Indonesian Journal of Statistics and Its Applications
DOI: https://doi.org/10.29244/ijsa.v10i1p37-46

RESEARCH ARTICLE

A Comparative Study of Sequential Biclustering and Fuzzy C-Means, K-Nearest Neighbors, and Mean Imputation for Missing Value Estimation in Gene Expression Data

Yolanda Azzahra[1], Titin Siswantining[*1], Setia Pramana[2], Mogana Darshini Ganggayah[3], and Alhadi Bustamam[1]

[1]Department of Mathematics, University of Indonesia, Indonesia
[2]Department of Statistics, Politeknik Statistika STIS, Indonesia
[3]Department of Econometrics and Business Statistics, School of Business, Monash University, Malaysia

OPEN ACCESS

Check for updates

Article History
Received 08 May 2026
Revised 18 June 2026
Accepted 30 June 2026

Keywords
biclustering, gene expression data, mean squared residue

Corresponding Author:
Titin Siswantining

Email:
titin@sci.ui.ac.id

Department of Mathematics, University of Indonesia, Depok, West Java, Indonesia.

ABSTRACT
Missing values are a common issue in gene expression data and can significantly affect downstream analysis. This study aims to compare the performance of a hybrid sequential biclustering and centroid-based clustering method with conventional imputation approaches for handling missing values. The proposed method integrates sequential biclustering based on mean squared residue to identify coherent submatrices, followed by centroid-based clustering to estimate missing entries. The dataset used in this study consists of gene expression data of patients with type 2 diabetes mellitus, with missing values introduced under various proportions ranging from 5% to 55%. The performance of the proposed method is evaluated and compared with mean imputation and nearest neighbor imputation using mean squared error, root mean squared error, and mean absolute error. The experimental results show that the proposed method consistently achieves lower Mean Squared Error and Root Mean Squared Error values across all evaluated missing rates. Although K-Nearest Neighbors imputation generally produces lower Mean Absolute Error values, the proposed method demonstrates superior performance in reducing larger estimation errors and preserving local structural patterns within the data. These findings indicate that integrating biclustering and clustering techniques provides a robust framework for missing value estimation in high-dimensional gene expression datasets and contributes to improving the reliability of downstream biological analyses.

Introduction

Missing values are a common issue in high-dimensional biological data, particularly in gene expression datasets, where they can significantly affect downstream analysis and interpretation. Incomplete data may arise from experimental limitations, noise, or technical errors, leading to biased statistical results if not properly handled. Therefore, accurate imputation methods are essential to preserve the data's underlying structure and ensure reliable biological insights (Troyanskaya et al., 2001; Aittokallio, 2009).

Various imputation methods have been proposed, ranging from simple statistical approaches such as mean imputation to more advanced machine learning techniques, such as k-nearest neighbors and matrix factorization. Mean imputation is computationally efficient but often ignores data variability, while k-nearest neighbors can capture local similarity but may struggle with high-dimensional sparse data (Troyanskaya et al., 2001). Recent studies have shown that clustering-based and hybrid approaches can improve imputation accuracy by exploiting local similarity structures within the data (Dubey &

© 2026 Azzahra et al. This is an open-access article distributed under the terms of the Creative Commons Attribution (CC BY) license, allowing unrestricted use, distribution, and reproduction in any medium, provided proper credit is given to the original authors.
Think twice before printing this journal paper. Save paper, trees, and Earth!

Rasool, 2021; Zhu et al., 2021). Several comparative studies have investigated the performance of missing value imputation methods in biological and healthcare datasets. The findings generally indicate that K-Nearest Neighbors imputation performs better than simple statistical approaches because it utilizes neighborhood information and local similarity among observations. However, its effectiveness may decrease in high-dimensional datasets where distance measures become less discriminative (Zhang, 2012; Mastika et al., 2025; Sudrajat & Cholid, 2023; Laia, 2023; Pamuji et al., 2024). Furthermore, machine learning-based imputation methods have been reported to improve predictive performance compared with conventional approaches in complex datasets containing incomplete observations (Jerez et al., 2010; García-Laencina et al., 2010; Joel et al., 2025).

Biclustering is a powerful technique that simultaneously groups rows and columns to identify local patterns within subsets of data. The Cheng and Church algorithm is one of the most widely used biclustering methods, utilizing the mean squared error to identify coherent submatrices (Cheng & Church, 2000). By focusing on local structures rather than global patterns, biclustering is particularly suitable for gene expression data, where only subsets of genes are co-expressed under specific conditions (Madeira & Oliveira, 2004).

Recent comparative studies and review papers have highlighted the growing importance of biclustering techniques for gene expression analysis. Unlike traditional clustering methods, biclustering can identify local relationships between subsets of genes and experimental conditions, making it particularly suitable for high-dimensional biological datasets. Comparative evaluations have demonstrated that biclustering approaches often provide more meaningful biological interpretations than conventional clustering techniques when local structures are present in the data (Nicholls & Wallace, 2021; Castanho et al., 2024).

Several recent studies have further demonstrated the effectiveness of biclustering approaches in gene expression analysis and missing value estimation. Previous studies showed that biclustering methods are capable of preserving meaningful local gene relationships in high-dimensional biological data and improving pattern discovery in gene expression analysis (Siswantining et al., 2021). Recent studies on biclustering-based imputation methods also reported that exploiting local coherence structures within gene expression matrices can improve the accuracy of missing value estimation in biological datasets (Silalahi et al., 2025). These findings indicate that biclustering-based frameworks have strong potential for improving missing value imputation in gene expression analysis.

On the other hand, centroid-based clustering methods such as fuzzy clustering provide a flexible framework for handling uncertainty in data. Unlike hard clustering, fuzzy clustering assigns membership degrees to each data point, allowing a more nuanced representation of data structure. This property makes it particularly suitable for imputing missing values, as it can capture overlapping patterns within the data (Bezdek et al., 1984).

Despite the advantages of biclustering and fuzzy clustering individually, their combined application for missing values imputation remains relatively underexplored. In particular, integrating sequential biclustering with centroid-based imputation can improve imputation accuracy by leveraging both local pattern discovery and a soft-clustering mechanism. Furthermore, there is a need to systematically compare such hybrid approaches with conventional imputation methods to evaluate their effectiveness across different levels of missing data (Yosboon et al., 2024).

Therefore, this study aims to compare the performance of a hybrid sequential biclustering and fuzzy clustering approach with conventional methods, including mean imputation and k-nearest neighbors imputation, for handling missing values in gene expression data. The evaluation uses multiple error metrics across varying levels of missing data to provide a comprehensive assessment of method performance.

Material and Methods

Data Source and Preprocessing

The dataset used in this study consists of gene expression data obtained from the Gene Expression Omnibus repository with accession number GSE278204. The dataset contains high-dimensional gene expression measurements with thousands of genes and multiple samples. Prior to analysis, the data were cleaned by removing rows and columns containing only missing values and converting all entries into numeric format.

To ensure comparability across features, the dataset was normalized using min-max scaling, mapping all values to the range 0-1. This normalization step is essential for distance-based methods such as clustering and imputation, as it prevents features with larger scales from dominating the analysis.

Missing Data Simulation

Since the original dataset is complete, missing values were artificially introduced using the Missing Completely at Random mechanism. In this approach, missing entries are randomly distributed across the data matrix independent of observed or

unobserved values.

The proportion of missing data ranged from 5% to 55% in increments of 5%. For each missing rate, the experiment was repeated 5 times with different random seeds to ensure robustness and reduce the effects of randomness. This setup enables a comprehensive evaluation of imputation performance across varying levels of data incompleteness.

Proposed Method: Sequential Biclustering with Fuzzy Clustering Imputation

The proposed method combines sequential biclustering based on the Cheng and Church algorithm with centroid-based imputation using fuzzy clustering. The overall procedure consists of two main stages, namely bicluster identification and missing value estimation. In the first stage, sequential biclustering is applied to identify a coherent subset of columns from the gene expression matrix. The coherence of a biclsuter is evaluated using the Mean Squared Residue (MSR) criterion proposed by Cheng and Church (2000), which is defined as

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

where

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

denotes the residue of element (c_{ij}), (c_{iJ}) represents the mean of row, (c_{Ij}) represents the mean of column, and (c_{IJ}) denotes the overall mean of the bicluster. Before biclustering is performed, missing values are temporarily replaced using column mean imputation. The column means are calculated only from the available observed values while missing entries are ignored during the calculation. This temporary replacement is used solely to facilitate the computation of MSR values and bicluster construction.

The biclustering process begins with the complete set of columns. The MSR value is calculated ad compared with a predefined threshold of $(\delta = 0.005)$. if the MSR exceeds the threshold, the column with the largest residue contribution is removed. The MSR is then recalculated using the remaining columns, and the procedure is repeated iteratively until the MSR value of the remaining columns satisfies the threshold criterion. The term sequential refers to this stepwise process of recalculating MSR and removing columns iteratively. The proposed approach generates a single final bicluster rather than multiple biclusters. The final bicluster is defined as the subset of columns remaining after the stopping criterion is satisfied and represents the most coherent local structure identified in the gene expression data. Because the objective of this study is to evaluate the imputation performance within the most coherent feature subset, only the final bicluster satisfying the MSR criterion is retained for subsequent analysis. This bicluster is subsequently used as the analysis space for all imputation experiments.

After the final bicluster has been identified, artificial missing values are introduced using the Missing Completely at Random (MCAR) mechanism. Missing rates ranging from 5% to 55% with increments of 5% are considered, and each missing rate is replicated five times using different random seeds. The missing-value simulation is performed only after the biclustering stage has been completed. Therefore, the biclustering procedure does not use any of the values intentionally removed during evaluation, preventing information leakage throughout the experimental process.

In the second stage, missing values are estimated using the Fuzzy C-Means clustering algorithm (Bezdek, 1981). The number of clusters is set to (c = 2), and the fuzziness parameter is fixed at (m = 2). Cluster centroids are computed according to

$$v_{kj} = \frac{\sum_{i\in I}(\mu_{ik})^m x_{ij}}{\sum_{i\in I}(\mu_{ik})^m} \tag{3}$$

where (v_{kj}) denotes the centroid value of variable in cluster, μ_{ik} represents the membership degree of observation in cluster, and m is the fuzziness parameter. Euclidean distance is employed to update the membership matrix and cluster centroids iteratively. The optimization process continues until the change in the membership matrix becomes smaller than 10^{-6} or the maximum number of iterations reaches 100. After convergence, each observation is assigned to the cluster for which it has the highest membership degree. Missing values are subsequently estimated using the corresponding centroid values of that cluster. By combining local pattern identification through sequential biclustering with the flexibility of fuzzy clustering, the proposed method is expected to provide more accurate missing value estimates while preserving the underlying structure of gene expression data.

Comparison Methods

To evaluate the effectiveness of the proposed method, two conventional imputation techniques were used as benchmarks. Mean imputation replaces missing values with the average value of each column. Although simple and computationally efficient, this method does not preserve data variability and often leads to biased estimates. K-nearest neighbors imputation

estimates missing values based on the similarity between observations. For each missing entry, the algorithm identifies the nearest neighbors and computes the imputed value based on their observed values. In this study, the number of neighbors was set to five.

Evaluation Metrics

The performance of each imputation method was evaluated using three commonly used error metrics: mean squared error, root mean squared error, and mean absolute error. These metrics are computed only on the artificially removed values to ensure a fair evaluation.

Mean Squared Error is a commonly used metric to evaluate the accuracy of imputation results. It measures the average squared difference between the actual and estimated values (Hastie et al., 2009). The mean squared error is defined as:

$$MSE = \frac{1}{|I||J|}\sum_{i\in I}\sum_{j\in J}(x_{i,j} - \hat{x}_{i,j})^2 \tag{4}$$

Root Mean Squared Error is widely used metric to measure the magnitude of prediction error. It 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 (Hastie et al., 2009). The root mean squared error is defined as:

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

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 (Khan, 2024):

$$MAE = \frac{1}{|I||J|}\sum_{i\in I}\sum_{j\in J}\left|x_{i,j} - \hat{x}_{i,j}\right| \tag{6}$$

Lower values of these metrics indicate better imputation performance. All experiments were conducted using Python, and the final results were obtained by averaging the performance across repetitions for each missing rate.

Result and Discussion

Results

The experimental results for the imputation methods are presented in Table 1, which summarizes average performance across different missing rates in terms of mean squared error, root mean squared error, and mean absolute error. The results are obtained from five replications for each missing rate, ensuring robustness and consistency of the evaluation.

Table 1. Performance comparison of imputation methods under varying missing rates. It presents the average values of mean squared error, root mean squared error, and mean absolute error for the proposed method, mean imputation, and k-nearest neighbors imputation across missing rates ranging from 5% to 55%.

Missing Rate	MSE (Proposed)	RMSE (Proposed)	MAE (Proposed)	MSE (Mean)	RMSE (Mean)	MAE (Mean)	MSE (KNN)	RMSE (KNN)	MAE (KNN)
5%	0.000050	0.006520	0.000154	0.000113	0.010391	0.000315	0.000072	0.008250	0.000118
10%	0.000058	0.007318	0.000179	0.000114	0.010472	0.000320	0.000072	0.008323	0.000124
15%	0.000049	0.006766	0.000172	0.000115	0.010557	0.000320	0.000073	0.008427	0.000126
20%	0.000062	0.007710	0.000189	0.000113	0.010540	0.000317	0.000080	0.008843	0.000131
25%	0.000067	0.008177	0.000202	0.000113	0.010549	0.000319	0.000084	0.009044	0.000135
30%	0.000076	0.008659	0.000217	0.000113	0.010563	0.000318	0.000090	0.009394	0.000140
35%	0.000083	0.009085	0.000229	0.000111	0.010517	0.000317	0.000091	0.009540	0.000142
40%	0.000091	0.009511	0.000244	0.000110	0.010474	0.000318	0.000092	0.009600	0.000144
45%	0.000086	0.009219	0.000246	0.000104	0.010177	0.000316	0.000089	0.009422	0.000140
50%	0.000097	0.009797	0.000260	0.000106	0.010289	0.000317	0.000095	0.009730	0.000148
55%	0.000094	0.009625	0.000264	0.000103	0.010129	0.000318	0.000097	0.009788	0.000149

Table 2. Standard deviation of imputation performance metrics across five replications. The proposed method consistently achieves lower MSE and RMSE values than both mean imputation and KNN imputation across all missing rates.

Missing Rate	MSE SD (Proposed)	RMSE SD (Proposed)	MAE SD (Proposed)	MSE SD (Mean)	RMSE SD (Mean)	MAE SD (Mean)	MSE SD (KNN)	RMSE SD (KNN)	MAE SD (KNN)
5%	0.000036	0.003136	0.000056	0.000057	0.002597	0.0000055	0.000040	0.002216	0.000049
10%	0.000034	0.002288	0.000054	0.000045	0.002290	0.0000041	0.000027	0.001777	0.000034
15%	0.000026	0.001919	0.000041	0.000041	0.002054	0.0000036	0.000026	0.001596	0.000032
20%	0.000027	0.001704	0.000033	0.000034	0.001638	0.0000026	0.000027	0.001556	0.000030
25%	0.000013	0.000782	0.000021	0.000032	0.001513	0.0000023	0.000028	0.001506	0.000032
30%	0.000017	0.000922	0.000020	0.000026	0.001271	0.0000016	0.000024	0.001323	0.000027
35%	0.000012	0.000656	0.000015	0.000017	0.000810	0.00000077	0.000013	0.000678	0.000014
40%	0.000020	0.001052	0.000020	0.000011	0.000502	0.00000038	0.000009	0.000481	0.000010
45%	0.000015	0.000809	0.000015	0.000014	0.000692	0.00000029	0.000012	0.000605	0.000013
50%	0.000026	0.001311	0.000021	0.000018	0.000880	0.000000087	0.000019	0.000938	0.000019
55%	0.000024	0.001210	0.000016	0.000019	0.000953	0.00000035	0.000022	0.001107	0.000022

Table 1 presents the average values of mean squared error, root mean squared error, and mean absolute error for the proposed method, mean imputation, and k-nearest neighbors imputation across missing rates ranging from 5% to 55%. Each result represents the average of five replications, while the corresponding standard deviations are reported in Table 2 to assess the variability of the results. The proposed method consistently achieves lower MSE and RMSE values than both mean imputation and KNN imputation across all missing rates. For example, at a 5% missing rate, the proposed method yields an MSE of 0.000050 compared with 0.000113 for mean imputation and 0.000072 for KNN imputation. Similar trends are observed across higher missing rates, indicating that the proposed method is more effective in reducing squared estimation errors and preserving the underlying structure of the data.

When compared to the k-nearest neighbors method, the proposed method approach maintains lower MSE and RMSE values throughout the evaluated missing-rate range, suggesting that the integration of sequential biclustering and Fuzzy C-Means clustering provides a robust mechanism for handling missing values. However, a different pattern is observed for MAE, where KNN imputation generally produces lower values than the proposed method across most missing rates. This finding indicates that KNN yields slightly smaller average absolute deviations, whereas the proposed method is more effective in reducing larger estimation errors that contribute more heavily to MSE and RMSE.

Furthermore, as the missing rate increases, all methods exhibit an increasing trend in error values. Nevertheless, the proposed method maintains consistently lower MSE and RMSE values than the benchmark methods, even at higher levels of missingness. The standard deviations reported in Table 2 indicate that the performance of all methods remains reasonably consistent across the five replications. These results suggest that the proposed method is robust in preserving local data structures and maintaining imputation accuracy under varying degrees of data incompleteness.

To provide a clearer visualization of the results, the performance trends are illustrated in Figures 1–3. These figures show the relationship between missing rate and each evaluation metric.

This journal article is © Azzahra et al. 2026



Figure 1. Relationship between missing rate and mean squared error for all imputation methods. The proposed method consistently achieves lower mean squared errors across all missing rates, indicating better accuracy in estimating missing values. The gap between methods becomes more evident as the missing rate increases.



Figure 2. Relationship between missing rate and root mean squared error for all imputation methods. the proposed method maintains lower root mean squared errors, confirming its robustness to increasing levels of missing data

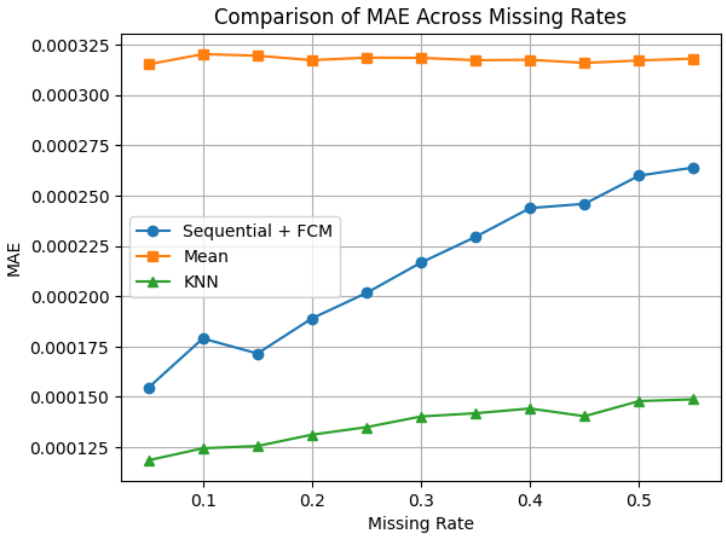

Figure 3. Relationship between missing rate and mean absolute error for all imputation methods. KNN imputation generally achieves lower Mean Absolute Error values than the proposed method across the evaluated missing rates. Nevertheless, the proposed method maintains competitive MAE performance while achieving substantially lower MSE and RMSE values, indicating better control of larger estimation errors.

Discussion

The superior performance of the proposed method in terms of MSE and RMSE can be attributed to the integration of sequential biclustering and Fuzzy C-Means clustering. Biclustering enables the identification of locally coherent patterns within subsets of genes, which is particularly important in gene expression datasets where global patterns are often weak or inconsistent. By selecting columns that satisfy the Mean Squared Residue criterion, the method focuses on relevant feature subsets and reduces the influence of noisy variables during the imputation process.

In addition, the use of Fuzzy C-Means clustering provides a flexible representation of data structure through membership degrees rather than hard cluster assignments. This allows observations to belong to multiple clusters simultaneously and enables the imputation process to capture overlapping biological patterns that commonly occur in gene expression data. Consequently, the proposed method is better able to estimate missing values while preserving local structural relationships among genes.

The comparison with mean imputation highlights the limitations of simple statistical approaches. Mean imputation fails to capture feature relationships and often leads to biased estimates, as reflected in the consistently higher error values. In contrast, k-nearest neighbors imputation exploits local similarity among observations and therefore generally provides better estimation accuracy. However, the relative performance between KNN and the proposed method varies depending on the evaluation metric, indicating that different methods may emphasize different aspects of imputation quality.

Compared with KNN imputation, the proposed method consistently achieves lower MSE and RMSE values across all evaluated missing rates. Since MSE and RMSE penalize large estimation errors more heavily, these results indicate that the proposed framework is more effective in reducing substantial deviations between the imputed and true values. However, the MAE results reveal a different pattern, where KNN imputation generally produces lower MAE values than the proposed method. Because MAE measures the average magnitude of prediction errors without emphasizing larger deviations, this finding suggests that KNN yields slightly smaller average absolute errors. Therefore, while KNN performs better according to MAE, the proposed method demonstrates greater effectiveness in controlling larger imputation errors, as reflected by its superior MSE and RMSE performance.

Another important finding is the stability of the proposed method across different missing rates. Although all methods experience performance degradation as the proportion of missing values increases, the proposed method maintains lower MSE and RMSE values throughout the evaluated range from 5% to 55%. Furthermore, the relatively small standard deviations reported in Table 2 indicate that the performance of all methods remains reasonably consistent across the five replications, supporting the reliability of the experimental results.

Overall, the results demonstrate that the proposed hybrid approach provides a robust solution for missing value imputation in gene expression data. While KNN imputation achieves lower MAE values, the proposed method consistently produces lower MSE and RMSE values, suggesting a stronger ability to reduce large estimation errors and preserve coherent local

structures within high-dimensional biological datasets. These findings highlight the importance of combining local pattern discovery and flexible clustering techniques to improve imputation performance and reduce large estimation errors in complex gene expression data.

Conclusion

This study demonstrates that integrating sequential biclustering with Fuzzy C-Means clustering provides an effective framework for handling missing values in high-dimensional gene expression data. By combining local pattern discovery through biclustering with the flexible representation of fuzzy clustering, the proposed approach is able to preserve important structural information within the data during the imputation process. Experimental results show that the proposed method consistently achieves lower MSE and RMSE values than mean imputation and KNN imputation across all evaluated missing rates, indicating a stronger ability to reduce large estimation errors. However, KNN imputation generally produces lower MAE values, suggesting slightly smaller average absolute deviations. These findings indicate that the proposed method offers a robust alternative for missing value imputation, particularly when minimizing large imputation errors is a primary objective. Overall, the study highlights the potential of hybrid biclustering and clustering-based approaches for improving missing value estimation in complex gene expression datasets. Future research may explore adaptive parameter optimization, alternative biclustering strategies, and statistical significance testing to further enhance the robustness and generalizability of the proposed framework.

Conflicts of Interest

There are no conflicts to declare.

Acknowledgements

None.

AI Writing Statement

During the preparation of this work, the author(s) used ChatGPT to assist with language editing, including grammar and clarity improvement. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

References

Aittokallio, T. (2009). Dealing with missing values in large-scale studies: Microarray data imputation and beyond. *Briefings in Bioinformatics*, *11*(2): 253–264. https://doi.org/10.1093/bib/bbp059

Bezdek, J. C. (1981). *Pattern recognition with fuzzy objective function algorithms*. Springer. https://doi.org/10.1007/978-1-4757-0450-1

Bezdek, J. C., Ehrlich, R., & Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, *10*(2–3): 191–203. https://doi.org/10.1016/0098-3004(84)90020-7

Castanho, E. N., Aidos, H., & Madeira, S. C. (2024). Biclustering data analysis: A comprehensive survey. *Briefings in Bioinformatics*, *25*(4): bbae342. https://doi.org/10.1093/bib/bbae342

Cheng, Y., & Church, G. M. (2000). Biclustering of expression data. *Proceedings of ISMB*, *8*: 93–103. https://doi.org/10.1093/bioinformatics/16.suppl_1.s93

Clifton, C., Hanson, E. J., Merrill, K., & Merrill, S. (2022). Differentially private k-nearest neighbor missing data imputation. *ACM Transactions on Privacy and Security*, *25*(3). https://doi.org/10.1145/3507952

Dubey, A., & Rasool, A. (2021). Efficient technique of microarray missing data imputation using clustering and weighted nearest neighbour. *Scientific Reports*, *11*: 24297. https://doi.org/10.1038/s41598-021-03438-x

Fadlil, A., Herman, & Praseptian M, D. (2022). K nearest neighbor imputation performance on missing value data graduate user satisfaction. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, *6*(4): 570–576.

https://doi.org/10.29207/resti.v6i4.4173

García-Laencina, P. J., Sancho-Gómez, J. L., & Figueiras-Vidal, A. R. (2010). Pattern classification with missing data: A review. *Neural Computing and Applications*, *19*: 263–282. https://doi.org/10.1007/s00521-009-0295-6

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer. https://doi.org/10.1007/978-0-387-84858-7

Hou, W., Ji, Z., Ji, H., & Hicks, S. C. (2020). A systematic evaluation of single-cell RNA-sequencing imputation methods. *Genome Biology*, *21*: 218. https://doi.org/10.1186/s13059-020-02132-x

Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., & Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial Intelligence in Medicine*, *50*(2): 105–115. https://doi.org/10.1016/j.artmed.2010.05.002

Joel, L. O., Doorsamy, W., & Paul, B. S. (2025). A comparative study of imputation techniques for missing values in healthcare diagnostic datasets. *International Journal of Data Science and Analytics*, *20*: 6357–6373. https://doi.org/10.1007/s41060-025-00825-9

Khan, M. A. (2024). A comparative study on imputation techniques: Introducing a transformer model for robust and efficient handling of missing EEG amplitude data. *Bioengineering*, *11*(8): 740. https://doi.org/10.3390/bioengineering11080740

Laia, M. (2023). Performance analysis of the k-nearest neighbor imputation (KNNI) algorithm for missing values in data mining classification. *Journal of Informatics, Electrical and Electronics Engineering*, *2*(3): 92–98. https://doi.org/10.47065/jieee.v2i3.891

Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.). Wiley. https://doi.org/10.1002/9781119482260

Madeira, S. C., & Oliveira, A. L. (2004). Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *1*(1): 24–45. https://doi.org/10.1109/TCBB.2004.2

Mastika, Siswantining, T., & Bustamam, A. (2025). Comparison of missing value imputation using mean, Bayesian KNN, and non-Bayesian KNN on TEP gene expression data. *Media Statistika*, *18*(1): 61–72. https://doi.org/10.14710/medstat.18.1.61-72

Nicholls, K., & Wallace, C. (2021). Comparison of sparse biclustering algorithms for gene expression datasets. *Briefings in Bioinformatics*, *22*(6): bbab140. https://doi.org/10.1093/bib/bbab140

Pamuji, F. Y., Muslikh, A. R., Arief, R. M., & Muti, D. (2024). Comparison of mean and KNN imputation methods for handling missing values in small datasets. *JIP (Jurnal Informatika Polinema)*, *10*(2): 257–264.

Qi, R., Ma, A., Ma, Q., & Zou, Q. (2020). Clustering and classification methods for single-cell RNA-sequencing data. *Briefings in Bioinformatics*, *21*(4): 1196–1208. https://doi.org/10.1093/bib/bbz062

Sanchez, T., et al. (2023). A deep learning framework for population genetics inference. *Bioinformatics*, *39*(1): btac765. https://doi.org/10.1093/bioinformatics/btac765

Silalahi, A. U., Siswantining, T., & Pramana, S. (2025). Evaluation of biclustering imputation methods for glioblastoma gene expression data. *International Journal of Applied Statistics and Data Science*, *5*(1): 67–77.

Siswantining, T., Aminanto, A. E., Sarwinda, D., & Swasti, O. (2021). Biclustering analysis using plaid model on gene expression data of colon cancer. *Austrian Journal of Statistics*, *50*: 101–114. https://doi.org/10.17713/ajs.v50i5.1195

Sudrajat, W., & Cholid, I. (2023). K-nearest neighbor (K-NN) for handling missing values in MSME data. *Jurnal Rekayasa Sistem Informasi dan Teknologi*, *1*(2): 54–63. https://doi.org/10.59407/jrsit.v1i2.77

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*(6): 520–525. https://doi.org/10.1093/bioinformatics/17.6.520

Yosboon, S., Iam-On, N., Boongoen, T., Keerin, P., & Kirimasthong, K. (2024). Optimised multiple data partitions for cluster-wise imputation of missing values in gene expression data. *Expert Systems with Applications*, *257*: 125040.

https://doi.org/10.1016/j.eswa.2024.125040

Zamri, N. A., Jaya, M. I., Irawati, I. D., Rassem, T. H., Rasyidah, & Kasim, S. (2024). Comparative analysis of imputation methods for enhancing predictive accuracy in data models. *International Journal on Informatics Visualization*, *8*(3): 1271–1276. https://dx.doi.org/10.62527/joiv.8.3.1666

Zhang, S. (2012). Nearest neighbor selection for iteratively KNN imputation. *Journal of Systems and Software*, *85*(11): 2541–2552. https://doi.org/10.1016/j.jss.2012.05.073

Zhu, X., Wu, X., & Chen, Q. (2021). An efficient ensemble method for missing value imputation in microarray gene expression data. *BMC Bioinformatics*, *22*: 188. https://doi.org/10.1186/s12859-021-04109-4

This journal article is © Azzahra et al. 2026