ENTHUSIASTIC
INTERNATIONAL JOURNAL OF APPLIED STATISTICS AND DATA SCIENCE
Volume 5, Issue 1, April 2025, pp. 67-77

Evaluation of Biclustering Imputation Methods for Glioblastoma Gene Expression Data

Agatha Ulina Silalahi [a,1], Titin Siswantining [a,2,*], Setia Pramana [b,3]

[a] Universitas Indonesia, Jl. Lingkar, Depok, 16424, Indonesia
[b] Politeknik Statistika STIS, Jl. Otto Iskandardinata No.64C 1, Jakarta 13330, Indonesia
[1] agatha.ulina@ui.ac.id; [2] titin@sci.ui.ac.id*; [3] setia.pramana@stis.ac.id
* Corresponding author

ARTICLE INFO

Keywords
Glioblastoma
Gene Expression
Missing Value Imputation
Soft Impute

ABSTRACT

Glioblastoma is a highly aggressive primary brain tumor with a low survival rate. One of the main challenges in analyzing glioblastoma gene expression data is the presence of missing values, which can reduce biclustering accuracy and affect biological interpretation. This research compared six imputation methods *k*-nearest neighbors (KNN), mean imputation, singular value decomposition, nonnegative matrix factorization, soft impute, and autoencoderon the GSE4290 gene expression dataset with missing values ranging from 5% to 50%. An evaluation using root mean square error (RMSE), mean absolute error (MAE), and structural similarity index measure (SSIM) showed that soft impute provided the best performance at all levels of missing values, with RMSE of 0.0076, MAE of 0.0073, and perfect SSIM of 1.0000 at 50% missing values. Meanwhile, deep learning-based autoencoder experienced significant performance degradation at high missing values. These findings indicate that more complex models are not always superior, and regularization-based approaches like soft impute are more effective in preserving the biological structure of the data. The results of this research contribute to the optimization of imputation strategies to improve the accuracy of biclustering analysis in glioblastoma studies.

1. Introduction

Glioblastoma is the most aggressive primary brain tumor with a poor prognosis, with a median survival rate of only 12–15 months after diagnosis [1]. Biclustering-based gene expression analysis has the potential to reveal molecular mechanisms and subtypes of this disease, which can support the development of more effective therapies. However, missing values in gene expression data present a significant challenge that can reduce analysis accuracy and hinder biological interpretation [2].

Various imputation methods have been developed to handle missing values, ranging from conventional approaches like singular value decomposition (SVD), which has shown promising results in several studies [3], [4] to deep learning-based methods, such as autoencoder, which can capture complex patterns in biological data [5], [6]. Although deep learning methods have been widely used in various types of biological data [7], [8], their effectiveness in the context of glioblastoma gene expression with varying levels of sparsity has not been comprehensively evaluated.

p-ISSN 2798-253X
e-ISSN 2798-3153

Regularization-based methods such as soft impute have been used in various matrix analysis applications and have shown resilience to missing values [4]. However, studies on its performance in glioblastoma gene expression data remain limited. Previous comparative studies [9], [10] have not included systematic evaluations of this method, so it is not yet known to what extent regularization approaches can maintain the biological structure of data in the context of biclustering.

Glioblastoma multiforme (GBM) is the most aggressive primary brain tumor, accounting for approximately 45% of all malignant gliomas. The molecular characteristics of GBM have been identified through comprehensive analysis of gene expression profiles, revealing at least four main molecular subtypes: proneural, neural, classical, and mesenchymal [11]. Each subtype has different genetic characteristics and therapeutic responses, emphasizing the importance of accurate gene expression analysis for patient stratification and the development of targeted therapies [12].

Biclustering, also known as co-clustering, is a data analysis technique that performs simultaneous clustering on rows and columns in a data matrix, which is particularly valuable in gene expression data analysis [13]. A systematic comparative evaluation of various biclustering techniques has shown their effectiveness in analyzing complex gene expression patterns [14], with recent advances incorporating proximity-graph approaches to improve analysis accuracy [15]. This technique has been introduced to overcome the limitations of traditional clustering methods that only group entire rows or columns. In a gene expression data matrix X with dimensions $n \times m$, where n represents the number of genes and m represents the number of samples or experimental conditions, a bicluster is defined as a pair $B = (I, J)$, consisting of a subset of rows $I \subseteq \{1, 2, \ldots, n\}$ and a subset of columns $J \subseteq \{1, 2, \ldots, m\}$. This bicluster forms a submatrix whose elements show coherent expression patterns [13]. The general formula for biclustering can be expressed as an optimization problem, as in (1).

$$B = \arg\min_{I,J} \sum_{i \in I} \sum_{j \in J} d(x_{ij}, \mu_{IJ}) \quad (1)$$

where d is a distance function, x_{ij} is the expression value of gene i in sample j, and μ_{IJ} is the representative value of the bicluster. Missing values in gene expression data can be categorized into three types based on their occurrence mechanism: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Understanding these mechanisms is important in selecting an appropriate imputation method [2].

MCAR occurs when the probability of missing data does not depend on other variables in the dataset, either observed or not. MAR occurs when missingness depends on other observed variables but not on the missing value itself. Integrative omics analysis provides systematic approaches for handling such complex missing data patterns [16]. MNAR occurs when missingness depends on the missing value itself, such as when genes with high expression tend to be undetected due to measurement technique limitations [17]. Sensitivity analyses can be conducted to evaluate the impact of this type of missingness on the results [18].

This research evaluates six imputation methods *k*-nearest neighbors (KNN), mean imputation, singular value decomposition, non-negative matrix factorization, soft impute, and autoencoder using the GSE4290 glioblastoma gene expression dataset with missing values levels of 5%–50%. By comparing the performance of deep learning-based methods and regularization-based approaches in maintaining data structure, this research aimed to provide deeper insights into the selection of optimal imputation strategies for biclustering analysis in glioblastoma studies.

2. Method

2.1. Research Methodology

The research methodology used in this study follows systematic steps as illustrated in Figure 1. In general, this research process consists of four main stages: data collection, preprocessing, implementation of imputation methods, and performance evaluation and results analysis.

Fig. 1 Research methodology flow chart for comparative analysis of data imputation methods.

2.2. Data Collection and Preprocessing

The dataset used in this research was GSE4290, acquired from Gene Expression Omnibus (GEO), a public repository providing high-quality gene expression data, on December 15, 2024. The GSE4290 dataset consisted of 180 samples, including 136 glioblastoma samples and 44 normal control samples, and was analyzed using the Affymetrix Human Genome U133 Plus 2.0 Array platform. Data integration and predictive modeling methods facilitate comprehensive analysis of such multi-omics datasets [12]. Glioblastoma, classified as a grade IV diffuse astrocytic tumor according to the 2021 WHO Classification [11], is a malignant brain tumor with aggressive

characteristics and a low survival rate, making a comprehensive understanding of its gene expression profile critical for the development of more effective therapies [1], [12].

The data preprocessing stage involved multiple sequential steps, commencing with initial cleaning and raw data preparation. Logarithmic transformation with base 2 was then applied to normalize data distribution using $x' = log_2(x + 1)$. Following this, quantile normalization was conducted to standardize expression value distribution between samples. Variance-based filtering was then performed to identify genes with significant contribution to data variability. Subsequently, the MCAR mechanism was applied to systematically introduce missing values, with the percentage ranging from 5% to 50%. This approach assumes that the probability of missing data does not depend on the missing values or observed values [17], with computational tools available for efficient processing [18].

2.3. Data Collection and Preprocessing

2.3.1. K-Nearest Neighbors (KNN)

KNN-imputation in biclustering estimates missing values by utilizing the local structure in gene expression data. This method calculates missing values based on k values using the Euclidean distance:

$$d(x_i, x_j) = \sqrt{\sum_{l=1}^{n}(x_{il} - y_{jl})^2} \quad (2)$$

Missing values were then estimated using a weighted average of existing values in KNN.

$$\hat{x}_{ij} = \frac{\sum_{l=1}^{k} w_l x_{lj}}{\sum_{l=1}^{k} w_l} \quad (3)$$

where w_l is the weight inversely proportional to the distance to neighbor l. For biclustering, KNN-imputation adapts by calculating similarity only in relevant gene and sample subsets, thus increasing imputation accuracy in certain biclusters [2]. Advanced implementations have incorporated temporal approaches [19] for improved performance in genomic datasets. Sensitivity analyses for missing data assumptions [20] provide additional validation for imputation quality.

2.3.2. Singular Value Decomposition (SVD)

SVD is a matrix decomposition method that breaks down the data matrix X into three main components.

$$X = U\Sigma V^T \quad (4)$$

with U and V denote orthogonal matrices, and Σ denotes a diagonal matrix containing singular values. For biclustering imputation, iterative SVD began by initializing missing values using initial estimates (usually row or column averages). Then, SVD decomposition was performed on the filled matrix. Next, the matrix was reconstructed using the first k components:

$$\hat{X} = U_k \Sigma_k V_k^T \quad (5)$$

Missing values were updated using the reconstruction results, and this process was repeated until convergence, defined as when the change in the Frobenius norm of successive approximations fell below a threshold ε:

$$\left\| X^{(t+1)} - X^t \right\|_F < \varepsilon \quad (6)$$

The SVD approach assumes that the data structure can be explained by a small number of orthogonal components, which is consistent with the nature of biclusters in gene expression data. This approach has proven effective in identifying relevant signal pathways in genomic data, including glioblastoma data [3].

2.3.3. Nonnegative Matrix Factorization (NMF)

NMF is particularly suitable for biclustering imputation of glioblastoma gene expression data due to its ability to maintain nonnegativity, which is an intrinsic property of gene expression data.

Recent advances have extended this approach to incorporate gene-gene interaction networks for improved dimensionality reduction and imputation [20]. This method decomposes the data matrix into two nonnegative matrices:

$$X \approx WH \tag{7}$$

where W and H are nonnegative matrices. Optimization was performed by minimizing the objective function with nonnegativity constraints:

$$min\ ||X - WH||^2\ subject\ to\ W, H \geq 0 \tag{8}$$

The implementation of NMF for biclustering analysis enables the identification of co-regulated gene modules related to tumor subtypes, which is very relevant in glioblastoma studies [21]. Multiple imputation frameworks provide flexible approaches for handling missing data in biological contexts [22].

2.3.4. Soft Impute

Soft impute uses a nuclear norm regularization approach for matrix reconstruction with missing values. This algorithm minimizes:

$$min\ ||P_\Omega(X - Z)||^2 + \lambda||Z|_* \tag{9}$$

with the solution given through the soft-thresholding operator:

$$\hat{Z} = S_\lambda(P_\Omega(X) + P_{\Omega^c}(\hat{Z})) \tag{10}$$

where S_λ is the soft-thresholding operator applied to singular values, P_Ω is the projection operator on observed elements, and P_{Ω^c} is the projection on missing elements. The parameter λ controls the strength of regularization, with higher values resulting in lower-rank solutions. The algorithm iteratively applies soft thresholding to singular values until convergence, defined as in (11).

$$\frac{\|Z^{(t+1)} - Z^{(t)}\|_F}{\|Z^{(t)}\|_F} < \delta \tag{11}$$

where δ is a small threshold value, typically set to 10^{-6}. Nuclear norm-based imputation is particularly effective for biclustering due to its ability to capture local low-rank structure in the gene expression matrix, which is an important characteristic in glioblastoma data [4]. Alternative sequential regression approaches [23] have also been explored for multivariate missing data scenarios.

2.3.5. Autoencoder-Based Imputation

Autoencoder uses a neural network architecture for compact data representation learning. With input x, encoding h, and decoding $\hat{x}$:

$$\begin{aligned} h &= \sigma(Wx + b) \\ \hat{x} &= \sigma(W'h + b') \end{aligned} \tag{12}$$

where h denotes hidden representation, x denotes input data, W, W' denotes weight matrices, b, b denotes biases, and σ denotes nonlinear activation function.

This study employed a rectified linear unit (ReLU) activation function for hidden layers and a sigmoid function for the output layer to maintain the range of gene expression values. The loss function was the mean squared error between the original and reconstructed values, with an additional L_2 regularization term to prevent overfitting.

$$L = \|x - \hat{x}\|^2 + \alpha\|W\|^2 + \alpha\|W'\|^2 \tag{13}$$

where α is the regularization parameter. The network was trained using the Adam optimizer with early stopping based on validation loss. Autoencoder approaches have shown promise in health data [7] and single-cell RNA sequencing applications [8] and can be applied to gene expression data for identifying clinically relevant subtypes of glioblastoma [5].

2.4. Evaluation Metrics

Evaluation frameworks using RMSE, MAE, and SSIM were implemented as the primary performance metrics. This comprehensive approach to evaluation drew inspiration from systematic benchmarking studies for imputation methods in biological data analysis [11], which was specifically adapted to address the unique challenges presented by glioblastoma gene expression datasets. Each metric provides complementary information about the quality of imputation, from error magnitude to structural preservation of the data.

2.4.1. Root Mean Square Error (RMSE)

RMSE measures the square root of the average squared difference between predicted values and actual values.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (14)$$

where y_i is the actual value, $\hat{y}_i$ is the imputed value, and n is the number of samples. RMSE gives greater weight to large errors, making it very sensitive to outliers [24].

2.4.2. Mean Absolute Error (MAE)

MAE measures the absolute average of the difference between predicted values and actual values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (15)$$

MAE gives equal weight to all errors, making it more robust against outliers compared to RMSE [24].

2.4.3. Structural Similarity Index (SSIM)

SSIM assesses the structural similarity between two matrices based on three components: luminance (l), contrast (c), and structure (s).

$$SSIM(x,y) = [l(x,y)]^{\alpha} \cdot [c(x,y)]^{\beta} \cdot [s(x,y)]^{\gamma} \quad (16)$$

The luminance component measures the difference in average gene expression between the original and imputed data. The contrast evaluates the variation in expression between genes in one condition, and the structure measures the spatial correlation between gene expressions in the bicluster. SSIM provides a more accurate assessment of data reconstruction quality in biological analysis applications, including gene expression analysis in cancers such as glioblastoma [25].

3. Method

3.1. Performance of Imputation Methods

Table 1 presents the quantitative evaluation results of six imputation methods based on RMSE, MAE, and SSIM metrics at various levels of missing values.

Table 1. Comparison of Imputation Methods Performance at Various Levels of Missing Values

Method	Metric	5%	10%	15%	20%	30%	50%
KNN	RMSE	1.4756	2.0879	2.5573	2.9550	3.6202	4.6723
	MAE	0.3139	0.6284	0.9421	1.2578	1.8887	3.1517
	SSIM	0.7757	0.6290	0.5250	0.4462	0.3341	0.1923
MEAN	RMSE	1.4004	1.9814	2.4248	2.8007	3.4313	4.4306
	MAE	0.3130	0.6265	0.9383	1.2519	1.8791	3.1336
	SSIM	0.7723	0.6219	0.5155	0.4353	0.3219	0.1868
SVD	RMSE	1.6907	2.1338	2.4582	2.7071	3.0452	3.2688
	MAE	1.1041	1.4498	1.7781	2.0770	2.5590	2.9487
	SSIM	0.6191	0.5016	0.4246	0.3648	0.2868	0.2149
NMF	RMSE	1.6935	2.1351	2.4575	2.7065	3.0443	3.2688
	MAE	1.1064	1.4502	1.7752	2.0756	2.5600	2.9559

Method	Metric	5%	10%	15%	20%	30%	50%
SOFT	SSIM	0.6141	0.4970	0.4192	0.3611	0.2835	0.1985
	RMSE	0.0065	0.0069	0.0069	0.0068	0.0064	0.0054
	MAE	0.0014	0.0021	0.0026	0.0029	0.0034	0.0036
AUTOENCODER	SSIM	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	RMSE	1.6508	2.0809	2.3807	2.6622	2.9712	3.1322
	MAE	1.0748	1.4426	1.7395	2.0390	2.4754	2.7396
	SSIM	0.6443	0.5394	0.4835	0.3977	0.3502	0.3185

Soft impute demonstrated the best performance with the lowest RMSE values, ranging between 0.0054 (at 50% missing values) and 0.0069 (at 15% missing values). MAE values were also observed to be consistent and low (ranging between 0.0014 and 0.0036), while SSIM remained at 1.0000 across all levels of missing values. This indicates that soft impute consistently maintains the data structure and provides high accuracy in predicting missing values.

In contrast, simple methods such as KNN and mean imputation experienced significant performance degradation. At 5% missing values, RMSE values for KNN and mean imputation were 1.4756 and 1.4004, respectively, with SSIM values of 0.7757 and 0.7723. However, when the level of missing values increased to 50%, RMSE values increased substantially to 4.6723 and 4.4306, and SSIM values decreased dramatically to 0.1923 and 0.1868, indicating significant damage to the data structure.

SVD and NMF methods exhibited better resilience with more controlled increases in RMSE and decreases in SSIM. Autoencoder, despite being based on deep learning, demonstrated less competitive performance with RMSE values ranging between 1.6508 and 3.1322 and MAE values ranging between 1.0748 and 2.7396, showing that model complexity does not guarantee high accuracy in the context of this data.

Fig. 2 shows the relationship between RMSE and the level of missing values for each imputation method. The pattern in the graph shows that soft impute has the most stable trend, with RMSE staying low and close to zero, even at the highest level of missing values (50%). This indicates that this method is able to reconstruct data with high accuracy without being affected by increased data sparsity.

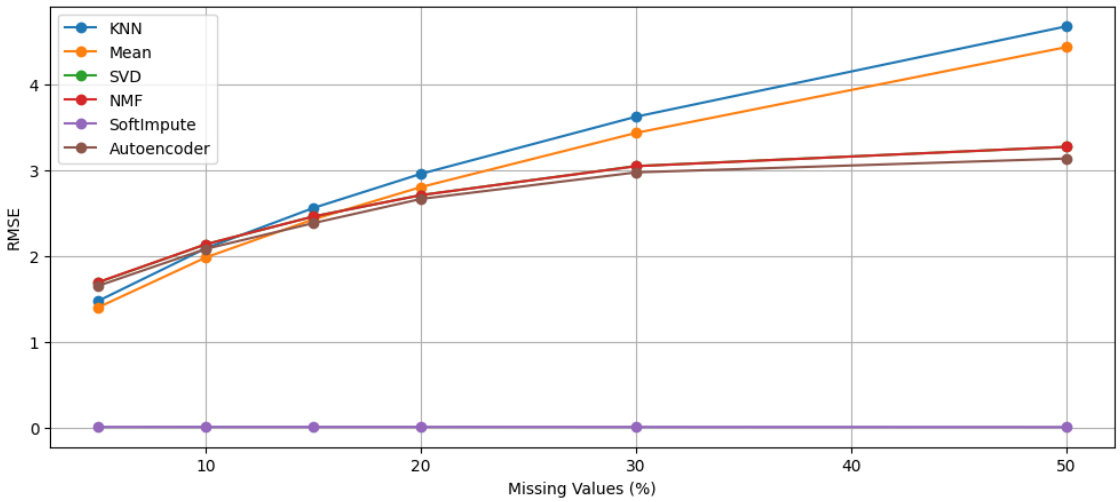

Fig. 2 Comparison of RMSE from various imputation methods at different levels of missing values.

In contrast, KNN and mean imputation showed exponential increases in RMSE as the level of missing values increased, reaching values of 4.6723 and 4.4306 at 50%. This shows significant accuracy degradation and the inability of these methods to handle data with high sparsity. Meanwhile, SVD, NMF, and autoencoder exhibited more gradual increases in RMSE (ranging

p-ISSN 2798-253X
e-ISSN 2798-3153

between 3.0 and 3.3 at 50%), indicating relative resilience to increases in missing values, although still lower compared to soft impute.

Fig. 3 shows similar patterns for MAE values, with soft impute maintaining consistently low error rates across all missing value percentages. The conventional methods (KNN and mean imputation) showed steeper increases in MAE at higher missing value rates, while decomposition-based and deep learning methods exhibited more moderate degradation.

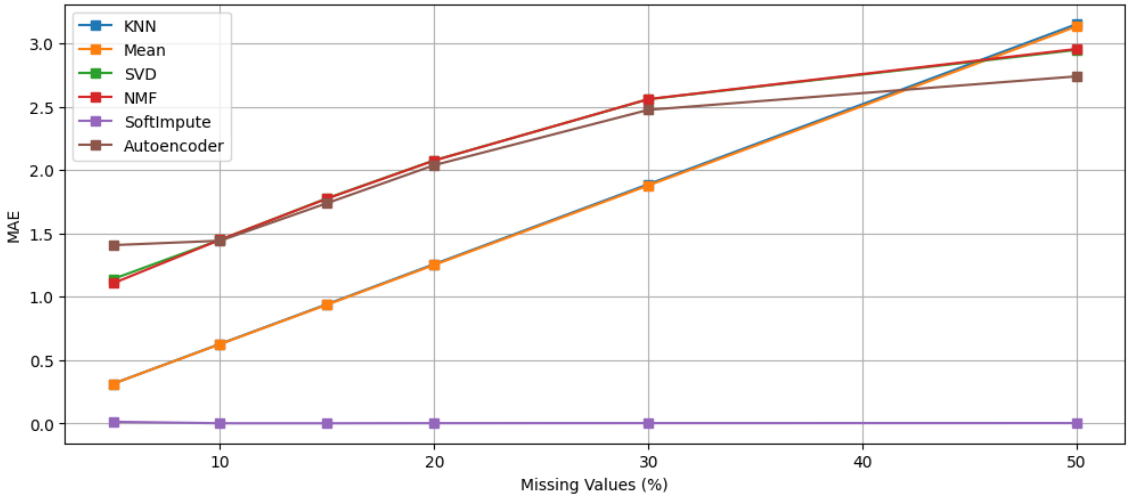

Fig. 3 Comparison of MAE from Various imputation methods at different levels of missing values.

Fig. 4 displays a visualization of SSIM between original data and imputed data for each method at various levels of missing values. Soft impute showed perfect SSIM values (1.0000) in all scenarios, affirming its ability to maintain structure and co-expression patterns between genes. These results indicate that soft impute not only provides imputation with minimal errors but also maintains the integrity of structural relationships in the data, which is important for further analysis, such as biclustering.

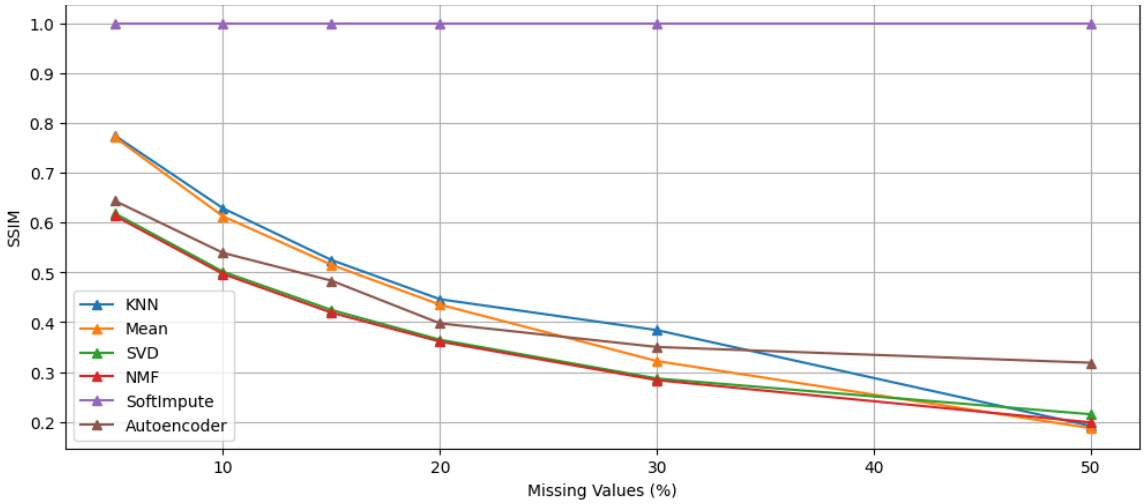

Fig. 4 Comparison of SSIM from various imputation methods at different levels of missing values.

In contrast, other methods demonstrated varying decreases in SSIM. SVD and NMF exhibited moderate resilience, with SSIM remaining above 0.2 at the highest level of missing values (50%). Conversely, KNN, mean imputation, and autoencoder experience significant decrease in SSIM, indicating substantial damage to the data structure.

3.2. Theoretical Implications of Research Findings

The results of this research demonstrate that soft impute consistently outperforms autoencoder and other methods at all levels of missing values, challenging the assumption that deep learning is always superior for complex biological data. These findings align with recent studies [26], suggesting that model complexity can cause overfitting and reduce generalization in cancer data analysis.

The superiority of soft impute can be attributed to the low-rank structure that is typical in glioblastoma gene expression data [4], [6], where the nuclear norm regularization used allows for more accurate data reconstruction by maintaining co-expression relationships between genes in glioblastoma molecular profiles. In contrast, autoencoder tends to struggle learning meaningful representations from glioblastoma gene expression patterns when missing values are high, as indicated by increased RMSE and decreased SSIM [5].

Besides minimizing numerical errors, the preservation of the biological structure of data becomes a crucial factor in glioblastoma biclustering analysis. The perfect SSIM (1.0000) achieved by soft impute demonstrates its ability to maintain inter-gene relationships in glioblastoma regulatory networks, aligning with principles proposed by [24] and [14] that maintaining biological data structure is more important than mere imputation accuracy, especially in identifying molecular subtypes of brain cancer.

These findings highlight that model selection for imputation should be guided primarily by data characteristics rather than model complexity alone. While this study focused on classical imputation methods and neural network approaches, future research can benefit from exploring ensemble methods like random forests [27] and gradient boosting frameworks such as XGBoost [28], which might provide complementary strengths in capturing both linear and nonlinear relationships in the heterogeneous expression patterns characteristic of glioblastoma subtypes. Advanced computational approaches, including graph neural networks [29], may provide additional insights for single-cell genomic data analysis.

4. Conclusion

This research evaluated six imputation methods for glioblastoma gene expression data. Soft impute demonstrated the best performance, with the lowest RMSE values (ranging between 0.0054 and 0.0069) and MAE values (ranging between 0.0014 and 0.0036), as well as perfect SSIM values (1.0000) at all levels of missing values (ranging from 5% to 50%). This demonstrates soft impute's ability to maintain co-expression relationships between genes and biological structural patterns in glioblastoma gene expression data, which is very important for accurate biclustering processes in identifying functional gene modules.

Conventional methods (KNN and mean imputation) experienced significant performance degradation at high levels of missing values, shown by substantial increases in RMSE values (from approximately 1.4 to more than 4.4) and MAE values (from approximately 0.3 to more than 3.1), as well as sharp decreases in SSIM values (from approximately 0.77 to less than 0.2) at 50% missing values. Meanwhile, decomposition-based methods (SVD and NMF) showed better resilience with RMSE values around 3.2 and SSIM values above 0.2 at 50% missing values, but they were still not as effective as soft impute. Autoencoder, despite being a complex deep learning-based method, did not provide significant advantages with higher RMSE values (between 1.65 and 3.13) and MAE values (between 1.07 and 2.74) compared to soft impute.

The superiority of soft impute comes from its ability to utilize low-rank structure in gene expression data, important for biclustering analysis. These findings confirm soft impute as the optimal method for glioblastoma gene expression data imputation, while also opening opportunities for the development of integrative methods based on regularization and deep learning for future multi-omics analysis.

p-ISSN 2798-253X
e-ISSN 2798-3153

References

[1] A.C. Tan, D.M. Ashley, G.Y. López, and M. Malinzak, "Management of glioblastoma: State of the art and future directions," *CA Cancer J. Clin.*, vol. 70, no. 4, pp. 299–312, Jul. 2020, doi: 10.3322/caac.21613.

[2] O. Troyanskaya et al., "Missing value estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, Jun. 2001, doi: 10.1093/bioinformatics/17.6.520.

[3] O. Alter, P.O. Brown, and D. Botstein, "Singular value decomposition for genome-wide expression data processing and modeling," *Proc. Natl. Acad. Sci.*, vol. 97, no. 18, pp. 10101–10106, Aug. 2000, doi: 10.1073/pnas.97.18.10101.

[4] R. Mazumder, T. Hastie, and R. Tibshirani, "Spectral regularization algorithms for learning large incomplete matrices," *J. Mach. Learn. Res.*, vol. 11, pp. 2287–2322, Mar. 2010.

[5] G.P. Way and C.S. Greene, "Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders," in *Pac. Symp. Biocomput.*, 2018, vol. 23, pp. 80–91.

[6] D. Risso, F. Perraudeau, S. Gribkova, S. Dudoit, and J.P. Vert, "A general and flexible method for signal extraction from single-cell RNA-seq data," *Nat. Commun.*, vol. 9, Jan. 2018, Art. no 284 (2018), doi: 10.1038/s41467-017-02554-5.

[7] B.K. Beaulieu-Jones and J.H. Moore, "Missing data imputation in the electronic health record using deeply learned autoencoders," in *Pac. Symp. Biocomput*, 2017, vol. 22, pp. 207–218.

[8] G. Eraslan, L.M. Simon, M. Mircea, N.S. Mueller, and F J. Theis, "Single-cell RNA-seq denoising using a deep count autoencoder," *Nat. Commun.*, vol. 10, Jan. 2019, Art. no 390 (2019), doi: 10.1038/s41467-018-07931-2.

[9] C. Lazar, L. Gatto, M. Ferro, C. Bruley, and T. Burger, "Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies," *J. Proteome Res.*, vol. 15, no. 4, pp. 1116–1125, Apr. 2016, doi: 10.1021/acs.jproteome.5b00981.

[10] G.N. Brock, J.R. Shaffer, R.E. Blakesley, M.J. Lotz, and G.C. Tseng, "Which missing value imputation method to use in expression profiles: a comparative study and two selection schemes," *BMC Bioinform.*, vol. 9, Jan. 2008, Art. no. 12, doi: 10.1186/1471-2105-9-12.

[11] D. N. Louis et al., "The 2021 WHO classification of tumors of the central nervous system: A summary," *Neuro Oncol.*, vol. 23, no. 8, pp. 1231–1251, Jun. 2021, doi: 10.1093/neuonc/noab106.

[12] M. Kim and I. Tagkopoulos, "Data integration and predictive modeling methods for multi-omics datasets," *Mol. Omics*, vol. 14, no. 1, pp. 8–25, Feb. 2018, doi: 10.1039/C7MO00051K.

[13] R.G. Verhaak et.al., "Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1," *Cancer Cell*, vol. 17, no. 1, pp. 98–110, Jan. 2010, doi: 10.1016/j.ccr.2009.12.020.

[14] V.A. Padilha and R.J.G.B. Campello, "A systematic comparative evaluation of biclustering techniques," *BMC Bioinform.*, vol. 18, Jan. 2017, Art. no 55 (2017), doi: 10.1186/s12859-017-1487-1.

[15] Y. Cheng and G.M. Church, "Biclustering of expression data," in *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 2000, pp. 93–103.

[16] M. Kim and I. Tagkopoulos, "Data integration and predictive modeling methods for multi-omics datasets," *Mol. Omics*, vol. 14, no. 1, pp. 8–25, Feb. 2018, doi: 10.1039/C7MO00051K.

[17] C. Lazar et al., "Batch effect removal methods for microarray gene expression data integration: A survey," *Brief. Bioinform.*, vol. 14, no. 4, pp. 469–490, Jul. 2013, doi: 10.1093/bib/bbs037.

[18] P. Orzechowski, A. Pańszczyk, X. Huang, and J.H. Moore, "Runibic: A bioconductor package for parallel row-based biclustering of gene expression data," *Bioinformatics*, vol. 34, no. 24, pp. 4302–4304, Dec. 2018, doi: 10.1093/bioinformatics/bty512.

[19] G.P. Way and C.S. Greene, "Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders," in *Pac. Symp. Biocomput.*, 2018, pp. 80–91.

[20] R. Elyanow, B. Dumitrascu, B.E. Engelhardt, and B. J. Raphael, "netNMF-sc: Leveraging gene-gene interactions for imputation and dimensionality reduction in single-cell expression analysis," *Genome Res.*, vol. 30, no. 2, pp. 195–204, Feb. 2020, doi: 10.1101/gr.251603.119.

[21] R. Gaujoux and C. Seoighe, "A flexible R package for nonnegative matrix factorization," *BMC Bioinform.*, vol. 11, Jul. 2010, Art. no 367 (2010), doi: 10.1186/1471-2105-11-367.

[22] P. Li, E.A. Stuart, and D.B. Allison, "Multiple imputation: A flexible tool for handling missing data," *JAMA*, vol. 314, no. 18, pp. 1966–1967, Nov. 2015, doi: 10.1001/jama.2015.15281.

[23] Nurzaman, T. Siswantining, S.M. Soemartojo, and D. Sarwinda, "Application of sequential regression multivariate imputation method on multivariate normal missing data," in *2019 3rd Int. Conf. Inform. Comput. Sci. (ICICoS)*, 2019, pp. 1–6, doi: 10.1109/ICICoS48119.2019.8982423.

[24] A. Subramanian et al., "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proc. Natl. Acad. Sci.*, vol. 102, no. 43, pp. 15545–15550, Oct. 2005, doi: 10.1073/pnas.0506580102.

[25] G. Eraslan, L.M. Simon, M. Mircea, N.S. Mueller, and F.J. Theis, "Single-cell RNA-seq denoising using a deep count autoencoder," *Nat. Commun*., vol. 10, Jan. 2019, Art. no. 390, doi: 10.1038/s41467-018-07931-2.

[26] T. Chai and R.R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? - Arguments against avoiding RMSE in the literature," *Geosci. Model Dev.*, vol. 7, no. 3, pp. 1247–1250, Jun. 2014, doi: 10.5194/gmd-7-1247-2014.

[27] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5-32, Oct. 2001, doi: 10.1023/A:1010933404324.

[28] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.293978.

[29] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Adv. Neural Inf. Process. Syst.*, 2017, pp. 1024–1034.

p-ISSN 2798-253X
e-ISSN 2798-3153