The 2nd FORSTAT International Conference

Implementation of The Sequential Biclustering Method and Centroid-Based Imputation (Fuzzy C-Means) for Missing Values Imputation

Yolanda Azzahra[a], Titin Siswantining[b*], Setia Pramana[c]

[a,b] Department of Mathematics, University of Indonesia, Depok, Indonesia.
[c] Department of Statistics, Politeknik Statistika STIS, Jakarta, Indonesia.

*Corresponding author: titin@sci.ui.ac.id

A B S T R A C T

Missing values are a common issue in gene expression data and may significantly affect the accuracy of subsequent analyses if not properly handled. This study aims to implement and evaluate a missing value imputation method based on Sequential Biclustering and Centroid-Based Imputation with Fuzzy C-Means (FCM) on gene expression data of Type 2 Diabetes Mellitus patients. Missing values were generated under the Missing Completely At Random (MCAR) mechanism with missing rate ranging from 5% to 55%, and five replications were conducted for each missing rate. Biclustering was performed using the Cheng and Church algorithm to identify biclusters based on the Mean Squared Residue (MSR), followed by missing value estimation using FCM with 2 to 5 biclusters. The performance of the imputation method was evaluated using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) by comparing the imputed values with the original data. The results indicate that the configuration with 2 biclusters consistently produces the lowest and most stable MSE, RMSE, and MAE across all missing rates.

Keywords: Cheng and Church; Mean Squared Residue; Missing Completely At Random; Missing Rate; Root Mean Squared Error

1. Introduction

Type 2 Diabetes Mellitus (T2DM) is a chronic metabolic disease whose prevalence continues to increase globally, making it a serious public health problem in many countries. The latest report from the International Diabetes Federation (IDF), 11th edition in 2025, states that there are currently approximately 589 million adults (aged 20-79 years) worldwide living with diabetes. This number is projected to rise dramatically to 853 million by 2050, meaning that 1 in 8 adults is expected to suffer from diabetes if no significant health interventions are implemented. Diabetes is also a major cause of mortality, with an estimates 3.4 million deaths annually. With its continuously increasing prevalence, T2DM is not only an individual health issue but also a major challenge for global healthcare systems, driven by lifestyle changes, urbanization, and nutritional transitions [1].

In this context, gene expression analysis has become increasingly important to better understand the biological mechanisms of T2DM and to support the development of more precise prevention and treatment strategies. However, gene expression data obtained from platforms such as Gene Expression

Omnibus (GEO) or RNA sequencing are typically large-scale, complex, and often contain missing values due to technical limitations in biological experiments. The presence if missing values can disrupt analyses such as clustering, biclustering, and gene identification, and may lead to biased or misleading conclusions if not properly handled [2]. Therefore, imputation methods are required to estimate missing values in gene expression data so that analysis results remain valid and reliable.

Missing values in gene expression data are generally categorized into Missing Completely at Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR) [3]. In studies of gene expression data from T2DM patients, missing values are often assumed to follow the MCAR mechanism due to signal loss or technical failures in microarray or RNA-seq platforms. Several studies have shown that a high proportion of missing values in gene expression data can reduce sensitivity in detecting differentially expressed genes and gene interactions [4]. This highlights that selecting an appropriate imputation strategy is crucial to maintaining the quality of genomic analysis results in T2DM studies.

Various imputation methods have been developed to handle missing values, ranging from simple approaches to more complex machine learning and statistical models. Simple methods such as mean and median imputation tend to ignore the underlying data structure [5]. To address these limitations, several studies have developed biclustering-based imputation methods. Several biclustering-based imputation methods have been developed to address missing values in gene expression data. The SBi-MSREimpute method proposed in [6] successfully generated coherent biclusters using Mean Squared Residue and Euclidean Distance but relied on weighted-average estimation. Furthermore, the FuBiCMPSO method presented in [7] enhanced imputation accuracy through the integration of fuzzy centroids and Particle Swarm Optimization. Nevertheless, these approaches did not simultaneously incorporate sequential bicluster extraction and centroid-based imputation using Fuzzy C-Means while evaluating performance using MSE, RMSE, and MAE. Therefore, a hybrid framework integrating Sequential Biclustering and Fuzzy C-Means is proposed in this study.

In this study, a hybrid approach is proposed by combining Sequential Biclustering with Centroid-Based Imputation using the Fuzzy C-Means (FCM) algorithm. Sequential Biclustering is employed to identify local structures in gene expression data, where variation patterns often appear within specific subsets of genes and conditions rather than globally. This method extracts coherent biclusters sequentially by eliminating rows or columns with the highest residue contribution, ensuring that missing value estimation is performed on homogeneous subsets. The quality of biclusters is evaluated using Mean Squared Residue (MSR), where lower MSR values indicate higher coherence within biclusters [8], [9], [10].

After bicluster formation, missing values are estimated using Fuzzy C-Means (FCM), a fuzzy clustering algorithm that assigns data points to multiple clusters with certain membership degrees [11]. This approach allows flexible handling of uncertainly in biological data and provides adaptive estimation through cluster centroid [12]. Several studies have shown that FCM have demonstrated competitive performance in imputing missing values in gene expression data [13], [14]. The integration of biclustering and FCM is important because biclustering alone cannot provide precise numerical estimation, while FCM without biclustering may lose important local structural information. By combining both methods, the imputation process can capture coherent patterns while producing smoother and more realistic estimates. Previous studies have also demonstrated the effectiveness of hybrid approaches have been shown to improve data quality for downstream analysis [15]. The quality of imputation results is evaluated using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). These metrics are used to assess the accuracy of estimated values, in contrast to MSR, which is used during bicluster formations. Therefore, this study aims to implement and evaluate a hybrid method combining Sequential Biclustering and Centroid-Based Imputation (FCM) for imputing missing values in gene expression data of T2DM patients, with the expectation of improving data quality and supporting more reliable biological analysis.

2. Theoritical Framework

2.1. Missing Values

Missing values refer to condition where some observations in a dataset are not recorded completely. The presence of missing values is common across various research fields and may lead to biased estimation, invalid conclusions, and reduced representativeness of the sample. According to [16], missing values can be classified into three categories, Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR). MCAR occurs when missingness is completely independent of any variables in the dataset. MAR occurs when the probability of missingness depends on observed variables, while MNAR occurs when missingness depends on the unobserved values themselves.

2.2. Min-Max Normalization

Gene expression data typically contain values with large numerical ranges, which may affect the performance of analytical methods. Therefore, normalization is required to scale the data into a uniform range. Min-max normalization is commonly used method that transforms the data into a predefined interval, typically $[0,1]$, without altering the original data distribution [17]. Mathematically, the formula for min-max normalization is:

$$x'_{ik} = \frac{x_{ik} - min(x_i)}{max(x_i) - min(x_i)} \tag{1}$$

where x'_{ik} represents the normalized value, x_{ik} is the original value, $min(x_i)$ and $max(x_i)$ denote the minimum and maximum values of row i, respectively.

2.3. Clustering Analysis

Clustering is a data analysis technique used to group objects into clusters such that objects within the same cluster are more similar to each other than to those in different clusters [18]. In gene expression analysis, clustering is widely used to identify groups of genes with similar expression patterns across experimental conditions. The data are typically represented as a matrix where rows correspond to genes and columns correspond to conditions [19].

2.4. Biclustering Analysis

Biclustering extends traditional clustering by simultaneously grouping rows and columns of a data matrix. This allows the identification of subsets of genes that exhibit similar behavior under specific subsets of conditions [8].

2.5. Cheng and Church Biclustering Model

The Cheng and Church model is one of the earliest biclustering methods designed for gene expression data [20]. This model evaluates the quality of a bicluster using the Mean Squared Residue (MSR), which measures the coherence of values within the bicluster. The residue of an element is defined as:

$$R(c_{ij}) = c_{ij} - c_{iJ} - c_{Ij} + c_{IJ} \tag{2}$$

where c_{iJ} represents the mean of row, c_{Ij} the mean of row, and c_{IJ} the overall mean of the data.

The MSR value of a bicluster is calculated as:

$$H(\boldsymbol{C}) = \frac{1}{|I||J|} \sum_{i \in I, j \in J} R^2(c_{ij}) \tag{3}$$

where $|I|, |J|$ represents the number of rows and columns in the data.

2.6. Centroid-Based Imputation using Fuzzy C-Means (FCM)

Centroid-based imputation estimates missing values using cluster centroids. In this study, the Fuzzy C-Means (FCM) algorithm is used due to its ability to assign data points to multiple clusters with varying degrees of membership. The centroid of cluster is calculates as:

$$c_{k,j} = \frac{\sum_{i \in I} \mu_{i,k}^{m} x_{i,j}}{\sum_{i \in I} \mu_{i,k}^{m}} \tag{4}$$

where $c_{k,j}$ is the centroid of the k-th cluster for the j-th column in the bicluster, $x_{i,j}$ is the data value at the i-th column in the bicluster, $\mu_{i,k}$ is the membership degree of row i to cluster k, m is the fuzziness parameter $(m > 1)$, I denotes the set of row indices in the bicluster, and J denoted the set of column indices in the bicluster.

2.7. Mean Squared Error (MSE)

Mean Squared Error (MSE) is a commonly used metric to evaluate the accuracy of imputation results. It measures the average squared difference between the actual and estimated values [21]. The formula for MSE is:

$$MSE = \frac{1}{|I||J|} \sum_{i \in I} \sum_{j \in J} (x_{i,j} - \hat{x}_{i,j})^2 \tag{5}$$

where $x_{i,j}$ represents the true value and $\hat{x}_{i,j}$ is the predicted value.

2.8. Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) is a widely used metric to measure the magnitude of prediction error. It represents the square root of the average squared differences between the actual values and estimated values. RMSE provides an interpretable measure of error in the same unit as the original data and is more sensitive to large errors due to the squaring operation [21]. The formula for RMSE is:

$$RMSE = \sqrt{\frac{1}{|I||J|} \sum_{i \in I} \sum_{j \in J} (x_{i,j} - \hat{x}_{i,j})^2} \tag{6}$$

where $x_{i,j}$ represents the true value and $\hat{x}_{i,j}$ is the predicted value.

2.9. Mean Absolute Error (MAE)

The Mean Absolute Error represents the average magnitude of the errors in a set of predictions without considering their direction. It is calculated in Equation [22]:

$$MAE = \frac{1}{|I||J|} \sum_{i \in I} \sum_{j \in J} \left| x_{i,j} - \hat{x}_{i,j} \right| \tag{7}$$

where $x_{i,j}$ represents the true value and $\hat{x}_{i,j}$ is the predicted value.

3. Methods

3.1. Data

This study employed gene expression data from patients with Type 2 Diabetes Mellitus (T2DM) obtained from the Gene Expression Omnibus (GEO) database under accession number GSE278204. The dataset consists of 13,466 genes and 30 experimental conditions represented in the form of a gene expression matrix. Gene expression data were selected because they commonly contain missing values

arising from technical limitations during data acquisition, making them suitable for evaluating missing value imputation methods.

Prior to analysis, data preprocessing was conducted to remove irrelevant information and prepare the dataset for subsequent analysis. Min-Max normalization was then applied to transform all gene expression values into the range of 0–1, ensuring a uniform scale among variables and preventing attributes with large numerical values from dominating the biclustering and clustering processes.

To evaluate the performance of the proposed imputation method, missing values were artificially generated using the Missing Completely At Random (MCAR) mechanism. Missing rates of 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, and 55% were considered. Each missing-rate scenario was replicated five times to obtain stable and reliable performance measurements.

3.2. Research Design

This study applied a quantitative experimental design to evaluate the effectiveness of a hybrid imputation method that combines Sequential Biclustering and Centroid-Based Imputation using the Fuzzy C-Means (FCM) algorithm. The research focused on assessing the ability of the proposed method to estimate missing values in gene expression data while preserving the underlying local data structure. The study utilized complete gene expression data as the reference dataset. Artificial missing values were generated under the MCAR mechanism, and the resulting incomplete datasets were subsequently imputed using the proposed method. The estimated values were then compared with the original values to assess imputation accuracy. The variables considered in this study consisted of the original values ($x_{i,j}$), the generated missing values, and the imputed values ($\hat{x}_{i,j}$). These variables were used to evaluate the performance of the imputation method through several error metrics.

3.3. Imputation Procedure

The proposed imputation framework integrates Sequential Biclustering based on the Cheng and Church model with Centroid-Based Imputation using Fuzzy C-Means. The imputation process began by generating missing values under the MCAR mechanism. To facilitate bicluster formation, temporary mean imputation was applied to missing entries during the biclustering stage. Sequential Biclustering was then performed using the Cheng and Church algorithm to identify coherent subsets of genes and conditions. The quality of each bicluster was evaluated using the Mean Squared Residue (MSR) criterion, and rows or columns with the highest residue contribution were iteratively removed until a coherent bicluster was obtained. After bicluster formation, the Fuzzy C-Means algorithm was applied within each bicluster to determine cluster membership degrees and centroid values. Missing values were subsequently estimated using centroid-based imputation weighted by fuzzy membership values. This approach enables the estimation process to account for uncertainty and local similarity patterns in the gene expression data.

3.4. Performance Evaluation

The performance of the proposed method was evaluated by comparing the imputed values with the corresponding original values. Three commonly used error metrics were employed, namely Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). Lower values of these metrics indicate higher imputation accuracy and better preservation of the original data structure. The evaluation was conducted for each missing-rate scenario and each bicluster configuration. The average values obtained from five replications were used as the final performance measures to ensure the stability and reliability of the experimental results.

4. Results and Discussion

The performance of the proposed Sequential Biclustering–FCM imputation method was evaluated under different missing rates ranging from 5% to 55% and different bicluster configurations consisting of 2, 3, 4, and 5 biclusters. The evaluation was conducted using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). The average results obtained from five replications for each missing-rate scenario are presented in Table 1.

Table 1. Average Results of MSE, RMSE, and MAE

Missing Rate (%)	Average Error Metrics	2 Biclusters	3 Biclusters	4 Biclusters	5 Biclusters
5%	MSE	0.000050375	0.000087344	0.000093058	0.000093104
5%	RMSE	0.006519842	0.008554819	0.008863959	0.008865604
5%	MAE	0.000154472	0.000178108	0.000173273	0.00017059
10%	MSE	0.000057738	0.000113101	0.000114977	0.000115243
10%	RMSE	0.007317887	0.010396052	0.010494192	0.010505145
10%	MAE	0.000178986	0.000217612	0.000209076	0.000206731
15%	MSE	0.000048724	0.000099946	0.000101283	0.000101475
15%	RMSE	0.006766013	0.009756982	0.009833166	0.009841853
15%	MAE	0.000171526	0.000206201	0.000197329	0.000193724
20%	MSE	0.000061772	0.0001034054	0.000104311	0.000099508
20%	RMSE	0.007710421	0.010004522	0.01004729	0.009746283
20%	MAE	0.000188954	0.0002146281	0.000205732	0.000200555
25%	MSE	0.000067359	0.000107586	0.000108529	0.000108604
25%	RMSE	0.008177421	0.010287617	0.010331357	0.010334852
25%	MAE	0.000201612	0.000227221	0.00022132	0.000219847
30%	MSE	0.000075659	0.000110932	0.000111976	0.000112048
30%	RMSE	0.008659030	0.010478636	0.010526529	0.010529920
30%	MAE	0.000216698	0.000237528	0.000234375	0.000235651
35%	MSE	0.000082882	0.000111423	0.000112330	0.00011246
35%	RMSE	0.009085091	0.010541808	0.010584113	0.010590296
35%	MAE	0.000229477	0.000244599	0.000243490	0.000242578
40%	MSE	0.000091342	0.000108132	0.000109629	0.000109822
40%	RMSE	0.009510845	0.010380509	0.010450825	0.010459745
40%	MAE	0.000243804	0.00025284	0.000250535	0.000249806
45%	MSE	0.000085521	0.000104713	0.00010583	0.000105974
45%	RMSE	0.009219382	0.010211736	0.01026545	0.010272058
45%	MAE	0.000245905	0.000256995	0.000256278	0.000253429
50%	MSE	0.000097348	0.000110053	0.000111289	0.0001115073
50%	RMSE	0.009796617	0.010462386	0.010521578	0.010531308
50%	MAE	0.000259939	0.000272054	0.00026554	0.000263005
55%	MSE	0.000093818	0.000104634	0.00010007	0.000105935
55%	RMSE	0.009625339	0.01020110	0.00992793	0.010263919
55%	MAE	0.000263969	0.00027505	0.000274017	0.000269416

The results presented in Table 1 indicate that the proposed method was able to maintain relatively low error values across all missing-rate scenarios. Although MSE. RMSE. and MAE generally increased as the percentage of missing values increased. the magnitude of the increase remained relatively small. This finding suggests that the proposed hybrid method is robust in handling incomplete gene expression data. even when more than half of the observations are missing. A notable result is the consistent superiority of the two-bicluster configuration. Across all missing-rate levels. the use of two biclusters produced the lowest MSE. RMSE. and MAE values compared with the three-. four-. and five-bicluster configurations. For instance. at a missing rate of 5%. the MSE obtained using two biclusters was 0.000050375. whereas the corresponding values for three. four. and five biclusters were 0.000087344. 0.000093058. and 0.000093104. respectively. Similar patterns were observed for RMSE and MAE. indicating that increasing the number of biclusters did not necessarily improve imputation accuracy. This phenomenon may be explained by the characteristics of gene expression data. A smaller number of biclusters allows the Cheng and Church algorithm to capture broader and more coherent local patterns among genes and experimental conditions. Conversely. increasing the number of biclusters tends to partition the data into smaller subsets. potentially reducing bicluster coherence and affecting the quality of centroid estimation during the Fuzzy C-Means stage. Consequently. the imputation process becomes less stable and produces larger prediction errors. The effectiveness of the proposed approach can also be attributed to the complementary roles of Sequential Biclustering and Fuzzy C-Means. Sequential

Biclustering identifies local structures within subsets of genes and conditions based on the Mean Squared Residue criterion. while Fuzzy C-Means provides flexible estimation through fuzzy membership values. Unlike conventional clustering methods that assign observations exclusively to a single cluster. Fuzzy C-Means allows genes to belong to multiple clusters with varying membership degrees. thereby producing smoother and more adaptive estimations.

The findings of this study are consistent with previous studies reporting that biclustering-based approaches are effective for identifying local gene expression patterns and improving missing value estimation [8]-[10]. [15]. Similarly. fuzzy clustering techniques have been shown to provide reliable imputation performance due to their ability to accommodate uncertainty in biological data [11]-[14]. However. most previous studies applied biclustering and clustering separately. The novelty of the present study lies in integrating Sequential Biclustering based on the Cheng and Church model with Centroid-Based Imputation using Fuzzy C-Means. This integration enables the method to simultaneously preserve local gene-expression structures and improve estimation accuracy. Another important finding is the stability of the proposed method across increasing missing rates. Even at a missing rate of 55%. the MSE remained below 0.0001 for the two-bicluster configuration. This result indicates that the proposed method can effectively recover missing information while preserving the underlying structure of high-dimensional gene expression data. Such robustness is particularly important for biological datasets. where missing values frequently occur due to experimental limitations and measurement errors. Overall. the experimental results demonstrate that the proposed Sequential Biclustering–FCM framework is capable of producing accurate and stable imputations for gene expression data. The method provides a promising alternative for handling missing values in high-dimensional biological datasets and may support downstream analyses such as gene clustering. biomarker discovery. and disease characterization in Type 2 Diabetes Mellitus studies.

5. Conclusion

This study implemented a hybrid missing value imputation framework that integrates Sequential Biclustering based on the Cheng and Church model with Centroid-Based Imputation using the Fuzzy C-Means algorithm for gene expression data of Type 2 Diabetes Mellitus patients. The results demonstrate that the proposed method is capable of producing accurate and stable imputations across various missing-rate scenarios. The integration of biclustering and fuzzy clustering effectively preserves local gene expression patterns while providing reliable estimation of missing values. The main finding of this study indicates that the two-bicluster configuration achieves the best imputation performance. suggesting that a smaller number of coherent biclusters is more effective in maintaining the underlying structure of gene expression data. These findings highlight the potential of the proposed Sequential Biclustering–FCM framework as an alternative approach for handling missing values in high-dimensional biological datasets and supporting subsequent bioinformatics analyses. Future studies may evaluate the proposed method on other omics datasets and compare its performance with advanced imputation techniques. including deep learning-based approaches.

REFERENCE

[1] International Diabetes Federation. *IDF Diabetes Atlas*. 11th ed. Brussels. Belgium: International Diabetes Federation. 2025. doi: https://diabetesatlas.org.

[2] T. Aittokallio. “Dealing with missing values in large-scale studies: microarray data imputation and beyond.” *Briefings in Bioinformatics*. vol. 11. no. 2. pp. 253–264. 2010. doi: https://doi.org/10.1093/bib/bbp059.

[3] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. 3rd ed. Hoboken. NJ. USA: Wiley. 2019. doi: https://doi.org/10.1002/9781119482260.

[4] O. Troyanskaya. M. Cantor. G. Sherlock. P. Brown. T. Hastie. R. Tibshirani. D. Botstein. and R. B. Altman. “Missing value estimation methods for DNA microarrays.” *Bioinformatics*. vol. 17. no. 6. pp. 520–525. 2001. doi: https://doi.org/10.1093/bioinformatics/17.6.520.

[5] H. Junninen. H. Niska. K. Tuppurainen. J. Ruuskanen. and M. Kolehmainen. “Methods for imputation of missing values in air quality data sets.” *Atmospheric Environment*. vol. 38. no. 18. pp. 2895–2907. 2004. doi: https://doi.org/10.1016/j.atmosenv.2004.02.026.

[6] A. D. Putri. "Sequential Biclustering Based on Mean Square Residue and Euclidean Distance for Missing Value Imputation in Gene Expression Data." Undergraduate Thesis. Department of Mathematics. Universitas Indonesia. Depok. Indonesia. 2021.

[7] N. S. Sianipar. "Fuzzy Biclustering Means with Particle Swarm Optimization for Missing Value Imputation in Gene Expression Data." Undergraduate Thesis. Department of Mathematics. Universitas Indonesia. Depok. Indonesia. 2025.

[8] S. C. Madeira and A. L. Oliveira. "Biclustering algorithms for biological data analysis: A survey." *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. vol. 1. no. 1. pp. 24–45. 2004. doi: https://doi.org/10.1109/TCBB.2004.2.

[9] B. Pontes. R. Giráldez. and J. S. Aguilar-Ruiz. "Biclustering on expression data: A review." *Journal of Biomedical Informatics*. vol. 57. pp. 163–180. 2015. doi: https://doi.org/10.1016/j.jbi.2015.06.028.

[10] R. Henriques and S. C. Madeira. "Biclustering and Triclustering: A review of computational approaches and applications." *Journal of Biomedical Informatics*. vol. 88. pp. 28–46. 2018.

[11] J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York. NY. USA: Plenum Press. 1981. doi: https://doi.org/10.1007/978-1-4757-0450-1.

[12] D. Dembélé and P. Kastner. "Fuzzy C-means method for clustering microarray data." *Bioinformatics*. vol. 19. no. 8. pp. 973–980. 2003. doi: https://doi.org/10.1093/bioinformatics/btg119.

[13] F. Meng. H. Gong. and Y. Wang. "A Fuzzy C-Means Based Missing Value Imputation Method for Gene Expression Data." *Applied Soft Computing*. vol. 18. pp. 155–165. 2014.

[14] Z. Cai. X. Wang. and Y. Yin. "Gene expression data imputation using fuzzy clustering techniques." *Computational Biology and Chemistry*. vol. 65. pp. 75–82. 2016.

[15] K. Eren. M. Deveci. O. Küçüktunç. and Ü. V. Çatalyürek. "A comparative analysis of biclustering algorithms for gene expression data." *Briefings in Bioinformatics*. vol. 14. no. 3. pp. 279–292. 2013. doi: https://doi.org/10.1093/bib/bbs032.

[16] D. B. Rubin. "Inference and Missing Data." *Biometrika*. vol. 63. no. 3. pp. 581–592. 1976. doi: https://doi.org/10.1093/biomet/63.3.581.

[17] T. Adeyomo. S. A. Abdulhamid. and M. S. Ibrahim. "Performance Evaluation of Min-Max and Z-Score Data Normalization Techniques." *International Journal of Computer Network and Information Security*. vol. 10. no. 10. pp. 1–7. 2018.

[18] G. Gan. C. Ma. and J. Wu. *Data Clustering: Theory. Algorithms. and Applications*. 2nd ed. Philadelphia. PA. USA: SIAM. 2020. doi: https://doi.org/10.1137/1.9781611971507

[19] K. Y. Yeung. D. R. Haynor. and W. L. Ruzzo. "Validating clustering for gene expression data." *Bioinformatics*. vol. 17. no. 4. pp. 309–318. 2001. doi: https://doi.org/10.1093/bioinformatics/17.4.309.

[20] Y. Cheng and G. M. Church. "Biclustering of expression data." in *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology*. San Diego. CA. USA. 2000. pp. 93–103.

[21] T. Hastie. R. Tibshirani. and J. Friedman. *The Elements of Statistical Learning: Data Mining. Inference. and Prediction*. 2nd ed. New York. NY. USA: Springer. 2009. doi: https://doi.org/10.1007/978-0-387-84858-7.

[22] M. A. Khan. "A Comparative Study on Imputation Techniques: Introducing A Transformer Model for Robust and Efficient Handling of Missing EEG Amplitude Data." *Bioengineering*. vol. 11. no. 8. pp. 740. 2024. doi: https://doi.org/10.3390/bioengineering11080740.