ComTech: Computer, Mathematics and Engineering Applications, 15(1), June 2024, 29−40 P-ISSN: 2087-1244
DOI: 10.21512/comtech.v15i1.11002 E-ISSN: 2476-907X

Performance of Fuzzy C-Means (FCM) and Fuzzy Subtractive Clustering (FSC) on Medical Data Imputation

Sri Kusumadewi[1]*, Linda Rosita[2], and Elyza Gustri Wahyuni[3]

[1,3]Department of Informatics, Faculty of Industrial Technology, Universitas Islam Indonesia
Yogyakarta, Indonesia 55584
[2]Departement of Clinical Pathology, Faculty of Medicine, Universitas Islam Indonesia
Yogyakarta, Indonesia 55584
[1]sri.kusumadewi@uii.ac.id; [2]linda.rosita@uii.ac.id; [3]elyza@uii.ac.id

Received: 12th December 2023/ **Revised:** 19th April 2024/ **Accepted:** 22nd April 2024

How to Cite: Kusumadewi, S., Rosita, L., & Wahyuni, E. G. (2024). Performance of Fuzzy C-Means (FCM) and Fuzzy Subtractive Clustering (FSC) on Medical Data Imputation. *ComTech: Computer, Mathematics and Engineering Applications, 15*(1), 29−40. https://doi.org/10.21512/comtech.v15i1.11002

Abstract **-** Missing values or incomplete data are frequently encountered in medical records. These issues will be a serious problem if the data must be provided completely for analysis. The research aimed to prove the performance of the Fuzzy Subtractive Clustering (FSC) and Fuzzy C-Means (FCM) methods for solving imputation problems. Both methods were implemented using medical data. It had been conducted using K-Means as a crisp clustering approach for imputation. In the research, fuzzy clustering—a distinct methodology—was applied. The primary research contribution was the suggested fuzzy logic imputation method, which took uncertainty under consideration. The data sample consisted of patients who were at least 40 years old and had a history of hypertension, diabetes, heart disease, stroke, or chronic kidney disease. The test was carried out by taking random portions of data from the entire medical record. The randomization technique used a probability of 10%–50%. The results of the ANOVA test show that the p-value is greater than $\propto$(=0.05). It means that the imputed value does not differ from the original value, whether implemented in the FSC or FCM method. The algorithm's performance is evaluated using the Pearson correlation coefficient. According to the t-test results, the FCM method has a higher correlation coefficient than the FSC method. It implies that FCM is superior to FSC.

Keywords**:** Fuzzy C-Means (FCM), Fuzzy Subtractive Clustering (FSC), medical data imputation

I. INTRODUCTION

Missing data in medical records are very probable because there are occasions when patients' observations are not performed routinely or updated accurately (Nancy et al., 2017). When the value of a desired variable is not measured or recorded for all subjects in the sample, it is referred to as missing data (Austin et al., 2020). Clinical research is fraught with missing data. Suppose people want to know someone's hemoglobin, triglyceride, High Density Lipoprotein (HDL), Low Density Lipoprotein (LDL), total cholesterol, blood sugar, creatinine, and other clinical data. In medical data, some attribute values are not always accessible. The root cause of the issue often lies in the existence of some missing medical data. Ignoring missing data has the potential to produce bias in parameter estimations when full data analysis is performed on all data attributes, such as in the case of predicting metabolic syndrome (Kusumadewi et al., 2020, 2022a, 2022b). The small number of samples also exacerbates this issue (Blazek et al., 2021; Nishanth & Ravi, 2016).

Missing data can be classified into three categories: 1) missing values do not have a relationship or are independent of other data sets; 2) the missing value depends on another variable, but the value can be obtained by estimating the other complete variable; and 3) missing values depend on other missing values. Therefore, missing values cannot be estimated from existing data (Kumaran et al., 2019). The traditional approaches for missing data are deleting data points

*Corresponding Author

with missing values during the pre-processing phase or estimating the missing values using the mean or zero values (Afghari et al., 2019). However, the first approach has consequences for data loss. It can be problematic if there is a significant amount of missing data, and it is extremely difficult to complete the data because the data cannot be retrieved simultaneously.

Solving the imputation problem in the traditional method has been studied using the mean method (Ferrer et al., 2021; Luo & Paal, 2021; Pandey et al., 2021), linear interpolation (Nobach, 2019; Yang & Hu, 2018), and regression (Crambes & Henchiri, 2019; Roeling & Nicholls, 2020). The disadvantage of traditional methods is that they are deterministic, so they are not suitable for solving problems that contain uncertainty. Currently, there are several algorithms that are able to handle the problem of uncertainty, imprecision, and partial truth, known as soft computing. Fuzzy systems, neural networks, evolutionary computing, probabilistic reasoning, and swarm intelligence are part of soft computing.

Research on imputation using soft computing has been carried out by several researchers. The results of all the studies indicate that the soft computing method is capable of imputing well. Fuzzy logic is used for imputation by Khan et al. (2021) and Sefidian and Daneshpour (2019). However, the number of clusters cannot be obtained optimally. Then, neural network imputation is used by Choudhury and Pal (2019), Gautam and Ravi (2015), and Verpoort et al. (2018). Meanwhile, the example of swarm intelligence for imputation is Nekouie and Moattar (2019). Evolutionary computation is used for imputation in Gautam and Ravi (2015).

Imputation using the nearest neighbor idea has been shown to be particularly effective in dealing with missing data. When handling missing values, some imputation methods are applicable (Audigier et al., 2018). Multiple imputation has been proven to be an effective method for dealing with missing data and imputation ambiguity. Compared to existing methods, sequential imputation using weighted nearest neighbors may be successfully applied to various data circumstances and is close to the best (Faisal & Tutz, 2021). The imputation process is also implemented on medical data (stroke dataset). The previous researchers use k-Nearest Neighbors (k-NN) and compare it with J48, Multilayer Perceptron (MLP), Random Forest (RF), and Support Vector Machine (SVM) methods. The results show that k-NN has the best accuracy (Cheng et al., 2020).

The research aims to implement Fuzzy clustering as a method for imputation. The implementation of clustering is intended to group data points with similar characteristics, so imputation is expected to focus more on a similar group. The research presents three primary contributions: 1) provide an innovative approach to imputation challenges, particularly with regard to medical data; 2) use Fuzzy logic in the proposed framework to adjust for uncertainty; and 3) apply the cluster approach to imputation to preserve the characteristics of the data properly. Two methods are used: Fuzzy Subtractive Clustering (FSC) and Fuzzy C-Means (FCM). According to Khan et al. (2021) and Sefidian and Daneshpour (2019), FCM is a dependable method for imputation. On the other hand, FSC is utilized to determine the ideal number of clusters. Imputation is performed on medical record data for patients with a history of hypertension, diabetes mellitus, cardiovascular disease, stroke, or chronic kidney disease. These two methods will be compared to choose the best method to determine which is the best method for dealing with missing data in medical records.

II. METHODS

The research is carried out in several stages, as shown in Figure 1. Research begins with data collection. A total of 104 medical records are used in the research. Data are gathered from a number of hospitals in the Daerah Istimewa Yogyakarta (DIY) Province. The studied patients are at least 40 years old. The age requirements are chosen according to the beginning stages of metabolic syndrome risk. There are 61 complete data points in the 104 medical record data points used as reference data (X). The remaining 43 data points with missing values are used as evaluation data (Y'). Age, HDL, LDL, triglycerides, total cholesterol, fasting blood sugar, systolic blood pressure, and diastolic blood pressure are the variables/dimensions/attributes used. The data types for the eight variables are numeric. A history of hypertension, diabetes, stroke, cardiovascular disease, and chronic kidney disease is also regarded. The five variables have a Boolean data type, which means they are true (1) if the patient has a related history and false (0) if they do not.

However, not all of the data from the samples will be clustered. The clustering data is chosen from a sample of data with complete values and no missing values. This particular data collection is also referred to as the reference data (X). If n^* reference data are obtained, $n–n^*$ data must be calculated to fill in the missing value. The term "evaluation data" refers to this incomplete dataset.

To obtain cluster center, both FSC and FCM are implemented on X. Let X be an ($n \times m$) matrix, where n is the number of data points and m is the number of data dimensions. Data must be normalized first before the clustering process begins. For each dimension, the normalization method requires a lower bound (x_{min}) and an upper bound (x_{max}). Equation (1) can be used to calculate normalization for data points on the j-th dimension (x_j).

$$x_j = \frac{x_j - x_{min}}{x_{max} - x_{min}}. \quad (1)$$

Algorithm 1 shows the clustering procedure using the FSC technique. The clustering process starts by calculating the density value of each data point. The

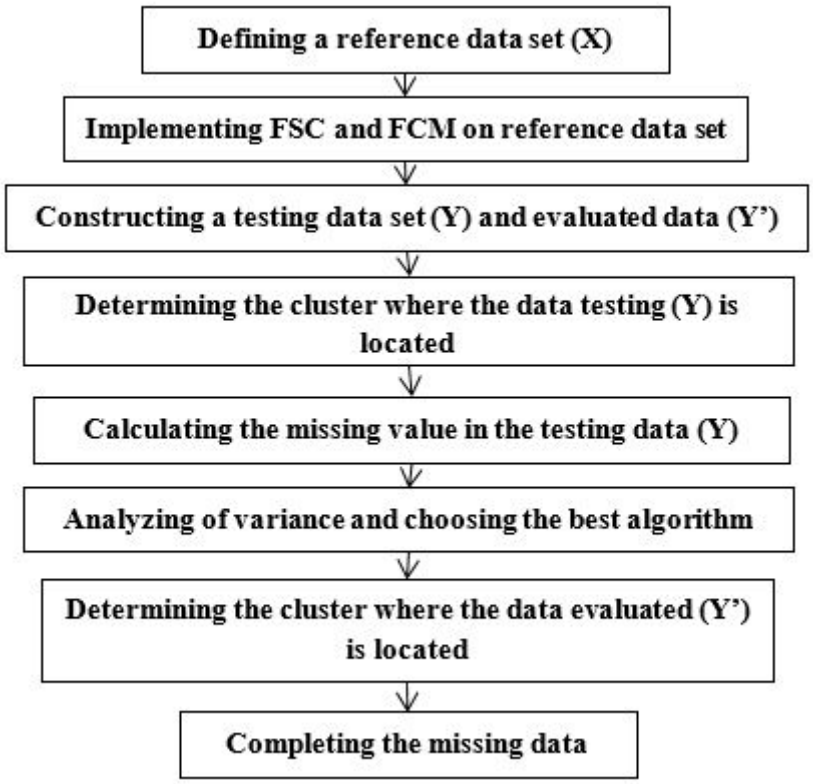

Figure 1 Research Stages

density of x_i (D_i) can be calculated using equitation by Yang et al. (2021). Equation (2) has $||x_t–x_i||$ as the distance between x_i and r (neighborhood radius) as a positive constant that shows how much the influence of the cluster center is on each variable. Then, according to Nancy et al. (2017), r can be calculated using Equation (3).

$$D_i = \sum_{t=1}^{n} exp\left(-\frac{4\|x_t - x_i\|^2}{r^2}\right), \quad (2)$$

$$r = \frac{1}{2}\min\{max\{\|x_t - x_i\|\}\}. \quad (3)$$

If a data point has a lot of close neighbors, it will have a lot of density. The data with the greatest density will be chosen as the center of the cluster. The input data with the biggest density value is denoted as x_{ti}, and its density value is denoted as D_{ti}. Next, x_{ti} is the first cluster (C_1), and the density of the surrounding data will be reduced. In Equation (4), r_b is a positive value that has the effect of reducing the data density in a cluster. Generally, r_b is greater than r. It is calculated as $(q)(r)$, where q is the squash factor in the range 1.2–1.5. It means that the close data to the cluster center will decrease in density very much. As a result, it will be impossible for the data to become the center of the next cluster.

$$D_i = D_i - D_{ti}\left(exp\left(-\frac{4\|x_i - x_{ti}\|^2}{r_b^2}\right)\right). \quad (4)$$

After the data density has been revised, the second cluster center (C_2) will be searched. After C_2 is obtained, the density of each data will be revised again, and so on. The researchers can use two fractions as a comparison factor: accept and reject ratios. Both accept and reject ratios are fractional numbers with values from 0 to 1. The accept ratio is the lower limit of a candidate cluster center being allowed to become a cluster center. Meanwhile, the reject ratio is the top limit for a prospective cluster center that cannot become a cluster center. If a data point with the highest potential (e.g., x_k with potential or density D_k) has been identified in an iteration, the iteration will be continued by finding the ratio of the potential data point with the highest potential of a data point at the start of the iteration (e.g., x_h $\frac{D_k}{D_h}$th potential D_h). The proportion is computed using $\frac{D_k}{D_h}$.

Three conditions can occur in an iteration. First, if the ratio is > accept ratio, x_k will be accepted as the center of the new cluster. Second, if the reject ratio is < ratio ≤ accept ratio, x_k will be accepted as the center of the new cluster only if the data has a long distance enough from the other cluster centers. Otherwise, the data point will not be accepted as the center of the cluster and will no longer be considered the center of the new cluster (its potential is set to zero). Third, if the ratio is ≤ reject ratio, there are no more data points to be considered as candidates for the cluster center, and the iteration is stopped.

After the clustering process is complete,

the cluster center will be obtained in a normalized condition. However, the cluster center needs to be denormalized first before proceeding to the following process. The denormalization for the *k*-th cluster center on the *j*-th variable is calculated by Equation (5). Next, the range of influence of cluster centers for each data dimension or variable (σ) can be calculated using Equation (6). It shows σ_j as range of influence of cluster centers for *j*-th dimension.

$$c_{kj} = c_{kj}(x_{max} - x_{min}) + x_{min}, \quad (5)$$

$$\sigma = \frac{r(x_{max} - x_{min})}{\sqrt{8}}. \quad (6)$$

Algorithm 1: FSC Algorithm

Input: data set (X), neighborhood radius (r), squash factor (q),

Output: cluster center (C), range of influence cluster centers for each data dimension ().

1: Normalize X so the data points are between (0 and 1) or (-1 to 1).

2: Calculate the density of each data point based on Equation (2).

3: Choose a data point with the most potential to be the first cluster center (x_{c1}).

4: Improve the density of remaining data points around cluster center (x_{c1}) based on Equation (4).

5: Choose the remaining data point with the highest potential as the next cluster center (x_{c2}). To decide whether data are accepted as a new cluster center, use the ratio, accept ratio, and reject ratio.

6: Repeat steps 4 and 5 until all data are within the cluster center influence range.

7: Denormalize the cluster center and calculate based on Equations (5) and (6).

In the FCM data clustering method, the membership value indicates whether a particular point of data exists in a cluster. FCM is a semi-supervised clustering algorithm. It is necessary to first determine the number of clusters to determine the optimal number of clusters. The Average Silhouette Width (ASW) is a popular cluster validation index used for estimating the number of clusters (Batool & Hennig, 2021), as well as Condensed Silhouette (Naghizadeh & Metaxas, 2020). It is not examined specifically in the research how many clusters there are. The primary idea behind FCM is to find the cluster center, which will be used to compute the average location of each cluster. The cluster center is still not precise in the initial situation. A membership value for each cluster is assigned to each data point. It will be seen that the cluster center will shift to the correct location if the cluster center and the membership value of each data point are improved repeatedly. This iteration is based on minimizing the objective function that defines the distance between a particular data point and the cluster center, weighted by the data point's membership value.

Algorithm 2 shows the clustering procedure using the FCM technique. Suppose *X* is an ($n \times m$) matrix, with *n* representing the number of data points to cluster and *m* representing the number of variables. The data will be divided into *N* groups. The FCM technique starts by creating an ($n \times N$) matrix (U) containing random numbers. The membership value of a data point in a cluster is stored in the partition matrix (U). The elements of each matrix U (μ_{ik}) are calculated using Equation (7). Next, the researchers repair the cluster center (C) with Equation (8). Next, the objective value is calculated in the first iteration (t=1) with Equation (9). Last, the U matrix is repaired based on the new cluster center with Equation (10). In general, *w* is any positive number. In this case, the researchers set w = 2.

$$\mu_{ik} = \frac{u_{ik}}{\sum_{k=1}^{N} u_{ik}}, \quad (7)$$

$$c_{kj} = \frac{\sum_{i=1}^{n}\left((\mu_{ik})(x_{ij})\right)}{\sum_{i=1}^{n}(\mu_{ik})}, \quad (8)$$

$$P_t = \sum_{i=1}^{n}\sum_{k=1}^{N}\left(\left(\sum_{j=1}^{m}(x_{ij} - c_{kj})^2\right)(\mu_{ik})\right) \quad (9)$$

$$\mu_{ik} = \frac{\left(\sum_{j=1}^{m}\left(x_{ij} - c_{kj}\right)^2\right)^{\frac{-1}{w-1}}}{\sum_{k=1}^{N}\left(\sum_{j=1}^{m}\left(x_{ij} - c_{kj}\right)^2\right)^{\frac{-1}{w-1}}} \quad (10)$$

Algorithm 2: FCM Algorithm

Input: data set (X), number of clusters (N), threshold (ξ), maximum iteration (MaxIter),

Output: cluster center (x_c), partition matrix (U), objective value (P).

1: Create an n x N partition matrix (U) based on random numbers.

2: Calculate the membership value for each data point in each cluster with Equation (7).

3: Calculate the cluster center with Equation (8).

4: Improve partition matrix U with Equation (10).

5: Calculate the objective value (P) with Equation (9).

6: Check if (P < ξ) or (iteration > MaxIter). If that is true, stop iterating. Otherwise, go to step 3.

After the clustering process, the cluster center generated by FSC is named C1, and the cluster center generated by FCM is named C2. The software Matlab carries out the clustering process. Next, the researchers

build a set of test data (Y) and evaluation data (Y'). The testing dataset consists of a reference dataset from which some data have been randomly removed. The reference data, which contain some missing data, will be used to assess the model's performance. The following percentages of the data are chosen at random (p): 10%, 20%, 30%, 40%, and 50%. The process begins with creating a set of (*n x m*) random numbers, where *m* denotes the number of variables and *n* is the number of reference data. A random number denotes the possibility that a reference data point has been left out. The random number less than *p* is the value that needs to be eliminated.

Algorithm 3 illustrates this imputation procedure. The imputation process begins by determining a reference data set (X). X comes from a sample data set (P) that has complete values for each variable. Meanwhile, the researchers use incomplete samples, also known as missing data, as test data (Y'). Next, the researchers apply the FSC algorithm to X to determine the cluster center (C1). Similarly, the researchers apply the FCM algorithm to X to obtain the cluster center (C2). Next, the researchers find the cluster where the testing data is located (Y). Calculating the distance between each data point and the cluster center is to find a suitable cluster. The data point is located in the cluster with the shortest cluster center distance from the data point. The distance is calculated using the Euclidean distance formula in Equation (11). It shows d_k as distance of Y to k-th cluster center (C_k), C_{kj} as value of k-th cluster center on j-th variable, y_j as value of data evaluated on j-th variable, and $m*$ as number of variables or dimensions of cluster center minus 1.

$$d_k = \sqrt{\sum_{j=1}^{m*} (c_{kj} - y_j)^2} \quad (11)$$

The cluster center is used to complete the missing data in the data set after obtaining the relevant cluster. Thus, the data set's missing values can be filled in. The researchers label Y for this data set.

The analysis of variance is used to see if the means of two or more groups differ significantly. An ANOVA test with the same subject is used to see if there is a significant difference between the mean of the original data and the estimation findings using FSC and FCM. It selects the most appropriate algorithm. The Pearson correlation coefficient is calculated using the t-test. The optimal algorithm between FSC and FCM is chosen using this Pearson correlation coefficient. This correlation coefficient illustrates the statistical relationship that exists between the original and estimated values. A better algorithm is one with a higher Pearson correlation coefficient.

Next, the researchers can find the cluster where the data evaluated (Y') are located. Finding a suitable cluster by calculating the distance of the data point to each cluster center is obtained from the selected algorithm using Equation (11). Finally, the missing data can be filled in. The closest distance from each cluster center can be used to compute the imputation on the evaluation data. The value for missing data will be the cluster center value.

Algorithm 3: Imputation Algorithm

Input: sample data set (P), data evaluated (Y')

Output:

1: Define a reference data set (X), i.e., a sample dataset (P) that consists of complete values for each variable. Meanwhile, incomplete samples (missing data) will serve as testing data (Y').

2: Implement the FSC algorithm on the reference data set (X) to obtain cluster center (C1) and the FCM algorithm on the reference data set (X) to obtain cluster center (C2).

3: Create a testing data set (Y). This is a reference data set from which some data have been randomly removed.

4: Based on Equation (11), determine the cluster of each data point in Y. Implementation of the FSC and FCM algorithms will lead to each data point becoming a member of a cluster.

5: Use the ANOVA test with the same subject to determine whether there is a significant difference between the mean of the original data and the imputed results based on FSC and FCM.

6: Implement a t-test to obtain the Pearson correlation coefficient to select the best algorithm between FSC and FCM.

7: Determine the clusters of the individual data points in the set Y' using Equation (11) and the results from the selected algorithm. The cluster center will be the value for missing data.

III. RESULTS AND DISCUSSIONS

The researchers use 61 reference data consisting of 30 male data and 31 female data. Table 1 (see Appendices) shows the profile of the data. In the FSC method, the researchers use a squash factor = 2.0, an accept ratio = 0.9, and a rejection ratio = 0.025. The smaller the acceptance ratio is, the more clusters are generated. Similarly, the smaller the reject ratio is, the more clusters are generated. Equation (3) shows the same range of influence for all variables. The researchers need a different influence range for each variable. Therefore, by eliminating the maximum function, the researchers slightly modify Equation (3) so that a unique influence range can be determined for each variable.

Table 1 (see Appendices) shows the influence range of each variable. The smaller the influence range is, the more clusters are generated. Table 1 also shows that the influence range of the related variables is almost the same, between 0.4 and 0.5. According to Table 1, the influence range of total cholesterol (0.4) has the smallest value. It means that, in the total cholesterol variable, at least one data point has the closest distance to another data point. If no modifications are made to

Equation (3), the influence range of all variables is 0.5.

The FSC method generates 23 clusters. Table 2 (see Appendices) shows the cluster centers for each cluster. When using the FSC method, the cluster center is a part of the dataset rather than data obtained from arithmetic operations accomplished during the clustering process. According to the clustering results, there are 9 data points in Cluster 3, 8 in Cluster 2, 7 in Clusters 1 and 4, 6 in Cluster 5, 2 in Cluster 6, 6 in Cluster 7, and 1 point in the last cluster.

The clusters formed as a result of FSC show the optimum number of clusters. The researchers will base FCM on this number of clusters. Therefore, in FCM, the researchers start clustering with 23 clusters based on the number of clusters generated by FSC. Table 3 (see Appendices) provides the cluster center for each cluster. According to the FCM method, the cluster center is generated from the average of the cluster members, so there must be some adjustments to the data format. For the age variable, the researchers round the value so that it is an integer. If a decimal value is greater than or equal to 0.5, it is rounded up. Meanwhile, if it is less than 0.5, it is rounded down. Similarly, for chronic kidney disease, diabetes mellitus, hypertension, cardiovascular disease, and stroke, a score of 1 is assigned if the decimal value is more than or equal to 0.5. Then, a score of 0 is assigned if it is less than 0.5. According to the clustering results, Cluster 1 has the greatest number of members—8 data points—followed by Cluster 8 with 6 data points, Cluster 3 with 5 data points, Cluster 5, Cluster 10, Cluster 12, Cluster 13, Cluster 14 with 4 data points, Cluster 9 with 3 data points, Cluster 4, Cluster 7, Cluster 20, Cluster 21 with 2 data points, and the other Cluster with 1 data point.

Figure 2 shows the cluster visualization obtained from FCM. The dataset has 13 dimensions, so it is impossible to describe them simultaneously. Thus, the researchers describe only three visualizations. Figure 2a depicts the FCM results on total cholesterol and fasting blood sugar. Then, Figure 2b shows a visual representation of the FCM results on total cholesterol and systolic blood pressure. Figure 2c visualizes the FCM results on fasting blood sugar and systolic blood pressure. Different colors are used to represent the 23 clusters. For example the top legend (red) indicates that the data points in red are members of Cluster 1.

The test data is obtained using a random process of removing certain data points. Randomization begins by determining the probability of missing values (p), which are 10%, 20%, 30%, 40%, and 50%. Only seven variables have missing data: HDL, LDL, triglycerides, total cholesterol, fasting blood sugar, systolic blood pressure, and diastolic blood pressure. Then, the researchers generate a matrix (R) containing random numbers (between 0 and 1) with a size of (61 × 7). If $r_{ij} < p$, the researchers set y_{ij} as a missing value.

In the Y matrix, the missing values will be imputed by finding the closest distance from each data point to the cluster center. ANOVA analysis is performed with the same subject on the original data group: the set of imputed results with FSC and the set of imputed results with FCM. The researchers test whether there is a significant difference between the mean of the original data and the imputation results using FSC and FCM. Table 4 (see Appendices) shows the results of the ANOVA test.

According to Table 4 (see Appendices), all p-values for the ANOVA are greater than $\propto$ (=0.05). It means that the imputed value does not differ from the original value. Thus, the FSM or FCM can be determined to be the best choice. The Pearson correlation coefficient of the paired sample t-test can be used to determine the best method (Batool & Hennig, 2021). Table 4 shows the results of the t-test. Table 4 shows that for all conditions, the Pearson correlation coefficient obtained using the FCM method is greater than the Pearson correlation coefficient obtained using

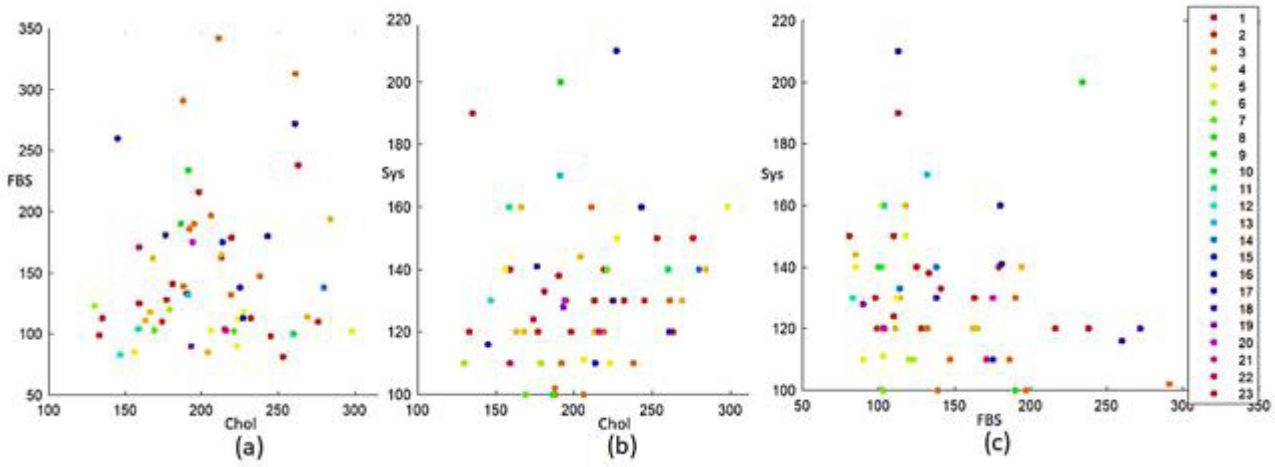

Figure 2 Cluster Visualization Using FCM: (a) Total Cholesterol and Fasting Blood Sugar; (b) Total Cholesterol and Systolic Blood Pressure; and (c) Fasting Blood Sugar and Systolic Blood Pressure

the FSC method. The FCM method shows a higher correlation coefficient than the FSC method. Mean Square Error (MSE) is another method of measuring the performance of FSC and FCM methods. For all conditions, the MSE for the FCM method is lower than that of the FSC method.

Incomplete medical record data is used as evaluated data (Y'). In this evaluation process, missing data is estimated using the FCM method. Table 5 (see Appendices) shows the evaluated data (Y'). The "-" symbol indicates that data are not available (missing values).

Then, Equation (11) uses the Euclidean formula, which takes the shortest distance, to get the correct cluster position. Table 6 (see Appendices) shows the evaluation results of the incomplete data set in Table 5 (see Appendices) after the imputation process. It also provides the resulting cluster number in the first column. For example, there is no FBS data in the first column. In this case, the closest cluster is Cluster 12. In the first row, FBS is estimated to have a missing value of 98 because the center of Cluster 12 for the FBS variable is 98.

Cluster centers in FCM are calculated by averaging data points that constitute the cluster. It means that the cluster center represents the new data representing the cluster's members. It promotes FCM as a better method of imputation. Cluster centers in FSC are derived from data points with the highest potential, so cluster centers do not represent new data.

There are two implications for employing fuzzy clustering as an imputation method. First, clustering parameters like accept and reject ratios need to be set, just like in other soft computing techniques. Second, Fuzzy clustering still needs to be improved to handle outliers. By hybridizing other soft computing components, such as Adaptive Neuro Fuzzy Inference System (ANFIS) or Fuzzy genetic algorithms, both of these issues can be resolved.

IV. CONCLUSIONS

FSC and FCM are effective imputation methods. According to the ANOVA test, both methods are capable of handling missing data well. A paired sample t-test shows that the FCM method is a better imputation method than the FSC. This is supported by the FCM's MSE value, which is better than FCS. The nearest neighbor concept can be implemented well to perform the imputation of the evaluated data. Clusters with close neighbors have the shortest distance between their centers and the data evaluated.

The research is limited by the haves to justify multiple clustering parameters, including the accept/reject ratios, just like with other soft computing techniques. Therefore, to determine the ideal parameters, hybridization with alternative techniques must be tested. Fuzzy genetic algorithms or the ANFIS can be applied in this way.

In the future, the research will be focused on predicting complications in patients by analyzing HDL, LDL, triglycerides, total cholesterol, fasting blood sugar, systolic blood pressure, and systolic blood pressure measurements. Hypertension, diabetes mellitus, cardiovascular disease, stroke, and chronic kidney disease are some of the possible complications. Medical records that have been imputation-equipped will be used as supporting data for predicting complications.

REFERENCES

Afghari, A. P., Washington, S., Prato, C., & Haque, M. M. (2019). Contrasting case-wise deletion with multiple imputation and latent variable approaches to dealing with missing observations in count regression models. *Analytic Methods in Accident Research*, *24*. https://doi.org/10.1016/J.AMAR.2019.100104

Audigier, V., White, I. R., Jolani, S., Debray, T. P. A., Quartagno, M., Carpenter, J., Van Buuren, S., & Resche-Rigon, M. (2018). Multiple imputation for multilevel data with continuous and binary variables. *Statistical Science*, *33*(2), 160–183. https://doi.org/10.1214/18-STS646

Austin, P. C., White, I. R., Lee, D. S., & Van Buuren, S. (2020). Missing data in clinical research: A tutorial on multiple imputation. *Canadian Journal of Cardiology, 37*(9), 1322–1331. https://doi.org/10.1016/J.CJCA.2020.11.010

Batool, F., & Hennig, C. (2021). Clustering with the average silhouette width. *Computational Statistics & Data Analysis*, *158*. https://doi.org/10.1016/J.CSDA.2021.107190

Blazek, K., Van Zwieten, A., Saglimbene, V., & Teixeira-Pinto, A. (2021). A practical guide to multiple imputation of missing data in nephrology. *Kidney International*, *99*(1), 68–74. https://doi.org/10.1016/J.KINT.2020.07.035

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*, *122*. https://doi.org/10.1016/J.COMPBIOMED.2020.103824

Choudhury, S. J., & Pal, N. R. (2019). Imputation of missing data with neural networks for classification. *Knowledge-Based Systems*, *182*. https://doi.org/10.1016/J.KNOSYS.2019.07.009

Crambes, C., & Henchiri, Y. (2019). Regression imputation in the functional linear model with missing values in the response. *Journal of Statistical Planning and Inference*, *201*, 103–119. https://doi.org/10.1016/J.JSPI.2018.12.004

Faisal, S., & Tutz, G. (2021). Multiple imputation using nearest neighbor methods. *Information Sciences*, *570*, 500–516. https://doi.org/10.1016/J.INS.2021.04.009

Ferrer, A. H., El Korso, M. N., Breloy, A., & Ginolhac,

G. (2021). Robust mean and covariance matrix estimation under heterogeneous mixed-effects model with missing values. *Signal Processing*, *188*. https://doi.org/10.1016/J.SIGPRO.2021.108195

Gautam, C., & Ravi, V. (2015). Data imputation via evolutionary computation, clustering and a neural network. *Neurocomputing*, *156*, 134–142. https://doi.org/10.1016/J.NEUCOM.2014.12.073

Khan, H., Wang, X., & Liu, H. (2021). Missing value imputation through shorter interval selection driven by Fuzzy C-Means clustering. *Computers & Electrical Engineering*, *93*. https://doi.org/10.1016/J.COMPELECENG.2021.107230

Kumaran, S. R., Othman, M. S., Yusuf, L. M., & Yunianta, A. (2019). Estimation of missing values using hybrid Fuzzy Clustering Mean and majority vote for microarray data. *Procedia Computer Science*, *163*, 145–153. https://doi.org/10.1016/J.PROCS.2019.12.096

Kusumadewi, S., Rosita, L., & Wahyuni, E. G. (2020). *Model sistem pendukung keputusan klinis untuk sindrom metabolik* (1st ed.). UII Press.

Kusumadewi, S., Rosita, L., & Wahyuni, E. G. (2022a). Development of a modified certainty factor model for prediction of metabolic syndrome. *International Journal of Innovative Computing, Information and Control (IJICIC)*, *18*(5), 1463–1475.

Kusumadewi, S., Rosita, L., & Wahyuni, E. G. (2022b). Selection of aggregation function in Fuzzy inference system for metabolic syndrome. *International Journal on Advanced Science, Engineering and Information Technology*, *12*(5), 2140–2146. https://doi.org/10.18517/IJASEIT.12.5.15552

Luo, H., & Paal, S. G. (2021). Advancing post-earthquake structural evaluations via sequential regression-based predictive mean matching for enhanced forecasting in the context of missing data. *Advanced Engineering Informatics*, *47*. https://doi.org/10.1016/J.AEI.2020.101202

Naghizadeh, A., & Metaxas, D. N. (2020). Condensed silhouette: An optimized filtering process for cluster selection in K-Means. *Procedia Computer Science*, *176*, 205–214. https://doi.org/10.1016/J.PROCS.2020.08.022

Nancy, J. Y., Khanna, N. H., & Arputharaj, K. (2017). Imputing missing values in unevenly spaced clinical time series data to build an effective temporal classification framework. *Computational Statistics & Data Analysis*, *112*, 63–79. https://doi.org/10.1016/J.CSDA.2017.02.012

Nekouie, A., & Moattar, M. H. (2019). Missing value imputation for breast cancer diagnosis data using tensor factorization improved by enhanced reduced adaptive particle swarm optimization. *Journal of King Saud University - Computer and Information Sciences*, *31*(3), 287–294. https://doi.org/10.1016/J.JKSUCI.2018.01.006

Nishanth, K. J., & Ravi, V. (2016). Probabilistic neural network based categorical data imputation. *Neurocomputing*, *218*, 17–25. https://doi.org/10.1016/J.NEUCOM.2016.08.044

Nobach, H. (2019). Note on nonparametric spectral analysis of wideband spectrum with missing data via sample-and-hold interpolation and deconvolution. *Digital Signal Processing*, *87*, 19–20. https://doi.org/10.1016/J.DSP.2019.01.008

Pandey, A. K., Singh, G. N., Sayed-Ahmed, N., & Abu-Zinadah, H. (2021). Improved estimators for mean estimation in presence of missing information. *Alexandria Engineering Journal*, *60*(6), 5977–5990. https://doi.org/10.1016/J.AEJ.2021.04.053

Roeling, M. P., & Nicholls, G. K. (2020). Imputation of attributes in networked data using Bayesian autocorrelation regression models. *Social Networks*, *62*, 24–32. https://doi.org/10.1016/J.SOCNET.2020.02.005

Sefidian, A. M., & Daneshpour, N. (2019). Missing value imputation using a novel grey based Fuzzy C-Means, mutual information based feature selection, and regression model. *Expert Systems with Applications*, *115*, 68–94. https://doi.org/10.1016/J.ESWA.2018.07.057

Verpoort, P. C., MacDonald, P., & Conduit, G. J. (2018). Materials data validation and imputation with an artificial neural network. *Computational Materials Science*, *147*, 176–185. https://doi.org/10.1016/J.COMMATSCI.2018.02.002

Yang, J., & Hu, M. (2018). Filling the missing data gaps of daily MODIS AOD using spatiotemporal interpolation. *Science of the Total Environment*, *633*, 677–683. https://doi.org/10.1016/J.SCITOTENV.2018.03.202

Yang, L. H., Ye, F. F., Liu, J., Wang, Y. M., & Hu, H. (2021). An improved Fuzzy rule-based system using evidential reasoning and subtractive clustering for environmental investment prediction. *Fuzzy Sets and Systems*, *421*, 44–61. https://doi.org/10.1016/J.FSS.2021.02.018

APPENDICES

Table 1 Lower Bound, Upper Bound, and Influence Range for Each Variable

Variable		Lower Bound	Upper Bound	Influence Range
V_1	Age	43	85	0.46
V_2	High-Density Lipoprotein (HDL)	22	80	0.49
V_3	Low-Density Lipoprotein (LDL)	50	276	0.49
V_4	Triglycerides (TG)	35	549	0.41
V_5	Total Cholesterol (Chol)	69	388	0.40
V_6	Fasting Blood Sugar (FBS)	65	342	0.47
V_7	Systolic blood pressure (Sys)	100	210	0.50
V_8	Diastolic blood pressure (Dias)	50	120	0.50
V_9	Hypertension (Hyp)	0	1	0.50
V_{10}	Diabetes Mellitus (DM)	0	1	0.50
V_{11}	Cardiovascular Disease (CVD)	0	1	0.50
V_{12}	Stroke	0	1	0.50
V_{13}	Chronic Kidney Disease (CKD)	0	1	0.50

Table 2 The Cluster Center on Fuzzy Subtractive Clustering (FSC) Method

Cluster	Variables												
	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}	V_{11}	V_{12}	V_{13}
1	60	47.00	93.05	94.77	159.00	125	140	79	1	1	0	0	0
2	59	57.00	155.88	160.90	245.00	98	130	70	1	0	0	0	0
3	72	48.90	112.73	154.00	192.00	186	110	70	0	1	0	0	0
4	53	42.00	136.48	127.60	204.00	85	144	78	0	1	0	0	0
5	66	54.86	140.94	136.00	223.00	90	110	70	0	0	0	1	0
6	73	36.03	66.27	137.00	129.70	123	110	70	1	0	1	0	0
7	66	78.30	123.70	95.14	221.00	102	140	90	0	0	1	0	1
8	75	31.00	97.28	290.10	186.30	190	100	50	0	1	0	1	1
9	51	37.00	86.96	336.70	191.30	234	200	120	1	1	1	0	0
10	45	69.69	150.00	202.00	260.00	100	140	90	1	0	1	1	0
11	66	32.00	71.54	275.30	158.60	104	160	87	0	1	1	0	1
12	56	31.20	111.30	177.00	146.70	83	130	90	0	0	0	0	1
13	46	27.36	96.00	335.00	191.00	132	170	100	1	1	0	1	0
14	65	49.00	275.86	312.80	387.20	114	133	73	1	1	0	0	0
15	60	31.00	160.92	108.40	213.60	175	110	90	0	0	1	1	0
16	56	40.84	153.00	163.00	227.00	113	210	110	0	0	0	1	0
17	59	33.60	80.00	154.00	145.00	260	116	86	0	1	1	0	0
18	59	59.70	102.60	493.30	261.00	272	120	80	1	1	0	0	0
19	65	35.00	145.26	163.70	193.00	90	128	69	1	1	0	0	1
20	68	34.25	109.00	253.00	194.00	175	130	90	0	0	1	0	0
21	64	54.30	144.07	88.71	216.00	103	120	80	1	1	0	1	0
22	73	34.74	110.00	72.00	159.00	171	110	70	1	1	1	0	0
23	75	62.50	55.50	85.00	135.00	113	190	90	1	0	0	0	0

Table 3 The Cluster Center on Fuzzy C-Means (FCM) Method

Cluster	Variables												
	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}	V_{11}	V_{12}	V_{13}
1	67	34.23	108.42	254.93	193.77	177	130	88	0	0	1	0	0
2	61	47.53	91.20	97.28	158.10	123	138	79	1	1	0	0	0
3	62	53.33	143.18	90.70	215.70	106	122	81	1	1	0	1	0
4	62	62.68	97.42	97.63	179.45	141	131	74	1	1	0	0	0
5	51	52.13	177.06	139.94	257.17	312	131	81	0	1	0	0	0
6	59	59.62	102.74	492.48	260.89	272	120	80	1	1	0	0	0
7	62	49.07	126.29	209.65	217.15	129	121	80	0	1	0	0	1
8	46	67.89	150.85	201.65	259.19	101	141	89	1	0	1	1	0
9	54	45.86	179.93	216.90	270.77	114	133	89	0	1	0	0	0
10	61	34.02	99.62	456.75	224.94	138	130	90	1	1	0	0	0
11	70	42.22	126.74	127.40	194.41	189	129	88	0	1	0	0	0
12	60	55.72	156.26	161.51	244.16	98	132	70	1	0	0	0	0
13	64	54.01	103.04	104.35	177.87	121	111	71	1	0	1	0	0
14	55	42.83	135.90	129.01	203.88	89	142	78	0	1	0	0	0
15	75	40.05	152.44	224.21	237.45	147	112	70	0	1	0	0	0
16	54	43.59	120.32	237.80	211.74	166	121	80	0	1	0	0	0
17	59	51.42	186.64	227.47	283.35	193	140	80	0	1	0	0	0
18	64	50.25	97.98	136.92	175.05	127	122	88	1	0	0	0	0
19	53	41.67	122.55	127.73	189.54	133	137	84	1	1	0	0	0
20	71	48.15	113.18	152.77	191.46	186	111	71	0	1	0	0	0
21	56	39.69	130.09	241.49	218.09	179	140	89	1	1	0	0	0
22	61	64.49	108.67	332.53	239.65	180	161	100	0	0	0	1	0
23	76	43.21	157.33	153.77	231.02	114	131	80	1	1	0	0	0

Table 4 The Result of the ANOVA Test, T-Test, and Mean Square Error (MSE)

Probability of missing data (p)	Number of missing data	Level of significance – ANOVA test (p-value)	Pearson correlation coefficient on paired two-sample t-test		Mean Square Error (MSE)	
			FSC	FCM	FSC	FCM
10%	31	0.995	0.991	0.995	0.8842	0.6987
20%	75	0.989	0.940	0.972	1.6680	1.6508
30%	124	0.943	0.920	0.967	1.8586	1.8460
40%	169	0.914	0.860	0.933	2.5440	2.3586
50%	210	0.798	0.831	0.910	2.6942	2.6284

Note: Fuzzy C-Means (FCM) and Fuzzy Subtractive Clustering (FSC)

Table 5 The Evaluated Datasets

No	HDL	LDL	TG	Chol	FBS	Sys	Dias
1	60.46	180.00	209.00	282.00	-	114	67
2	55.90	153.40	81.00	226.00	-	122	80
3	56.27	134.41	197.00	230.00	216	-	-
4	46.25	150.35	192.00	235.00	-	120	80
5	74.40	136.82	120.00	235.00	-	142	70
6	-	63.18	-	-	126	140	60
7	39.67	141.53	299.00	240.00	-	-	-
8	57.15	180.05	109.00	259.00	-	130	70
9	45.14	103.00	159.00	180.00	-	120	90
10	72.47	121.00	208.00	235.00	-	120	70
11	59.95	205.00	157.00	296.00	-	108	59
12	50.86	136.00	256.00	239.00	-	-	-
13	55.20	-	309.90	191.00	194	130	80
14	44.60	51.00	383.00	171.00	-	-	-
15	60.10	121.30	83.00	198.00	-	110	80
16	30.10	153.22	260.70	236.00	-	120	70
17	41.00	120.81	128.00	187.00	114	-	-
18	41.40	194.59	289.30	294.00	-	140	80
19	38.00	79.84	389.90	196.00	-	120	80
20	56.57	214.00	329.00	336.00	-	140	70
21	60.90	118.28	89.11	197.00	-	120	80
22	56.40	76.72	94.27	152.00	-	110	70
23	44.20	134.92	173.10	214.00	143	-	-
24	43.00	122.60	222.40	210.00	96	-	-
25	45.00	166.27	260.60	263.00	-	130	80
26	35.00	95.86	236.40	178.00	-	140	80
27	44.00	97.64	124.80	166.60	-	120	60
28	49.00	207.58	88.72	275.00	-	140	90
29	42.21	109.00	184.00	188.00	-	-	-
30	32.00	-	548.60	234.60	164	130	80
31	44.00	138.84	164.30	215.70	-	130	90
32	51.00	137.76	120.70	212.90	-	120	80
33	71.00	155.35	65.15	239.00	-	100	70
34	37.55	154.00	111.00	213.00	124	-	110
35	40.00	-	423.90	259.40	171	140	90
36	-	57.52	112.40	-	259	140	70
37	38.00	156.89	156.20	226.00	-	150	100
38	37.00	117.86	158.20	186.00	65	-	-
39	-	-	452.40	212.80	214	130	80
40	36.00	71.86	199.70	147.80	-	130	90
41	48.35	96.93	82.60	-	-	100	60
42	39.19	137.71	104.00	197.70	-	175	93
43	-	50.34	35.69	69.48	-	-	-

Table 6 The Completed Evaluated Dataset

No	Cluster number	HDL	LDL	TG	Chol	FBS	Sys	Dias
1	12	60.46	180.00	209.00	282.00	98	114	67
2	13	55.90	153.40	81.00	226.00	121	122	80
3	16	56.27	134.41	197.00	230.00	216	121	80
4	1	46.25	150.35	192.00	235.00	177	120	80
5	12	74.40	136.82	120.00	235.00	98	142	70
6	2	47.53	63.18	97.28	158.10	126	140	60
7	15	39.67	141.53	299.00	240.00	147	112	70
8	5	57.15	180.05	109.00	259.00	312	130	70
9	22	45.14	103.00	159.00	180.00	180	120	90
10	12	72.47	121.00	208.00	235.00	98	120	70
11	12	59.95	205.00	157.00	296.00	98	108	59
12	16	50.86	136.00	256.00	239.00	166	121	80
13	17	55.20	186.64	309.90	191.00	194	130	80
14	16	44.60	51.00	383.00	171.00	166	121	80
15	22	60.10	121.30	83.00	198.00	180	110	80
16	1	30.10	153.22	260.70	236.00	177	120	70
17	18	41.00	120.81	128.00	187.00	114	122	88
18	12	41.40	194.59	289.30	294.00	98	140	80
19	10	38.00	79.84	389.90	196.00	138	120	80
20	17	56.57	214.00	329.00	336.00	193	140	70
21	18	60.90	118.28	89.11	197.00	127	120	80
22	18	56.40	76.72	94.27	152.00	127	110	70
23	3	44.20	134.92	173.10	214.00	143	122	81
24	3	43.00	122.60	222.40	210.00	96	122	81
25	7	45.00	166.27	260.60	263.00	129	130	80
26	16	35.00	95.86	236.40	178.00	166	140	80
27	22	44.00	97.64	124.80	166.60	180	120	60
28	12	49.00	207.58	88.72	275.00	98	140	90
29	19	42.21	109.00	184.00	188.00	133	137	84
30	10	32.00	99.62	548.60	234.60	164	130	80
31	18	44.00	138.84	164.30	215.70	127	130	90
32	19	51.00	137.76	120.70	212.90	133	120	80
33	12	71.00	155.35	65.15	239.00	98	100	70
34	22	37.55	154.00	111.00	213.00	124	161	110
35	10	40.00	99.62	423.90	259.40	171	140	90
36	2	47.53	57.52	112.40	158.10	259	140	70
37	22	38.00	156.89	156.20	226.00	180	150	100
38	13	37.00	117.86	158.20	186.00	65	111	71
39	10	34.02	99.62	452.40	212.80	214	130	80
40	11	36.00	71.86	199.70	147.80	189	130	90
41	3	48.35	96.93	82.60	215.70	106	100	60
42	7	39.19	137.71	104.00	197.70	129	175	93
43	2	47.53	50.34	35.69	69.48	123	138	79