana J. (2016), Design and evaluation of a decision support system for pain management based on data imputation and statistical models, Measurement, 93, 480-489.
15. Rubin D.B. (1976), Inference and Missing Data, BIOMETRIKA, 63, 581–90.
16. Wagstaff K. (2004), Clustering with Missing Values: No Imputation Required. In: Banks D., McMorris F.R., Arabie P., Gaul W. (eds) Classification, Clustering, and Data Mining Applications. Studies in Classification, Data Analysis, and Knowledge Organisation. Springer, Berlin, Heidelberg
17. Forgy E.W. (1965). "Cluster analysis of multivariate data: efficiency versus interpretability of classifications". Biometrics. 21 (3), 768–769.
18. Ketchen D.J., Shook C.L., The application of cluster analysis in Strategic Management Research: an analysis and critique, Strategic Management Journal, 17 (6), 1996, 441-458.
19. Rousseeuw P.J. (1987), Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics, 20, 53-65.
20. Zhang Z. (2016), Multiple imputation with multivariate imputation by chained Equation (MICE) package. Ann Transl Med.;4(2):30.
21. Ondeck N.T., Fu M.C., Skrip L.A., McLynn R.P., Su E.P., Grauer J.N. (2018), Treatments

of Missing Values in Large National Data Affects Conclusions: the Impact of Multiple Imputation on Arthroplasty Research, The Journal of Arthroplasty, 33 661-667.

22. Batista G.E., Monard M.C. (2003), An analysis of four missing data treatment methods for supervised learning, Applied Artificial Intelligence, 17, 519-533.
23. Troyanskaya O., Cantor M., Sherlock G., Brown P., Hastie T., Tibshirani R., Botstein D., Altman R.B. (2001), Missing value estimation methods for DNA microarrays, Bioinformatics, 17, 520-525.
24. Keerin, P., Kurutach, W., Boongoen, T. (2016), A cluster-directed framework for neighbour based imputation of missing value in microarray data, International Journal of Data Mining and Bioinformatics 15 (2), 165-193.
25. Lee, J.Y. and Styczynski, M.P. (2018), NS-kNN: a modified k-nearest neighbors approach for imputing metabolomics data, Metabolomics 14, 1-12.
26. Cheng C.H., Chan C.P., Sheu Y.J. (2019), A novel purity-based k nearest neighbors imputation method and its application in financial distress prediction, Engineering Applications of Artificial Intelligence 81, 283–299
27. Cheng C.H., Chang J.R.Huang H.H. (2020), A novel weighted distance threshold method for handling medical missing values, Computers in Biology and Medicine, 2020, 122, 103824.
28. Lin, W.-C. and Tsai, C.-F. Missing Value Imputation: A Review and Analysis of the Literature (2006 - 2017). Artificial Intelligence Review, vol. 53 (2020), pp. 1487-1509.
29. Awan S.E., Bennamoun M., Sohel F., Sanfilippo F.M., Dwivedi G. (2021), Imputation of Missing Data with Class Imbalance using Conditional Generative Adversarial Networks, Neurocomputing, doi: https://doi.org/10.1016/j.neucom.2021.04.010.
30. Hathaway R. J., Bezdek J. C. (2001), "Fuzzy c-means clustering of incomplete data," in IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 31, no. 5, pp. 735-744, Oct. 2001, doi: 10.1109/3477.956035.
31. Al Shami A., Lotfi A., Coleman S. (2013), Intelligent synthetic composite indicators with application, Soft Comput 17, 2349–2364. https://doi.org/10.1007/s00500-013-1098-3.
32. Dinh D-T., Huynh V-N., Sriboonchitta S. (2021), Clustering Mixed Numerical and Categorical Data with Missing Values, Information Sciences, doi: https://doi.org/10.1016/j.ins.2021.04.076.
33. Andridge R.R., Little R.J.A. (2010), A Review of Hot Deck Imputation for Survey Non-response, 78, 40-64.
34. Shao J. (2000), Cold deck and ratio imputation, Survey Methodology, 26, 79-85.
35. Jerez J.M., Molina I., Garcia-Laencina P.J., Alba E., Ribelles N., Martin M., Franco L. (2010), Missing data imputation using statistical and machine learning methods in a real

breast cancer problem, Artif Intell Med, 50, 105-115.
36. Sandercock, P.A., Niewada, M. & Członkowska, A. (2011), The International Stroke Trial database. Trials 12, 101.
37. Moayedikia A., Ong K.L., Boo Y.L., Yeoh W.G., Jensen R. (2017), Feature selection for high dimensional imbalanced class data using harmony search, Eng. Appl. Artif. Intell., 57, 38-49.
38. Sammut C., Webb G.I. (2010), Encyclopedia of Machine Learning, Springer Publishing Company, Boston, MA.
39. Quinlan, J.R. (1992), C4.5 Programs for Machine Learning, San Mateo, CA: Morgan Kaufmann.
40. Mitra S., Pal S.K. (1995), Fuzzy multi-layer perceptron, inferencing and rule generation, IEEE Transactions on Neural Networks, 6, 51-63.
41. John G.H., langley P. (1995), Estimating Continuous Distributions in Bayesian Classifiers, In proceedings of the eleventh conference on uncertainty in artificial intelligence, pp. 338-345, San Mateo, CA: Morgan Kaufmann
42. Pearl J., Russell S. (2000), Bayesian Networks, TR R-277, University of California
43. Chang C.-C., Lin C.-J. (2011), LIBSVM: A library for support vector machines, ACM Transactions on Intelligent Systems and Technology (TIST), 2 (3), 1-27.
44. Li, D., Gu, H., & Zhang, L. (2010), A fuzzy c-means clustering algorithm based on nearest-neighbor intervals for incomplete data, Expert Systems with Applications, vol. 37 (10), 6942-6947.
45. Polit D.F., Beck C.T. (2012), Nursing Research: Generating and Assessing Evidence for Nursing Practice, 9th ed. Philadelphia, USA: Wolters Klower Health, Lippincott Williams & Wilkins.
46. Zhang Z., Yang X., Li H., Li W., Yan H., Shi F. (2017), Application of a novel hybrid method for spatiotemporal data imputation: A case study of the Minqin County groundwater level, Journal of Hydrology, 553, 384-397.
47. Dubey A., Rasool A. (2020), Clustering-Based Hybrid Approach for Multivariate Missing Data Imputation. International Journal of Advanced Computer Science and Applications, 11 (11), doi:10.14569/IJACSA.2020.0111186.
48. Sim J., Lee J.S., Kwon O. (2015), Missing values and optimal selection of an imputation method and classification algorithm to improve the accuracy of ubiquitous computing applications, Mathematical Problems in Engineering, 2015 (12) (2015), pp. 1-14

Appendix:

Table A. Internal comparison for the other UCI datasets in MAR

MD	Method	Banknote	Blood	Climate	Acute	Habermans	Liver	ILPD
5%	best k+PkNNI	**0.98**	**0.71**	**0.81**	**1.00**	**0.62**	0.67	0.68
	best k+DNNI	**0.98**	**0.71**	**0.81**	**1.00**	0.61	**0.68**	**0.69**
	3 clusters+PkNNI	**0.98**	**0.71**	**0.81**	**1.00**	**0.62**	**0.68**	0.68
	3 clusters+DNNI	**0.98**	**0.71**	**0.81**	**1.00**	0.60	0.65	**0.69**
	5 clusters+PkNNI	**0.98**	**0.71**	0.80	**1.00**	**0.62**	0.67	0.66
	5 clusters+DNNI	**0.98**	**0.71**	0.80	**1.00**	0.61	0.66	**0.69**
10%	best k+PkNNI	**0.98**	**0.73**	**0.86**	**1.00**	0.64	0.71	0.71
	best k+DNNI	**0.98**	**0.73**	0.84	**1.00**	0.64	0.69	0.71
	3 clusters+PkNNI	**0.98**	0.71	0.85	**1.00**	**0.65**	**0.72**	0.73
	3 clusters+DNNI	**0.98**	0.72	0.84	**1.00**	0.60	0.70	**0.75**
	5 clusters+PkNNI	**0.98**	0.71	0.85	**1.00**	0.64	0.66	0.68
	5 clusters+DNNI	**0.98**	0.72	0.80	**1.00**	0.61	0.67	0.72
15%	best k+PkNNI	**0.98**	**0.76**	0.83	**1.00**	0.63	0.72	0.74
	best k+DNNI	**0.98**	0.74	0.81	**1.00**	0.63	0.70	0.76
	3 clusters+PkNNI	**0.98**	0.75	**0.87**	**1.00**	0.63	**0.74**	0.74
	3 clusters+DNNI	**0.98**	**0.76**	0.82	**1.00**	**0.64**	0.73	**0.77**
	5 clusters+PkNNI	**0.98**	0.74	**0.87**	**1.00**	0.63	0.67	0.75
	5 clusters+DNNI	**0.98**	0.74	0.81	**1.00**	**0.64**	0.68	0.76
20%	best k+PkNNI	**0.98**	**0.78**	**0.92**	**1.00**	0.65	0.73	0.76
	best k+DNNI	**0.98**	**0.78**	0.86	**1.00**	0.65	0.71	**0.78**
	3 clusters+PkNNI	**0.98**	0.77	**0.92**	**1.00**	0.66	**0.75**	0.75
	3 clusters+DNNI	**0.98**	0.77	0.86	**1.00**	0.67	0.73	0.76
	5 clusters+PkNNI	**0.98**	0.76	0.88	**1.00**	0.66	0.70	0.73
	5 clusters+DNNI	**0.98**	0.77	0.83	**1.00**	**0.68**	0.69	**0.78**

Table B Internal comparison for the other UCI datasets in MCAR

MD	Method	Banknote	Blood	Climate	Acute	Habermans	Liver	ILPD
5%	best k+PkNNI	**0.98**	**0.70**	0.80	**1.00**	**0.63**	0.65	0.69
	best k+DNNI	**0.98**	**0.70**	0.79	**1.00**	0.62	0.65	0.69
	3 clusters+PkNNI	**0.98**	**0.70**	**0.81**	**1.00**	**0.63**	**0.66**	0.67
	3 clusters+DNNI	0.97	**0.70**	0.80	**1.00**	0.62	**0.66**	0.68
	5 clusters+PkNNI	**0.98**	0.69	0.80	**1.00**	**0.63**	0.65	0.69
	5 clusters+DNNI	0.97	**0.70**	0.79	**1.00**	**0.63**	0.64	**0.70**
10%	best k+PkNNI	**0.98**	0.71	**0.83**	**1.00**	**0.61**	0.66	0.70
	best k+DNNI	**0.98**	**0.72**	0.79	**1.00**	0.60	0.65	0.69
	3 clusters+PkNNI	**0.98**	0.71	0.82	**1.00**	**0.61**	**0.69**	**0.71**
	3 clusters+DNNI	**0.98**	0.71	0.79	**1.00**	0.60	0.64	**0.71**
	5 clusters+PkNNI	**0.98**	0.71	0.82	**1.00**	**0.61**	0.68	0.68
	5 clusters+DNNI	0.97	0.71	0.80	**1.00**	**0.61**	0.65	0.70
15%	best k+PkNNI	**0.98**	0.74	**0.84**	0.99	**0.65**	0.67	0.72
	best k+DNNI	**0.98**	**0.75**	0.78	0.99	0.64	0.68	0.73
	3 clusters+PkNNI	**0.98**	0.73	0.81	**1.00**	0.62	**0.71**	**0.74**
	3 clusters+DNNI	0.97	0.74	0.79	**1.00**	0.62	0.68	0.72
	5 clusters+PkNNI	**0.98**	0.73	0.83	**1.00**	0.64	0.69	0.70
	5 clusters+DNNI	0.97	0.73	0.79	**1.00**	0.62	0.66	0.73
20%	best k+PkNNI	**0.98**	0.76	0.82	**1.00**	**0.63**	0.69	**0.77**
	best k+DNNI	**0.98**	0.76	0.77	**1.00**	0.60	0.68	0.76
	3 clusters+PkNNI	**0.98**	**0.77**	**0.84**	**1.00**	0.62	**0.71**	0.75
	3 clusters+DNNI	**0.98**	**0.77**	0.77	**1.00**	0.60	0.67	0.71
	5 clusters+PkNNI	**0.98**	0.76	0.82	**1.00**	0.62	**0.71**	0.76
	5 clusters+DNNI	**0.98**	0.76	0.77	**1.00**	0.62	**0.71**	0.72

Figures

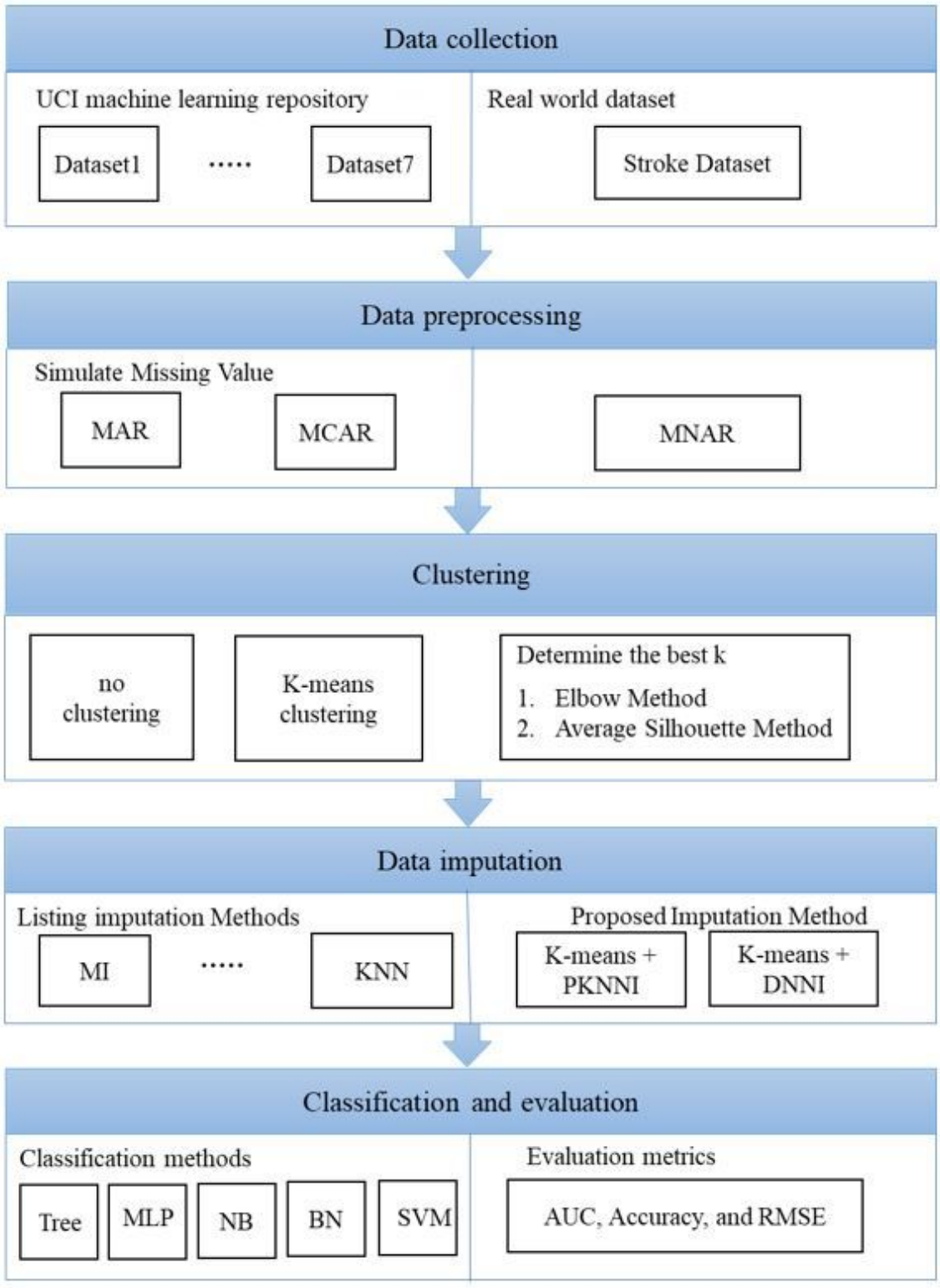

Figure 1

Computational procedure of this study

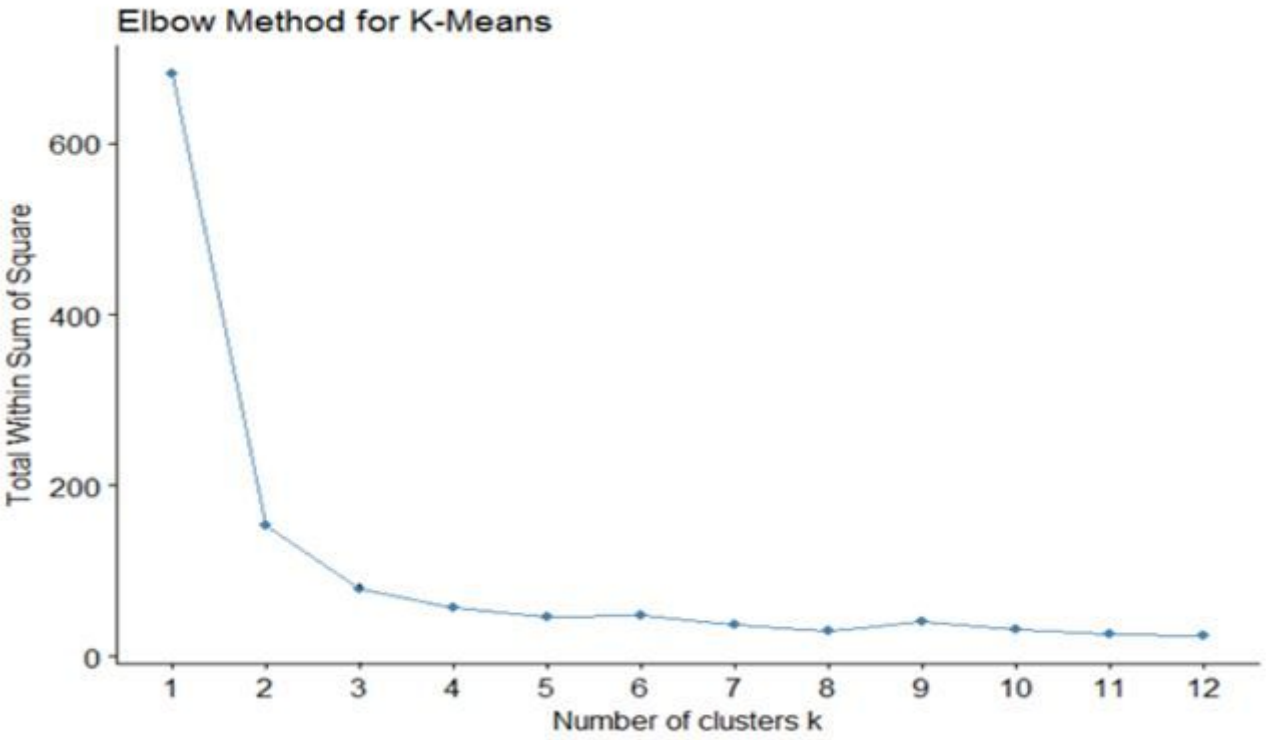

Figure 2

Elbow method showing the best k (liver-disorders dataset).

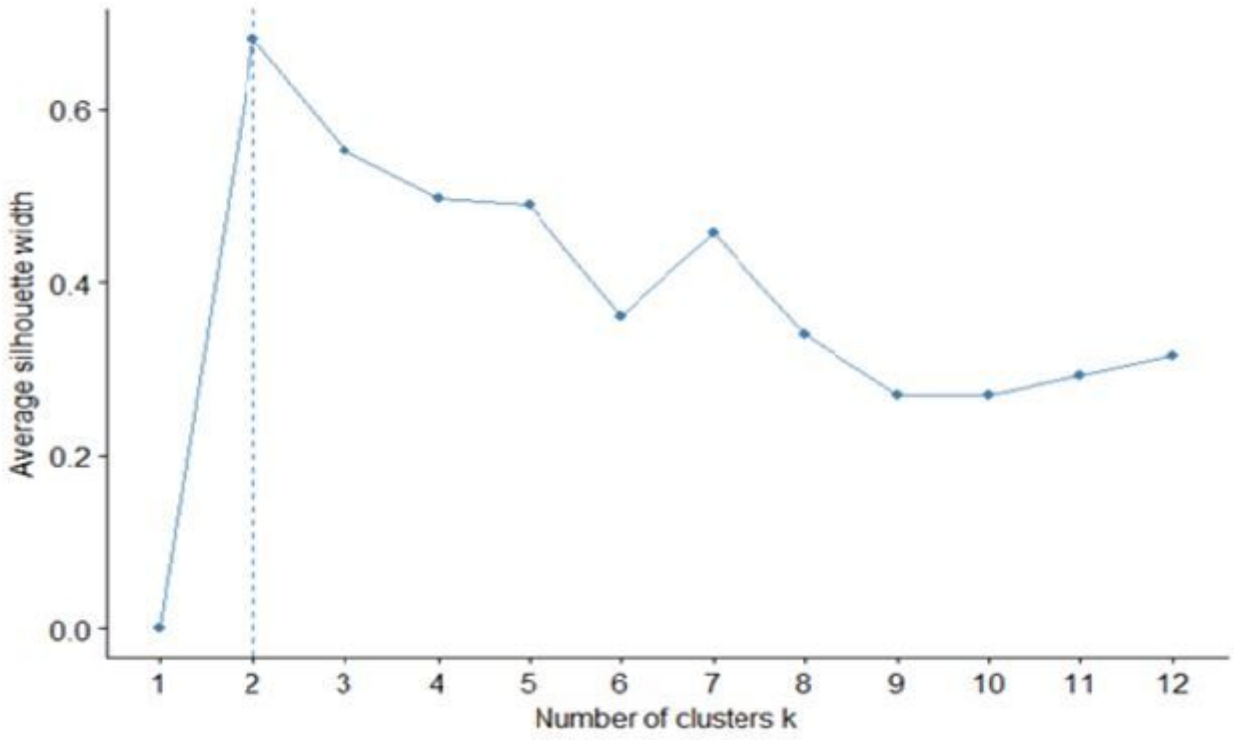

Figure 3

s(i) of different k for the average silhouette method (liver-disorders dataset).

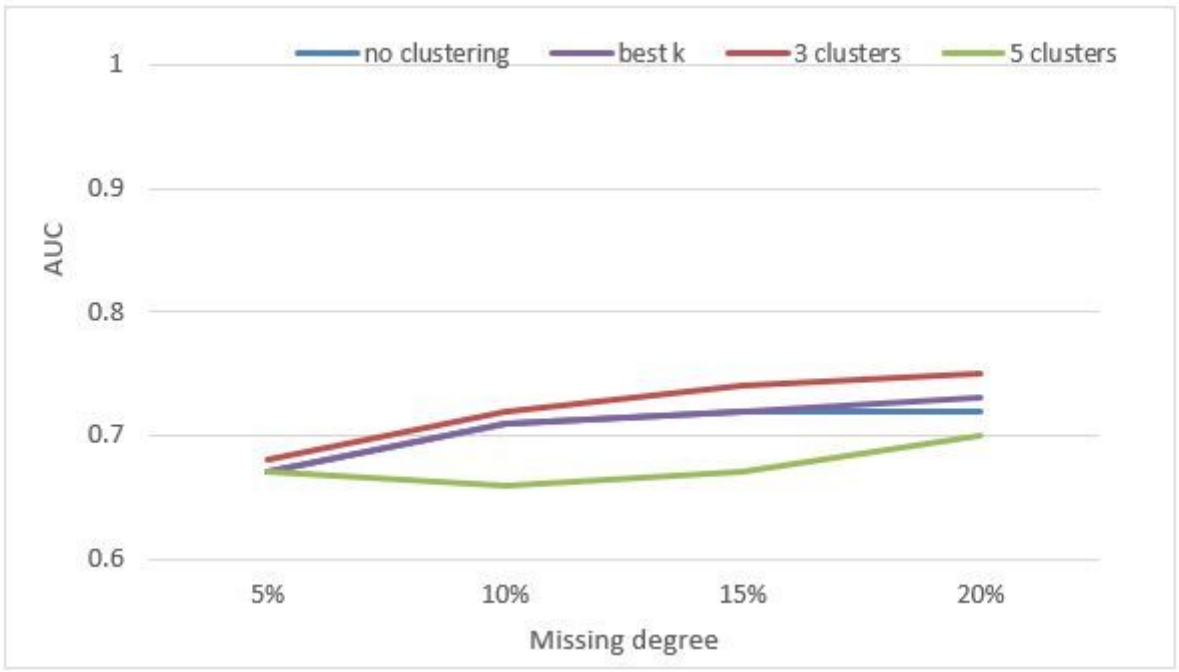

Figure 4

Average AUC of proposed method for different MDs and cluster number in MAR (liver disorders).

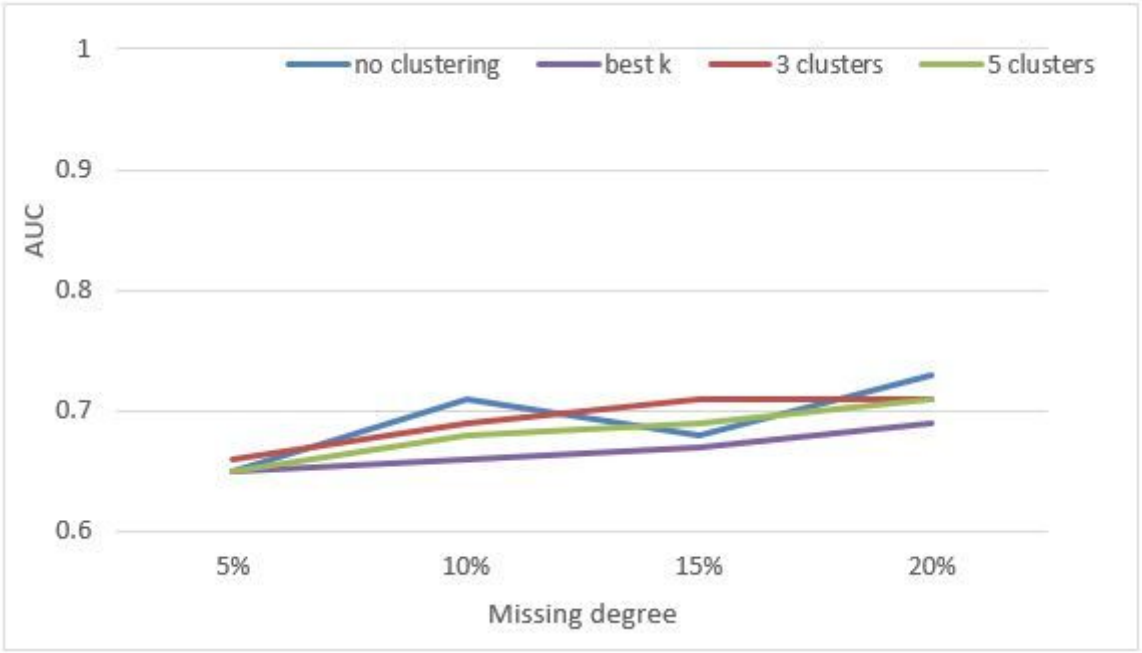

Figure 5

Average AUC of proposed method for different MDs and cluster numbers in MCAR (liver-disorders data)

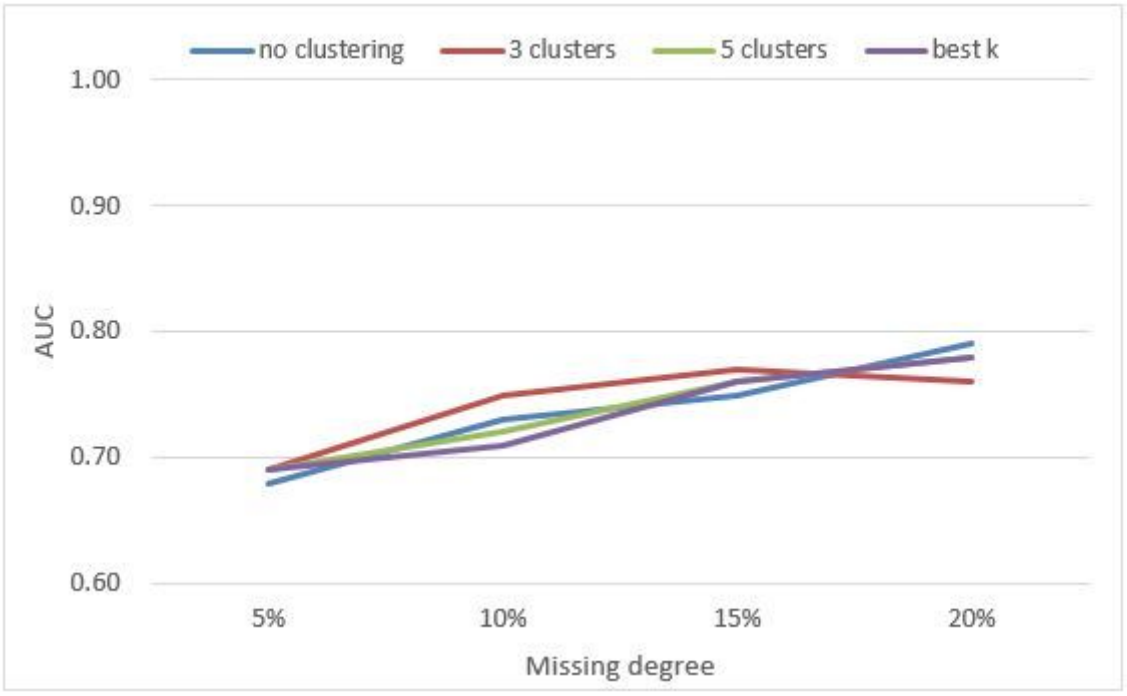

Figure 6

Average AUC of proposed method for different MDs and cluster numbers in MAR (ILPD data).

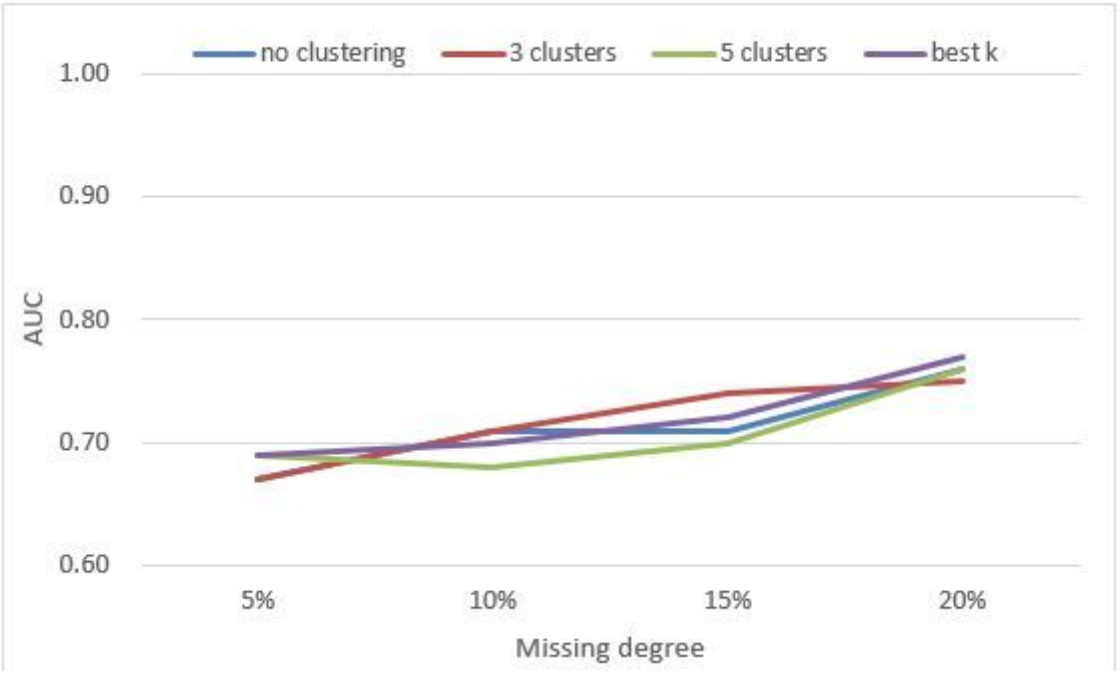

Figure 7

Average AUC of proposed method for different MDs and cluster numbers in MCAR (ILPD data).

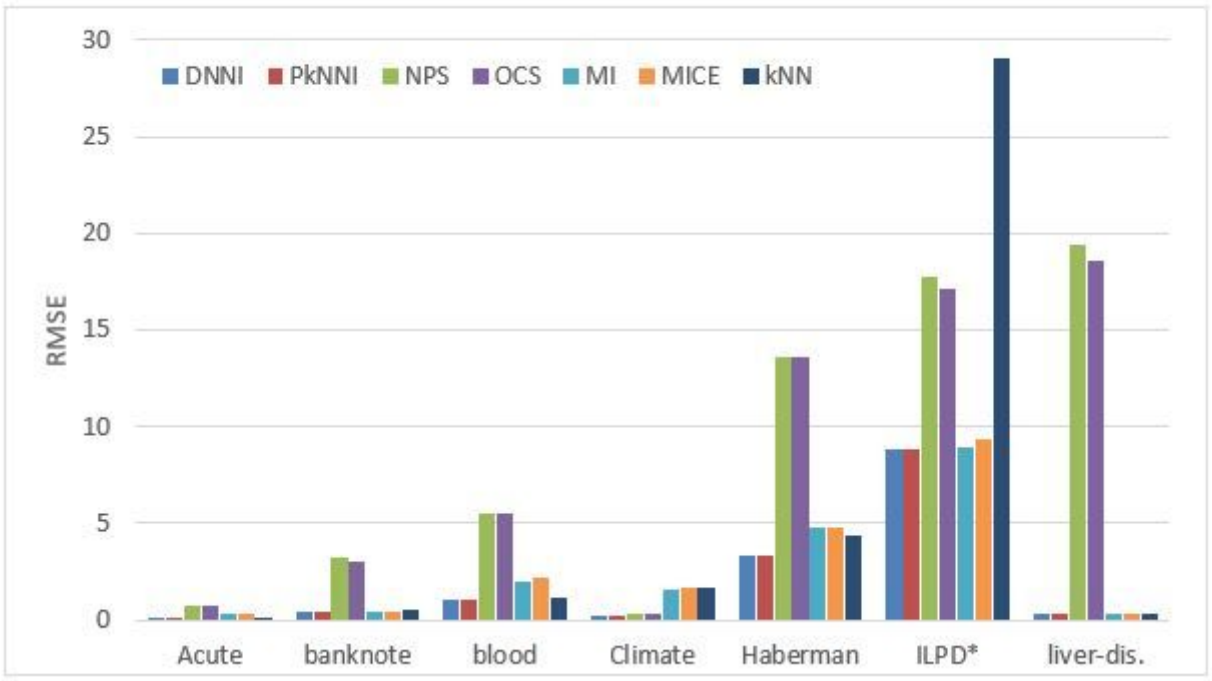

Figure 8

Minimal RMSE of seven imputation methods in MAR.

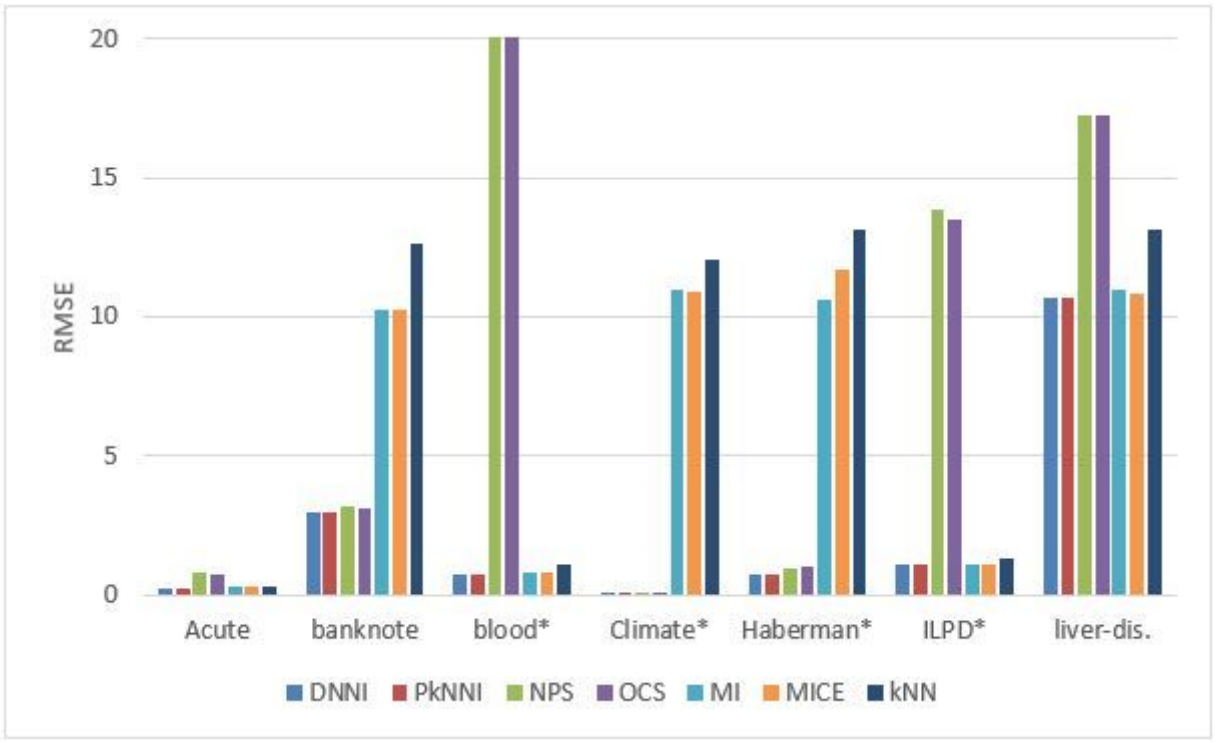

Figure 9

Minimal RMSE of seven imputation methods in MCAR.

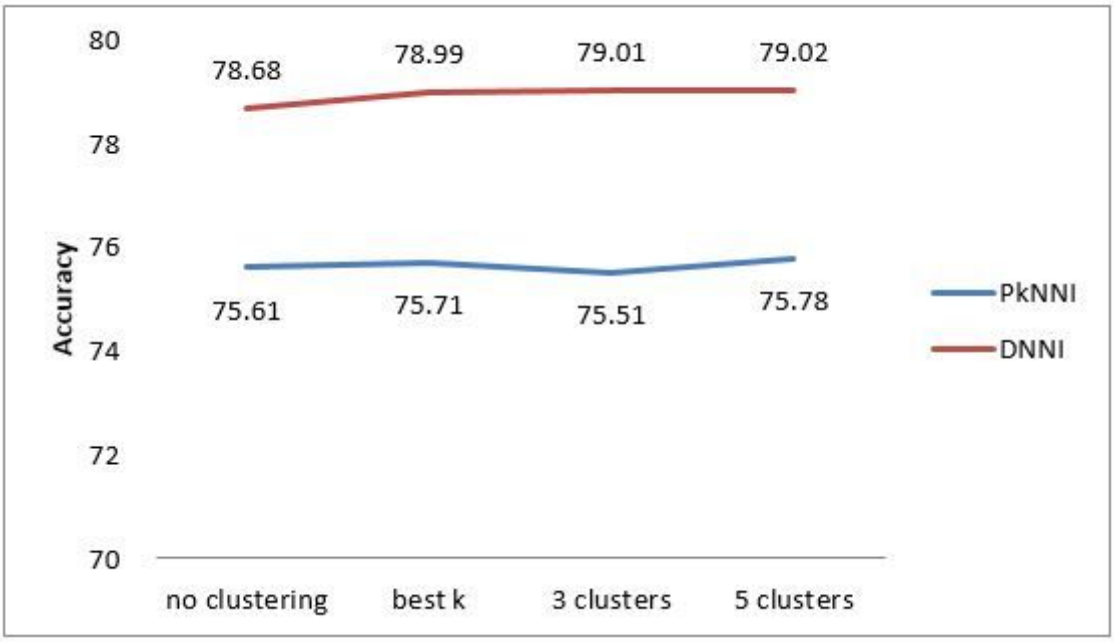

Figure 10

Average of the accuracy of the proposed method for different number of clusters in MNAR (stroke dataset)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Appendix.docx